Åbo Akademi University

Faculty of Social Sciences, Business and Economics,
and Law

September 2023

MASTER'S THESIS

CRYPTOCURRENCIES' PRICES DISCOVERY THROUGH MACHINE LEARNING ALGORITHMS: BITCOIN AND BEYOND.

Mominul Islam / Student ID 1900642
Master's Degree Program in
Governance of Digitalization

Mominul Islam: Cryptocurrencies' Price Discovery Through Machine Learning Algorithms: Bitcoin and Beyond

ÅBO AKADEMI UNIVERSITY – Faculty of Social Sciences, Business and Economics, and Law

Abstract for Master's Thesis

Subject: Information Systems

Author: Mominul Islam

Title: Cryptocurrencies' Prices Discovery Through Machine Learning Algorithms: Bitcoin and Beyond

Supervisor: Prof. Jozsef Mezei

The evolution of cryptocurrencies has emerged as a fundamental shift in the financial landscape, with price discovery being an area of intense interest and complexity. The thesis titled "Cryptocurrencies' price discovery through machine learning algorithms: Bitcoin and beyond" aims to investigate and unravel this complexity through the lens of machine learning.

In this comprehensive study, four major machine learning algorithms - Logistic Regression (LR), Decision Tree, Random Forest (RF), and Support Vector Machine (SVM) were applied to forecast the daily prices of four leading cryptocurrencies: Bitcoin, Ethereum, Cardano, and Solana, alongside an analysis of hourly Bitcoin price prediction.

The findings reveal distinct performance characteristics for each algorithm. Logistic Regression exhibited high accuracies for Bitcoin and Ethereum daily predictions at 0.86 and 0.85, respectively. Support Vector Machine proved particularly effective for Cardano and Solana with accuracies of 0.90 and 0.97. Conversely, the Decision Tree and RF algorithms demonstrated more modest performance across the examined cryptocurrencies. Besides, a specialized investigation into Bitcoin's hourly price prediction, employing the same set of algorithms, yielded varying results, with LR showing a standout accuracy of 0.98.

This research encompasses a journey from the foundational principles of cryptocurrency to the advanced techniques of machine learning, highlighting both the opportunities and challenges inherent in this rapidly evolving field. It acts as a roadmap for future investigations, offering the potential to deepen our understanding of cryptocurrencies' impact on the global financial landscape and to extend the boundaries of knowledge in the area of price discovery through machine learning.

Keywords: Bitcoin, Blockchain, cryptocurrency, fiat currency, machine learning, prediction, traditional financial systems

Date: 09.09.2023	**Number of pages:** 119 + III

Table of Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial intelligence
ARIMA Autoregressive integrated moving average
FBI Federal Bureau of Investigation
IBM International Business Machine
IS Information systems
LR Logistic Regression
LSTM Long short-term memory
ML Machine Learning
MSE Mean squared error
PoW Proof of work
PoS Proof of stake
DPoS Delegated proof of stake
QR Quick response
RF Random forest
RNN – Recurrent neural network
SAS Statistical analysis systems
SHA Secure hash algorithm
SPSS Statistical package for the social science
SVM Support vector machine
USA The United State of America

1. Introduction

Cryptocurrency is a digital currency that utilizes cryptography for transaction security, validation, and the generation of new units (Nakamoto, 2008). Bitcoin, the first cryptocurrency, was introduced in 2009 by an anonymous entity or collective referred to as Satoshi Nakamoto. Following this, a multitude of other cryptocurrencies have emerged, each with their own unique attributes (Narayanan, Bonneau, Felten, Miller, & Goldfeder, 2016). According to Swan (2015), the popularity of cryptocurrencies has surged in recent years, driven by the rapid proliferation of digital devices and a growing global demand for secure, decentralized systems of value exchange.

In contrast to conventional currencies such as the Euro, US Dollar, and British Pound, which are overseen and regulated by government institutions and central banks, cryptocurrencies operate on a decentralized technology called blockchain. This technology encompasses a distributed ledger system that documents every transaction across a network of computers (Böhme, Christin, Edelman, & Mooreet, 2015). Each transaction is verified and recorded by a network of users, referred to as "nodes," ensuring the system's transparency and security. While cryptocurrencies offer numerous potential advantages, including robust security, low transaction fees, and enhanced financial privacy, they also posse various challenges and risks (Narayanan et al., 2016).

1.1 Background

Traditional payment systems depend on third-party intermediaries, like banks and financial institutions, to process and validate transactions involving various forms of value, from cash to electronic funds. These intermediaries not only facilitate fund exchanges between parties but also exert substantial control over these transactions. Conventional financial systems often involve numerous intermediaries, necessitating multi-step communication exchanges that are both time-intensive and prone to errors. These systems also exhibit inherent drawbacks such as lapses in trust, security, privacy, transparency, and adaptability. Given these limitations, there exists an urgent imperative to develop a mechanism that can obviate the need for intermediaries in financial transactions and simultaneously address the above concerns. Ideally,

the proposed system should allow parties to transact funds directly, without external interference or oversight, be it from mediators or governmental bodies.

Peer-to-peer (P2P) distributed systems address the challenges of trust and transparency inherent in traditional payment systems. unlike conventional mechanisms, P2P systems bypass the involvement of third-party entities in the execution of financial transaction. By their very nature, these systems are transparent, distributing the transaction chain among nodes or peers. This framework set the stage for the emergence of a new form of digital currency: cryptocurrency (Patel, Tanwar, Gupta, & Kumar, 2020, p.1). Bitcoin, introduced by Nakamoto (2008) in 2009, was the pioneering decentralized digital cryptocurrency. Since its inception, the growth and adoption rate of cryptocurrencies have skyrocketed. As noted by Best (2022), by 2022 there were approximately ten thousand distinct cryptocurrencies in circulation worldwide, as depicted in Figure 1.

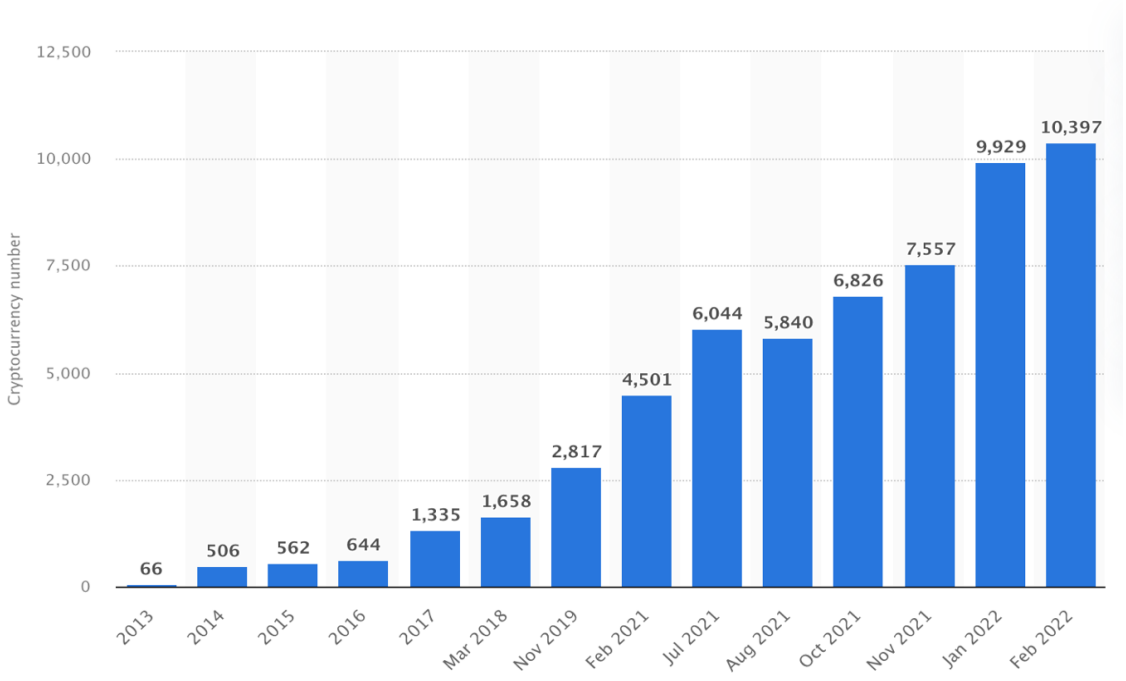

Figure 1: Number of cryptocurrencies worldwide from 2013 to February 2022 (Best, 2022)

Cryptocurrency has burgeoned as a favored investment avenue among both individual as well as institutional investors in recent years. This ascension to the forefront of the financial domain has solidified its status as one of the most vigorously traded financial instruments

globally. Consequently, the dynamics of cryptocurrency pricing have garnered widespread academic attention. Sovbetov (2018, pp. 6-7) underscores that the price of cryptocurrencies is marked by pronounced volatility. It's swayed by an array of determinants including market sentiments, oscillations in the stock market, mining complexities, transaction costs, its burgeoning popularity, the pricing trends of alternative coins, and an assortment of legal quandaries. These elements instill a capricious character in cryptocurrency prices, causing them to oscillate unpredictably over durations. As a result, forecasting these price trajectories morphs into a multifaceted undertaking, positioning it as a pivotal yet intricate research agenda

Cryptocurrency stands as one of the most intricate and enigmatic segments within the financial instruments realm, primarily due to its pronounced volatility. This research paper endeavors to apply machine learning algorithms to historical price data of leading cryptocurrencies, specifically Bitcoin, Ethereum, Cardano and Solana, with the ambition of predicting their future valuations. Following this, the research will juxtapose the precision of the posited methodologies against established models, thereby assessing the efficacy of the introduced frameworks.

1.2 Motivation

Cryptocurrencies have emerged as a global phenomenon in recent years, garnering substantial attention and adoption across the world. Best (2022) reports a staggering 190% increase in global cryptocurrency users from 2018 to 2020, with this momentum further accelerating in 2021, as illustrated in Figure 2. Additionally, Best (2022) highlighted the continued integration of cryptocurrency, notably Bitcoin, into the financial strategies of major public companies, including Tesla, Coinbase, Block, and MicroStrategy. In a landmark move, El Salvador became the first country to recognize Bitcoin as legal tender, setting a historic precedent (npr.org, 2021). Owing to their decentralized architecture, security features, and immutable nature, cryptocurrencies hold immense promise in reshaping the financial landscape.

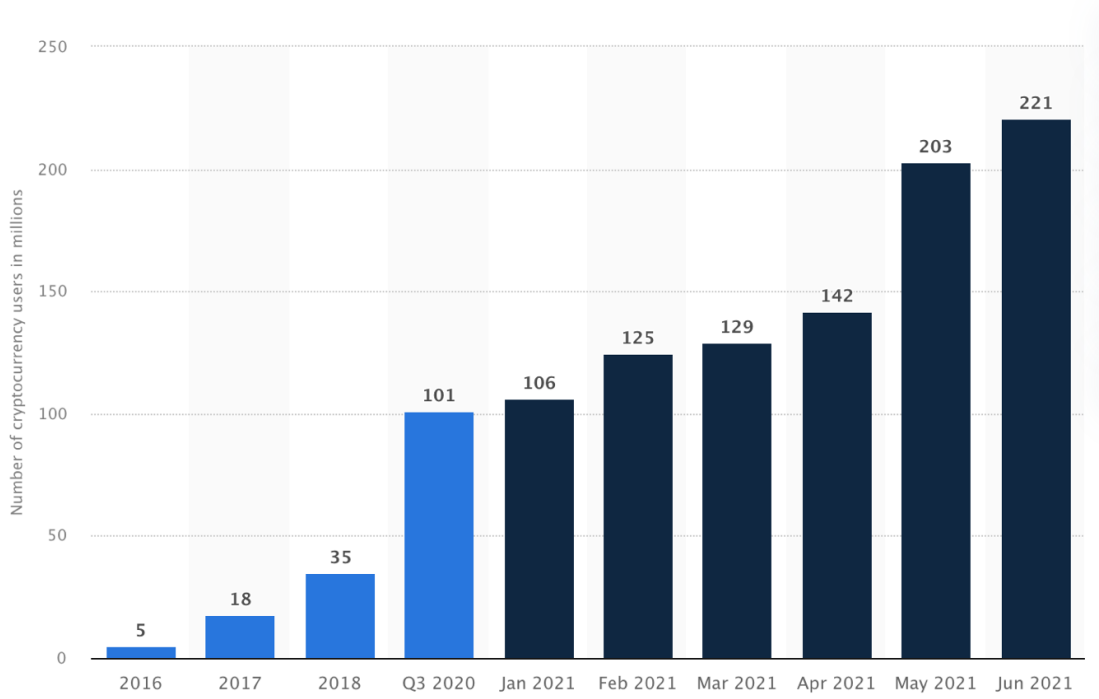

Figure 2: Number of identity-verified crypto asset users from 2016 to June 2021 (Best, 2022)

Owing to its pronounced volatility, cryptocurrency remains in a distinct transitional phase, setting it apart from more traditionally researched financial instruments like stocks, oil, and gold. As a result, forecasting cryptocurrency prices has emerged as a crucial research challenge. While a significant corpus of literature has delved into the use of sophisticated machine learning and deep learning techniques to forecast prices of prominent cryptocurrencies, namely Bitcoin and Ethereum, other potential-rich cryptocurrencies, such as Cardano and Solana, remain underexplored. This study endeavors to develop and test multiple machine learning algorithms, aiming to predict the prices of four specific cryptocurrencies: Bitcoin, Ethereum, Cardano, and Solana.

Machine learning methodologies have emerged as indispensable tools within the financial sector, especially in predicting the price movements of financial instruments, including cryptocurrencies. As a modality for forecasting cryptocurrency price variations, machine learning stands out for its efficacy. When incorporated into business intelligence platforms, these techniques can substantially enhance real-time decision-making, indicating that machine learning algorithms have the potential to revolutionize this sphere. To enable accurate

predictions and forecasts of future cryptocurrency closing prices, a machine must be effectively trained using historical data. Leveraging diverse algorithms, models can be crafted from this data, streamlining the prediction and forecasting endeavor.

Research in the realm of cryptocurrency, though still nascent, is witnessing burgeoning interest, primarily driven by the quest to understand its potential ramifications on the global financial landscape. Academic investigations are particularly attuned to the prospects of blockchain technology and cryptocurrency in crafting robust security measures and streamlining transaction mechanisms. Additionally, there's a growing emphasis on harnessing advanced machine learning techniques to forecast cryptocurrency prices (Böhme et al., 2015). Notably, studies conducted by Rathan, Sai, Manikanta (2019), Chih-Hung, Yu-Feng, Chih-Chiang & Ruei-Shan (2018), and Jain, Tripathi, Dwivedi, Saxena (2018) underscore the pivotal role of machine learning methodologies in predicting cryptocurrency market dynamics.

1.3 Outcome

The insights garnered from this research have the potential to significantly augment the strategies employed in managing and tracking cryptocurrency price movements. Further, this investigation will contrast the proposed machine learning (ML) methodologies with corresponding techniques in extant literature, where ML strategies have been leveraged for predicting cryptocurrency prices. The objective is to discern the most efficacious method that produces superior and competitive outcomes. Implementing such advanced forecasting methodologies can empower both individual and institutional investors by enhancing their understanding of both present and prospective cryptocurrency market dynamics. The hypotheses posited in this research seek to alleviate the challenges faced by stakeholders with existing cryptocurrency investments, as well as those contemplating entering the market shortly. Fundamentally, this research aspires to streamline the trading and investment process in a multifaceted and volatile financial domain like cryptocurrency.

1.4 Research Questions

This study predominantly utilizes a quantitative research method. Additionally, literature reviews and case studies will be taken into consideration during the composition of the

literature review section of this paper. The central theme of the investigation is dissected into four key research questions, encompassing the breadth of the topic. The research questions that this study seeks to address are:

- RQ1: How does cryptocurrency differ from fiat currency?
- RQ2: Which machine learning methods have been used to predict the prices of cryptocurrency in academic research?
- RQ3: What is the most effective machine learning method for predicting the price of cryptocurrency?
- RQ4: Can the same machine learning model be applied to other cryptocurrencies' price prediction?

1.5 Organization of this paper

The organization of this thesis is structured into six main sections, each examining different facets of the study and serving distinct roles in the overall narrative of the research.

The first section, "Introduction," provides a comprehensive background of the topic and articulates the motivation behind the study. This section sets the stage for the research and outlines the expected outcomes and research questions that the study intends to answer.

The second section, "State of the Art," provides a comprehensive review of the relevant literature, establishing a strong theoretical foundation for the study. It covers key topics such as the operation and advantages and disadvantages of fiat currency and cryptocurrency, the technology and operation of cryptocurrency, including blockchain and cryptography. It also outlines the differences between fiat currency and cryptocurrency and introduces the concept of machine learning, its types, and the most relevant research in this area.

The third section, "Methodology," details the research design and the quantitative research method used in the study. It also explains the data collection methods, the process of algorithm building, which includes data exploration and processing, feature selection and engineering, data normalization, algorithm selection, and parameter tuning, and algorithm validation. Furthermore, it presents the performance metrics for algorithm evaluation, such as

classification accuracy, confusion matrix, precision, recall, and F1. Finally, it explains the data analysis methods, which include different learning algorithms and data analysis tools.

The fourth section, "Empirical Results," presents the results of the analysis conducted using various machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machines (SVM). The results are divided by model and further subdivided based on the cryptocurrency for which the price predictions were made, both on daily and hourly timeframes.

The final section, "Discussion and Conclusion," synthesizes the findings and observations from the research and provides an interpretation of overall outcomes. These sections contribute to a better understanding of the study's implications and possible areas for future research.

In essence, this thesis paper is meticulously organized, adopting a structured and logical progression that starts from establishing the research groundwork, discussing relevant literature, detailing methodology, presenting the empirical results, and finally, drawing conclusions from the study.

2. State of the art

The subsequent chapter will elucidate the fundamental concepts forming the bedrock of this study, in addition to surveying the prevailing research on the topic. This foundation amalgamates insights from a confluence of disciplines, featuring investigations into cryptocurrencies, the dichotomies between fiat currency and cryptocurrency, the merits and demerits of cryptocurrency, and the technological underpinnings of cryptocurrency—detailing its creation and operation.

This section will also address the first research question: "How does cryptocurrency differ from fiat currency?" Subsequently, the latter part of this section will tackle the second research question, delving into pertinent scholarly works to discern: "Which machine learning methods have been employed to predict cryptocurrency prices in academic research?"

2.1 Fiat currency

The global currency landscape is both expansive and diverse. Over time, it has transformed from reliance on tangible assets, such as gold, to the adoption of digital currencies like Bitcoin for investments and peer-to-peer transactions. Yet, at the core of the modern financial system lies fiat currency. Hwang (2022) notes that approximately 180 different fiat currencies, including the dollar, euro, and pound circulate worldwide. Rosen (2022) asserts that the value of fiat currencies is dictated by the interplay of demand and supply. Central banks, exemplified by the Federal Reserve Bank in the USA, establish monetary policies to modulate currency supply, aligning it with economic demand (Hwang, 2022).

Currency serves as a medium of exchange for goods and services, a store of value, and a unit of account. Essentially, it is the paper money or coins issued by governments and universally accepted as a form of payment (Frankenfield, 2022). Traditional currencies, such as dollars, euros, and pounds, are often referred to as fiat currency. Unlike commodities like gold or silver, fiat currency holds no intrinsic value; its legitimacy and value are derived from governmental decree (Ginez, 2019, p. 1). Before the 20th century, currencies were typically backed by tangible commodities like gold or silver. However, as time progressed, governments

found it challenging to maintain sufficient reserves of these precious metals to back every note or coin. As a result, fiat money, not tied to physical reserves, gained prevalence. The transition to the current form of fiat currency was notably marked during the 20th century when many nations, including the USA, abandoned the gold standard (Hwang, 2022).

In 1933, President Franklin Delano Roosevelt enacted Executive Order 6102, which mandated U.S. citizens to exchange their gold for U.S. dollars. The primary goal of this directive was to enable the government to print or mint additional currency during times of economic distress (McDonald, 2022). Rosen (2022) posits that the value of fiat currency is in constant flux due to the dynamics of foreign exchange markets. However, one inherent vulnerability of fiat currency is the discretion it grants governments to print money in unlimited quantities. This unchecked power can precipitate hyperinflation and swift price escalations, leading to potential economic calamities, as witnessed in Zimbabwe.

2.1.1 Operations of fiat currency

Fiat currency transactions are characterized by intricate and sometimes protracted processes, as illustrated in Figure 3. These transactions, whether for everyday purchases or international trade, occur regularly. The procedure involves the banking or card network calculating the net settlement position. This denotes the amount the consumer's bank owes the merchant's bank. This information is then communicated to both banks and a settlement bank. The settlement bank subsequently compensates the merchant's bank. In return, the consumer's bank reimburses the settlement bank on the consumer's behalf. As a result, the merchant's bank receives a credit while the consumer's bank incurs a debit. The entire settlement process can span anywhere from 24 to 48 hours, contingent on working days and jurisdictional boundaries (Ginez, 2019, pp. 4-5).

Transferring money across international borders often entails additional time, costs, and the hassles of currency conversion. The process is cumbersome, heavily reliant on information, and involves multiple intermediaries to ensure completion. Despite these complexities, traditional financial systems are deeply entrenched and are generally perceived as trustworthy and reliable by their users (Ginez, 2019, p. 5). As Founders Guide (2021) notes, fiat currencies inherently compel owners to rely on governmental endorsements and third-party intermediaries

like banks or card companies (e.g., Mastercard) for transaction processing and settlements. This setup paradoxically means that while individuals work hard to earn their money, they often lack complete control over it.

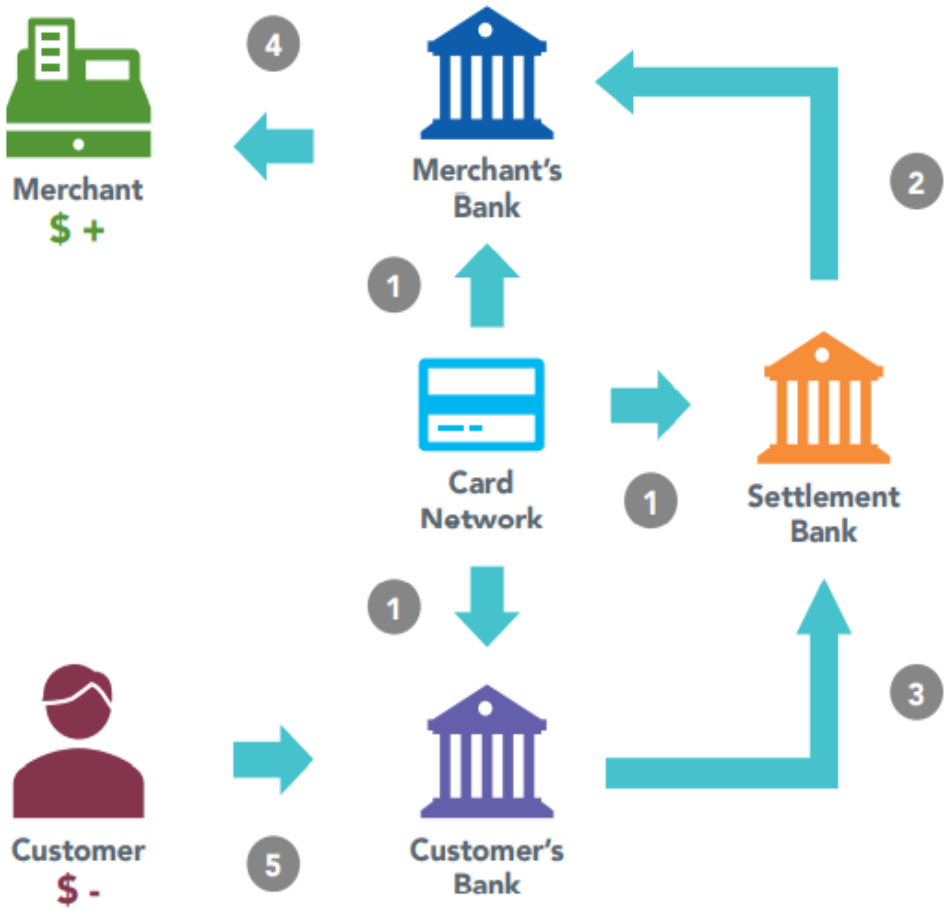

Figure 3: Fiat currency transfer and settlement process (Ginez, 2019, p. 5)

2.1.2 The pros and cons of fiat currency

Every commodity and currency, including fiat currency, inherently possesses advantages and drawbacks. The production, transportation, and exchange of fiat money are notably straightforward, facilitating both national and international trade. As cited by CFI (2022), the ascendancy of fiat currency in the 20th century can be attributed to proactive measures undertaken by governments and banking institutions, which sought to insulate their economies from recurrent downturns inherent in business cycles.

One of the paramount merits of fiat currency lies in its standardization. Within geopolitical boundaries, there is a prevalent consensus among citizens regarding its valuation, thereby facilitating transactions and fiscal planning (McDonald, 2022). Contrary to commodity-backed currencies, such as those anchored to gold or silver, fiat currency's valuation is relatively impervious to external manipulations. To elucidate, entities external to government and central banking structures cannot arbitrarily influence its demand and supply dynamics to skew its value. Hwang (2022) contends that this architecture grants governments and central banks enhanced authority over parameters such as monetary supply, interest metrics, and liquidity, essential tools for mitigating economic crises.

Fiat currency provides financial policymakers with the autonomy to modulate the money supply, aligning it with the economic demand. However, Founders Guide (2021) posits that the operations of fiat currency may not always be optimal. An abrupt surge in the money supply can precipitate sharp inflationary pressures and even hyperinflation. Following the fiscal interventions and economic dislocations engendered by the COVID-19 pandemic, governments worldwide have grappled with the challenge of inflationary containment, as underscored by Rosen (2022). This precariousness has led some critics to champion commodities like gold, which possess inherent supply constraints. They argue, as CFI (2022) highlights, that such intrinsic limitations render commodities more stable compared to fiat currencies, which are unconstrained in their issuance.

2.2 Cryptocurrency

Cryptocurrency is a digital currency that facilitates transfers between individuals and groups without the need for intermediaries. Kaspersky (2022) defines cryptocurrency as a virtual currency that relies on cryptographic algorithms to secure its transactions. Patel et al. (2020, p. 1) elucidate that, given the foundational reliance of cryptocurrencies on blockchain technology, they manifest properties inherent to the blockchain, such as transparency, immutability, and decentralization. Frankenfield (2021) contends that, in contrast to conventional financial systems, cryptocurrency transactions circumvent centralized authorities. Instead, they employ a decentralized mechanism, the blockchain, to both chronicle transactions and issue new units, as illustrated in Figure 4. This decentralized approach ameliorates trust concerns among system

stakeholders. Concisely, cryptocurrency can be characterized as a decentralized digital asset underpinned by a distributed network spanning numerous computers, thus operating autonomously from government and centralized institutional oversight.

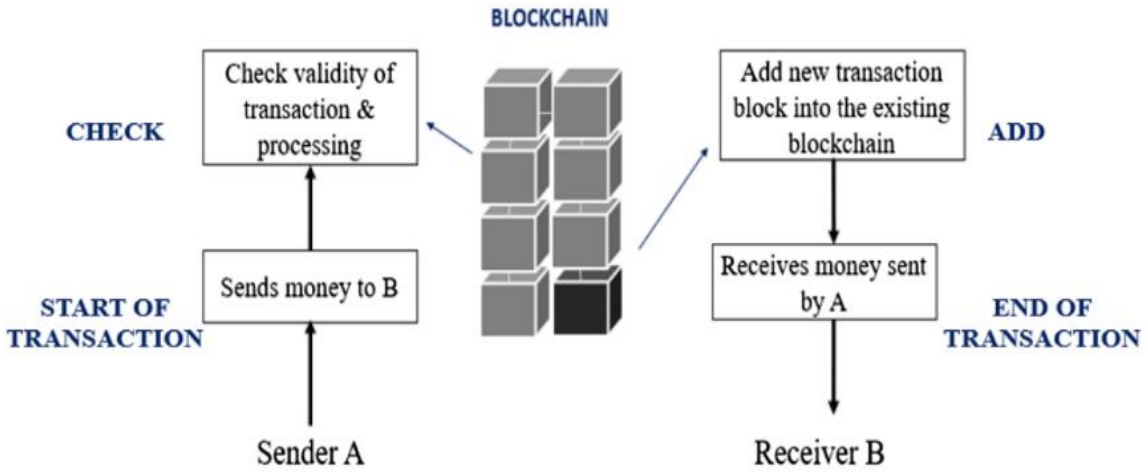

Figure 4: Cryptocurrency workflow using blockchain mechanism (Khedr, Arif, Raj, El-Bannany, Alhasmi & Sreedharan, 2021, P. 5)

Surbhi (2019) described that cryptocurrency is a system that enables secure online payment transactions, denominated as virtual tokens. These tokens represent ledger entries internal to the system. Essentially, such a token can be considered a form of digital cryptocurrency, which, although privately issued, facilitates online transactions. Frankenfield (2022) further elaborates that a crypto token is a specific type of digital currency. It represents a tradable asset or utility situated on its own blockchain. These tokens can serve a myriad of purposes, from investments and stores of value to facilitating purchases.

In 2008, an anonymous researcher going by the pseudonym Satoshi Nakamoto (2008) introduced a seminal paper titled "Bitcoin: A Peer-to-Peer Electronic Cash System." This paper presented a groundbreaking concept of transferring cash online directly between parties, bypassing intermediary financial institutions such as banks. Nakamoto proposed a transaction system that was entirely decentralized, utilizing a peer-to-peer ledger known as the blockchain. This innovative system involved a decentralized chain of validated transactions, or blocks, distributed across all nodes in the network. The integrity and chronological order of these

transactions were maintained using a proof-of-work (PoW) consensus mechanism, reliant on timestamps and cryptographic hashes.

Country	October survey	December survey
Argentina	-	15.2%
Australia	16.8%	22.9%
Brazil	10.1%	14.4%
Canada	10.0%	12.7%
Colombia	7.8%	14.5%
Germany	7.1%	7.7%
Ghana	-	17.3%
Hong Kong	15.7%	16.4%
India	15.4%	23.4%
Indonesia	16.4%	22.4%
Ireland	8.2%	12.1%
Japan	4.6%	5.8%
Kenya	-	15.8%
Malaysia	17.8%	19.9%
Mexico	11.8%	15.2%
New Zealand	6.8%	8.7%
Nigeria	23.6%	21.2%
Norway	4.4%	13.4%
Philippines	12.9%	21.6%
Russia	6.7%	19.6%
Singapore	15.6%	15.8%
South Africa	10.3%	11.3%
Sweden	8.3%	9.8%
United Kingdom	5.2%	6.1%
United States	10.2%	10.5%
Venezuela	-	14.6%
Vietnam	-	28.6%
Average	**11.2%**	**15.4%**

Table 1: Percentage of internet users who own crypto, by country: October 2021 vs. December 2021 (Laycock, 2022)

Since its inception, cryptocurrency has witnessed exponential growth in both usage and adoption. Laycock (2022) estimates that nearly 22 million Americans now own some form of cryptocurrency, with Bitcoin comprising the portfolios of 44.5% of these owners. Furthermore, Laycock highlights a significant surge in global adoption rates, noting that the proportion of internet users owning cryptocurrency rose from 11.2% in October to 15.5% by December 2020. An examination of specific national trends, as detailed in Table 1, reveals that Vietnam, India, and Australia boast the highest cryptocurrency ownership rates among internet users, at 28.6%, 23.4%, and 22.9%, respectively. In a testament to the burgeoning market, Bitcoin, which initially held no tangible value in 2008, achieved a record high of over $68,000 in November 2021. Correspondingly, the collective worth of the cryptocurrency sector has ballooned, now boasting a total market capitalization exceeding 2 trillion US dollars (DeMatteo, 2022).

2.2.1 Advantages cryptocurrency

Cryptocurrencies have garnered significant attention in recent times, largely attributed to their ease of use, trading capabilities, and inherent transactional speed and security. Moreover, their decentralized nature ensures that no single entity or consortium can exert undue control or manipulation over them. The subsequent sub-chapter delves into the advantages of cryptocurrency, drawing from a range of academic sources.

2.2.1.1 Eliminate centralize intermediaries

Cryptocurrencies herald a paradigm shift towards decentralized currency transactions. As posited by Ginez (2019), these digital assets obviate the need for centralized intermediaries, such as banks, in processing and finalizing transactions, a concept depicted in Figure 5. Frankenfield (2022) asserts that cryptocurrencies are designed to simplify online transactions between parties, thus bypassing the imperative for a trusted intermediary, like a bank or credit card company (e.g., Visa). Leveraging dedicated cryptocurrency wallets or prominent exchanges such as Coinbase, Binance, and Crypto.com, users can effortlessly convert traditional currencies like the US dollar into a myriad of cryptocurrencies (e.g., Bitcoin, Ethereum, Cardano) at relatively nominal transaction or network fees.

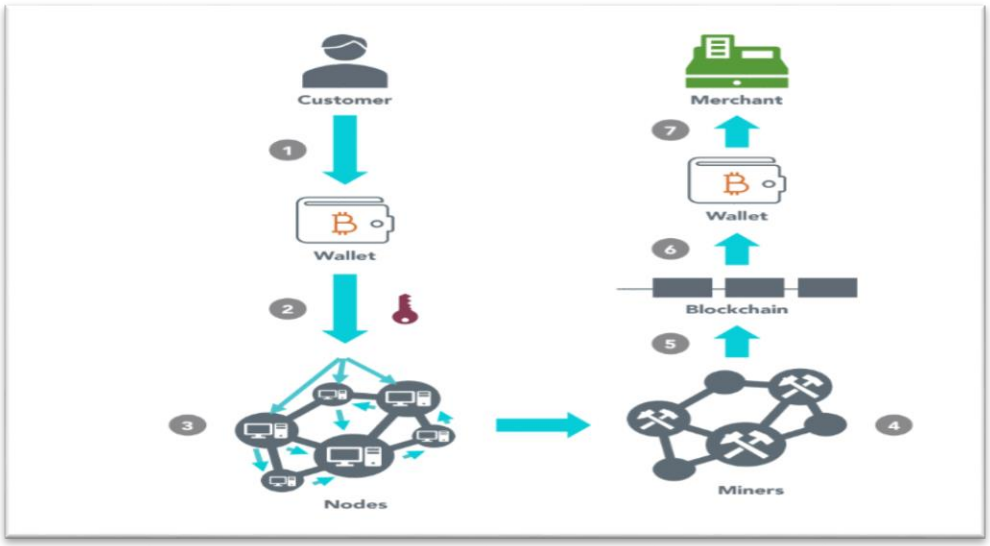

Figure 5: Bitcoin (cryptocurrency) transaction process (Ginez, 2019, p. 6)

2.2.1.2 Protection from inflation

Inflation often causes the value of many fiat currencies to diminish over time, a problem that seems almost inherent to physical currency. As noted by Benzinga (2022), inflationary tendencies predate the advent of paper money. For instance, numerous Roman emperors debased their coinage relative to precious metals to finance their extensive military campaigns. This issue persisted, even into the 20th century, leading to significant economic hardships. However, Geeksforgeeks (2022) posits that cryptocurrencies might offer the most robust safeguard against inflation seen to date. Supporting this assertion, Geeksforgeeks (2022) references the predetermined issuance cap of Bitcoin: per its foundational white paper, only 21 million bitcoins will ever be minted (Bitcoin, 2008). Thus, as demand for Bitcoin rises, its value is projected to escalate commensurately, serving as a bulwark against inflation.

2.2.1.3 Secure and privet

Security and privacy remain paramount considerations in the realm of cryptocurrencies. As delineated by Geeksforgeeks.org (2022), the underlying blockchain ledger is underpinned by intricate mathematical algorithms that are notably challenging to decipher, rendering cryptocurrency transactions generally more secure than their traditional electronic counterparts. Delving further into the aspect of security and privacy pertaining to cryptocurrency wallets, Cryptosecure (2022) emphasizes the significance of the so-called "paper wallet," also colloquially termed "cold storage." A paper wallet is essentially a tangible document containing both the public and private cryptographic keys. Often, for ease of transactional processing, this document features a QR code, enabling straightforward scanning and signing. The inherent security of a paper wallet stems from its physicality: unless one possesses the actual document or is privy to its specific details, accessing its associated funds becomes virtually impossible.

2.2.1.4 Cost effectiveness

Conducting cross-border transfers through conventional financial systems, like PayPal, can be notably expensive due to the substantial fees associated with transaction processing, as

highlighted by Britwise Technologies (2019). However, Geeksforgeeks (2022) posits a contrasting perspective when considering cryptocurrencies. For transactions involving cryptocurrencies across borders, users aren't encumbered by hefty commissions or charges typically levied by banks and other money transfer institutions. Instead, users typically bear a minimal network fee for initiating, processing, and finalizing the transaction. In essence, the decentralization inherent in cryptocurrencies precludes the need for intermediaries such as VISA or PayPal, thereby eliminating superfluous costs.

2.2.2 Disadvantages of cryptocurrency

While there are undeniable advantages to cryptocurrencies, they are not without their shortcomings. Just as every innovation has its set of pros and cons, cryptocurrencies are not exempt from this reality. The subsequent subsections will delve into some of the predominant challenges associated with cryptocurrencies.

2.2.2.1 Illegal transactions

One of the primary concerns surrounding cryptocurrency is its potential for facilitating illicit transactions. Given the high levels of security and privacy inherent to cryptocurrency transactions, tracing and tracking them becomes challenging for regulatory bodies such as the Federal Bureau of Investigation (FBI). As highlighted by Frankenfield (2022), cryptocurrencies, notably Bitcoin, have been instrumental in illegal activities, including drug trades on the dark web. Geeksforgeeks (2022) also points that malefactors exploit cryptocurrency as a conduit to launder money, effectively obfuscating the origins of their ill-gotten wealth. The decentralized and anonymous nature of cryptocurrencies makes them an attractive medium for illicit dealings, such as unauthorized purchases and extensive money laundering operations.

2.2.2.2 Centralize authority

While the founding ethos of cryptocurrencies, epitomized by Bitcoin, was to champion a decentralized financial model, many of its subsequent counterparts, often referred to as Altcoins, deviate from this principle. Cryptocurrencies such as Ethereum, Cardano, Binance

Coin, and Ripple have varying degrees of centralization embedded within their operations. Sanchez (2018) notes that these centralized cryptocurrencies typically have a single authority or a consortium at their helm, accountable for the trajectory and outcomes of the cryptocurrency's endeavors. In essence, the overarching organization retains control over pivotal aspects of the cryptocurrency's ecosystem, encompassing facets like security, privacy, and server operations. Sahu (2020) further expounds on the risks associated with such a model. Centralized entities have the potential to exercise undue influence over coin valuations and, in extreme cases, might even halt server operations unilaterally, sans any forewarning or eliciting stakeholder input.

2.2.2.3 Environmental impact

Mining cryptocurrencies, particularly Bitcoin, is notoriously energy-intensive. This process necessitates substantial computational power and sophisticated computer hardware to decipher complex algorithms and validate cryptocurrency transactions. Criddle (2021) emphasized Bitcoin's energy consumption, highlighting that the computational work to verify Bitcoin transactions consumes a vast amount of energy, making it one of the most energy-intensive operations in the crypto domain.

As per data from the University of Cambridge (2021), Bitcoin's energy consumption has exhibited an increasing trend over the years, with figures revealing consumptions of approximately 57.09 TW/h in 2019, 68.52 TW/h in 2020, and a staggering 104.89 TW/h in 2021, as depicted in Figure 6. Further exacerbating the environmental concerns is the fact that many cryptocurrency miners rely on non-renewable energy sources, such as coal and diesel, to power their mining operations. These energy sources significantly contribute to carbon emissions, leading to an enlarged carbon footprint. Geeksforgeeks (2022) highlighted the detrimental impact of these practices, underscoring the urgent need for more sustainable solutions in the cryptocurrency mining sphere.

Figure 6: Total Bitcoin energy consumption yearly (University of Cambridge, 2021)

2.2.3 Technology and operation of cryptocurrency

The primary aim of this section is to explore the construction and operation of cryptocurrencies. Specifically, it delves into the technologies underpinning the success and functionality of cryptocurrencies. Various pieces of literature will be referenced to substantiate the discussion.

2.2.3.1 Blockchain

Cryptocurrencies, including Bitcoin, Ethereum, Dogecoin, and others, are powered by a foundational technology known as the blockchain. This technology facilitates the transfer of value over the internet, eliminating the need for intermediaries such as banks or card companies like VISA. According to Coinbase (2022), transactions conducted via the blockchain are more secure than those through traditional financial systems. Similarly, Song (2018) pointed that employing the blockchain system for payments does not necessitate the disclosure of sensitive information, ensuring that users' financial and personal data remain uncompromised and protected from theft.

According to Hayes (2022), a blockchain is a distributed database or ledger shared among the nodes of a computer network, where data is stored electronically in a digital format. IBM (2022) offers further insights, suggesting that a blockchain serves as a shared and immutable ledger, enhancing the process of recording transactions and tracking assets within a business network. These assets can be tangible, such as cars, land, cash, gold, or houses, or intangible, like patents, intellectual properties, or copyrights. Essentially, any item of value can be traded and tracked on a blockchain network, reducing costs and risks for all parties involved (IBM, 2022).

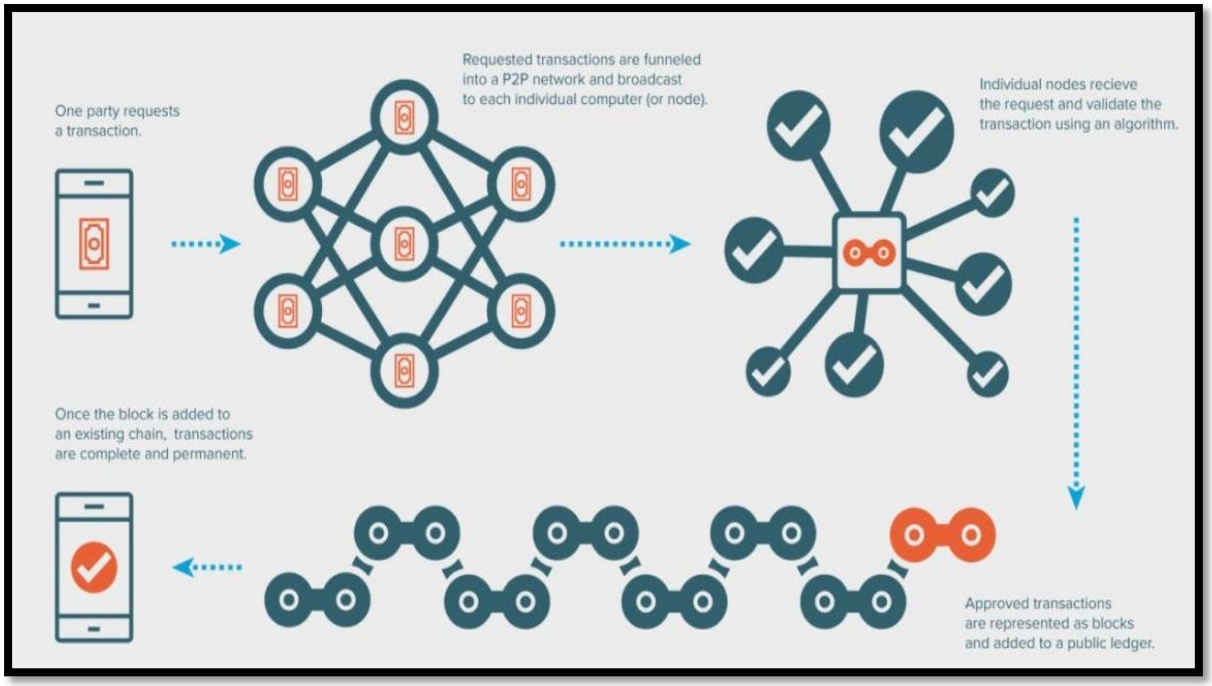

Figure 7: Functionalities of blockchain network (Ginez, 2019, p. 7)

Frankenfield (2022) explained that a blockchain is essentially a series of interconnected blocks or an online ledger. Each transaction is independently validated by every member of the network. Before a new block is added, it must be verified by each node, making it nearly impossible to forge transaction histories. The consensus for every entry on this online ledger must be achieved across the entire network, with every node or computer maintaining a copy of the ledger. The advent of blockchain technology ensures the security and integrity of

recorded data, fostering trust without the need for a centralized intermediary, as illustrated in Figure 7.

The fundamental distinction between a blockchain and a traditional database system lies in their data structures. Hayes (2022) highlighted that while traditional databases typically organize their data into tables, blockchain assembles data into groups, referred to as blocks, which hold specific pieces of information. This structuring of data results in an immutable timestamp when deployed in a decentralized manner. Each block has a set storage capacity. Once this capacity is reached, the block ceases to gather data, subsequently linking to the previously filled block. This creates a chain of data, termed a "blockchain." Every block added to the chain is assured a precise chronological placement.

2.2.3.2 Cryptography

Cryptography is fundamental to blockchain technology, serving as its cornerstone and rendering the technology exceptionally robust. It underpins the key attributes of blockchain, including immutability, reliability, and security. Handscomb (2022) elucidates that the term "crypto" derives from the Greek word meaning "secret," while "graphy" is rooted in the Greek word "Graphein," signifying "to write." Consequently, "cryptography" can be interpreted as the practice of crafting a concealed message, shielded from external observation.

Cryptography is renowned for its pivotal role in safeguarding information across various domains. It not only protects data but also ensures the authentication of both senders and receivers, guarding against repudiation. As Iredale (2021) notes, contemporary cryptography is an interdisciplinary field, drawing on mathematics, computer science, physics, and engineering, among others. According to GeeksforGeeks (2022), cryptography is characterized by techniques and protocols designed to prevent unauthorized third parties from accessing private messages during communication. Lai (2018) delineates several fundamental terms associated with cryptography, including:

- **Encryption**: Encoding texts into an unreadable form.
- **Decryption**: Retaining encryption – transforming a mess message into its original format.

- **Cipher**: An algorithm for functioning encryption or decryption. It is usually a well-explained set of functions that can be followed towards encryption or decryption.

2.2.3.3 Hash

Recently, blockchain technology has surged as a pioneering force across various industrial sectors. Selena (2022) posits that blockchain stands as one of the most defining technological innovations of the past century, shaping the trajectory of future advancements. Crucially, hashes have been instrumental in the evolution of blockchain technology. They also play a pivotal role in cybersecurity and foundational cryptocurrency protocols like Bitcoin. Within the blockchain network, transactions and data storage are fortified and rendered immutable through the application of hashing techniques (Ray, 2017).

A hash, or hashing, is a mathematical operation that transforms any type of input data into a fixed, unique string. Chen (2022) elucidates that a hash, integral to the blockchain's architecture, is constructed based on the information housed within a block header. Bybit (2020) defines hashing as a cryptographic process that underpins the efficacy of blockchain technology. In essence, hashing involves the conversion and generation of input data, irrespective of its length, into a distinctive string of a predetermined size. This transformation is executed by specific algorithms, such as Bitcoin's SHA (secure hash algorithm) 256 bits. Notably, hash algorithms operate as one-way cryptographic functions, meaning the original data cannot be retroactively retrieved (Pande, 2021).

2.2.3.4 Block

A block represents a record of transactional data captured during a specific timeframe within a blockchain network. This data can encompass various details as determined by users, such as the parties involved, transaction amounts, timestamps, and even specific conditions. Morris (2022) describes a block as a structured entity within the blockchain database where transactional data is indelibly recorded. A block continuously captures certain transactions that the network has yet to validate. Once the data within a block is verified, that block is finalized, and a subsequent block is initiated to document new transactions, thereby forming a continuous chain, as illustrated in Figure 8 (Pandey, 2019).

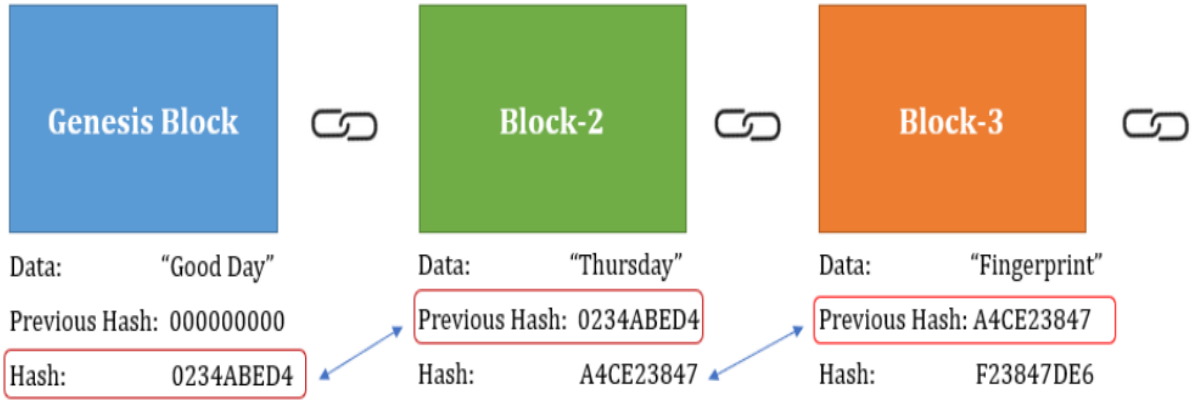

Figure 8: Creation of block (Thomas, 2020)

In a blockchain system, a block records data whenever a transaction takes place. These blocks are identified by extensive numerical sequences that incorporate transactional information from preceding blocks as well as new transaction details (Frankenfield, 2022). IBM (2022) emphasizes that each block not only validates the preceding block but fortifies the integrity of the entire blockchain, thereby manifesting its immutability. Vidrih (2018) contends that once a transaction is inscribed onto the shared digital ledger, no individual party can alter or tamper with it. However, Dutta (2021) points out that if an error is made in a recorded transaction, a corrective transaction must be appended to negate the initial error, making both transactions transparent to all network participants. Thus, a block contributes to the formation of a trustworthy digital ledger, curbing the potential for tampering or unauthorized alterations by malicious actors (Vidrih, 2018).

2.2.3.5 Node

Nodes are fundamental components and the very backbone of blockchain architecture. Their absence would render the storage, validation, and broadcasting of data to other nodes impossible. Tutorialspoint (2022) succinctly defines a node, within the context of digital currencies like cryptocurrency, as a computer connected to a cryptocurrency network, such as

blockchain. This connection facilitates vital functions, including data creation, storage, reception, and transmission. Cryptopedia (2021) further distinguishes that while some nodes focus on ensuring security, maintaining, verifying, and authenticating the public transaction ledger of a blockchain, others are tasked with monitoring network activities.

Nodes play a pivotal role in safeguarding the integrity and security of a blockchain network. Their primary function is to validate the transactions within each block on the network. Each node possesses a unique identification, setting it apart from others (Abrol, 2022). Zeeve (2021) posits that nodes are a hallmark of prominent cryptocurrencies like Bitcoin, Dogecoin, and Ethereum. Becher (2022) further elucidates that nodes essentially establish links among decentralized digital ledgers, which record cryptocurrency transactions. They also disseminate this transactional information across the network to all connected devices. These nodes, often computers, communicate with one another within the network, relaying transaction details and newly minted blocks. Given the inherently decentralized nature of blockchain, anyone globally can operate a node, provided they possess the necessary resources, a high-performance computer and an internet connection, and maintain a connection to the network (Worldcoin, 2022).

2.2.3.6 Consensus mechanism

Consensus mechanisms are pivotal for the accurate operation of blockchain. As posited by Blockgenic (2018), consensus mechanisms comprise a set of protocols or algorithms ensuring that all nodes, typically computers, within the blockchain network are synchronized. This alignment ensures unanimity on which transactions are valid and hence should be incorporated into the network. Through these consensus protocols, the foundational pillars of security, reliability, and trust are established within the blockchain ecosystem.

Elaborating on this, Rosenberg (2022) underscores that consensus denotes the methodologies or criteria employed by a network of interconnected peers, such as nodes, to discern valid from invalid transactions within the blockchain. These consensus algorithms fortify the network against malevolent behaviors and potential hacking onslaughts. Crypto.com (2022) further catalogues various consensus mechanisms prevalent in blockchain networks, like proof of work (PoW), proof of stake (PoS), and delegated proof of stake (DPoS). Their

applications are illustrated in Figure 9. While these mechanisms may differ in energy consumption, security robustness, reliability, and scalability, they are united in their overarching objective: ensuring recorded transactions are both transparent and authentic.

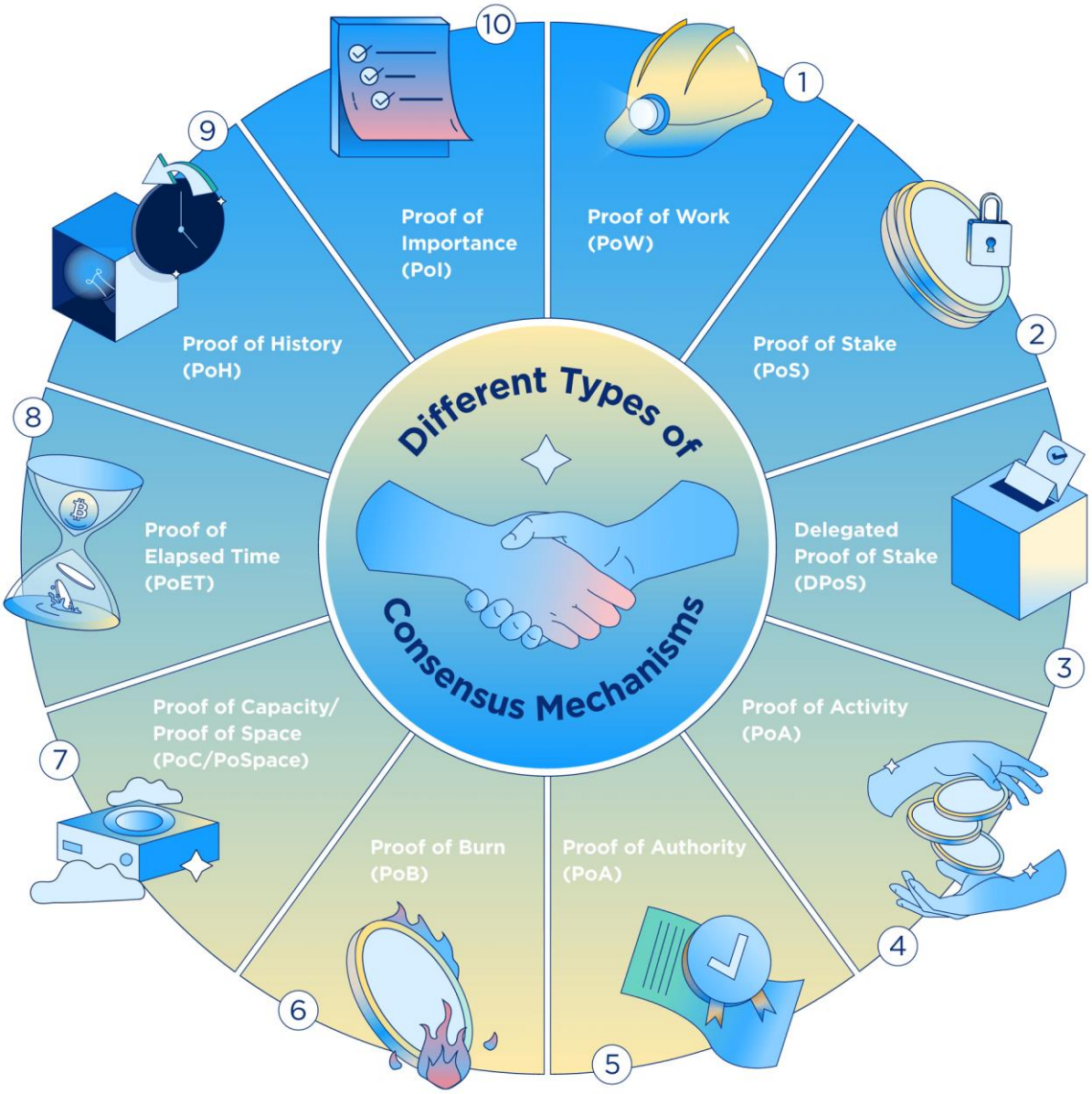

Figure 9: Types of consensus mechanisms (Crypto.com, 2022)

2.3 Differences between fiat currency and cryptocurrency

The difference between fiat currency and cryptocurrency are discussed in the below based on several academic articles as well as research papers.

- Surbhi (2019) explained that fiat refers to the currencies including paper money, coins bills etc. which are accepted nationwide or worldwide such as dollars by the government regulation. On the contrary, cryptocurrency is decentralized and digital exchange of medium which facilitate the transactions between the parties by using encryption technique.
- Fiat money is regulated as well as controlled by the central bank whereas decentralized cryptocurrencies work independently.
- Coin cloud (2021) claimed that the major distinguish between the two currencies is that the transactions or transferring currencies between the parties are direct in cryptocurrencies, since it removes mediators such as bank, which is mandatory in fiat currencies.
- Bitcoin, Ethereum, Ripple, Cardano, Doge Coin, Binance etc. are the most common example of cryptocurrency. As against, Dollars, Euro, Yen, Pound, Ruble etc. are the most popular fiat currency.
- The supply of fiat currency is unlimited, as it can be printed as per need, mentioned Surbhi (2019). Conversely, most of the cryptocurrencies has the limited supply such as Bitcoin has the maximum supply of 21 millions of coin, discussed in the Bitcoin (2008) white paper.
- The transaction fees are negligeable in cryptocurrency in comparison to fiat systems, as it removes the third parties and their fees towards transferring currency.
- According to Coin cloud (2021), cryptocurrency is stored in individual's digital wallet, paper wallet and in some cases in exchange such as Coinbase. On the other hand, fiat money is stored in bank through opening an individual's bank account.

2.4 Machine learning technique

In this section of the study, a variety of advanced machine learning techniques will be introduced to support and enhance the research objectives. Furthermore, a crucial research question for this study, namely, the types of machine learning methods employed in academic research, will be discussed in this chapter.

2.5 Machine learning

Referring to Figure 10, machine learning (ML) is a subfield of artificial intelligence (AI) (SAP, 2022) that endeavors to utilize mathematical data models in enabling computers to learn autonomously, without explicit instructions (Azure, 2022). Brown (2021) contends that artificial intelligence (AI) systems are harnessed to execute intricate tasks in a manner akin to human problemsolving. Building on this, Pryadharshini (2022) elaborates that ML techniques empower computer systems to learn and refine themselves through experiences or data, facilitated by the development of programming languages like Python and R. These languages facilitate automated data access and task execution through detection and prediction mechanisms. The principal aim of ML techniques lies in the identification of data patterns, which are subsequently employed to construct data models for informed decision-making (Burns, 2022). As the volume of data and experiences grows, the precision of ML outcomes also improves; analogous to how human proficiency advances through repeated practice (MathWorks, 2022). Notably, ML techniques rely on computational methodologies to glean insights from data, operating independently of preordained equations as models. These ML algorithms progressively adapt and refine their functions with the accumulation of sample data, exhibiting enhanced performance as the learning dataset expands.

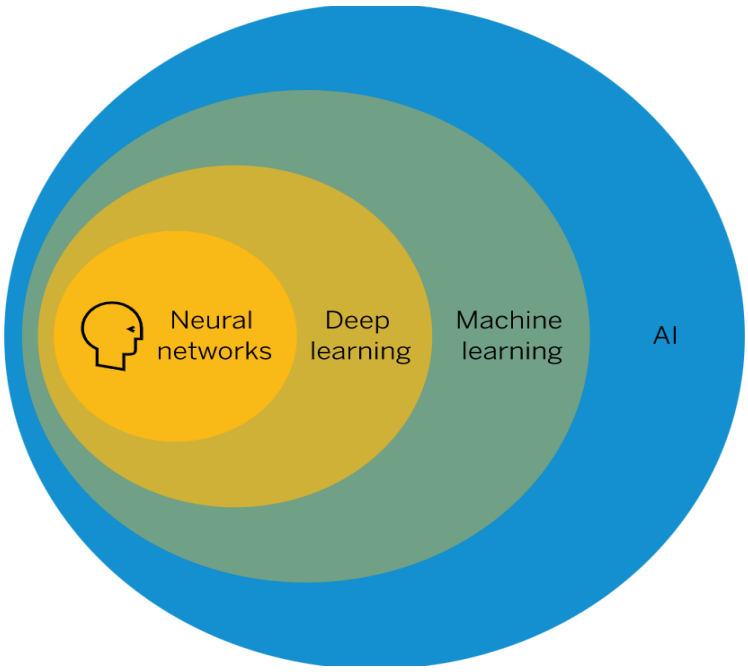

Figure 10: Relationship between AI and machine learning (SAP, 2022)

Machine learning constitutes an indispensable constituent of the continuously burgeoning realm of data science. Through the utilization of refined statistical methodologies, ML algorithms are conditioned to yield classifications and predictions, while also unveiling pivotal data insights, as depicted in Figure 11. These discernments wield a significant influence on crucial growth metrics, contributing to informed decision-making across both commercial enterprises and various application domains (IBM, 2020). Brown (2021) expounds that as the expanse of big data continues to proliferate, the concomitant demand for ML is poised to escalate in parallel. This demand will necessitate businesses and management entities to proficiently discern the most pertinent business insights latent within the data. Consequently, the malleability inherent to ML assumes paramount importance for businesses, organizations, and management entities confronted with dynamic data evolution and growth, coupled with the perpetually shifting landscape of tasks.

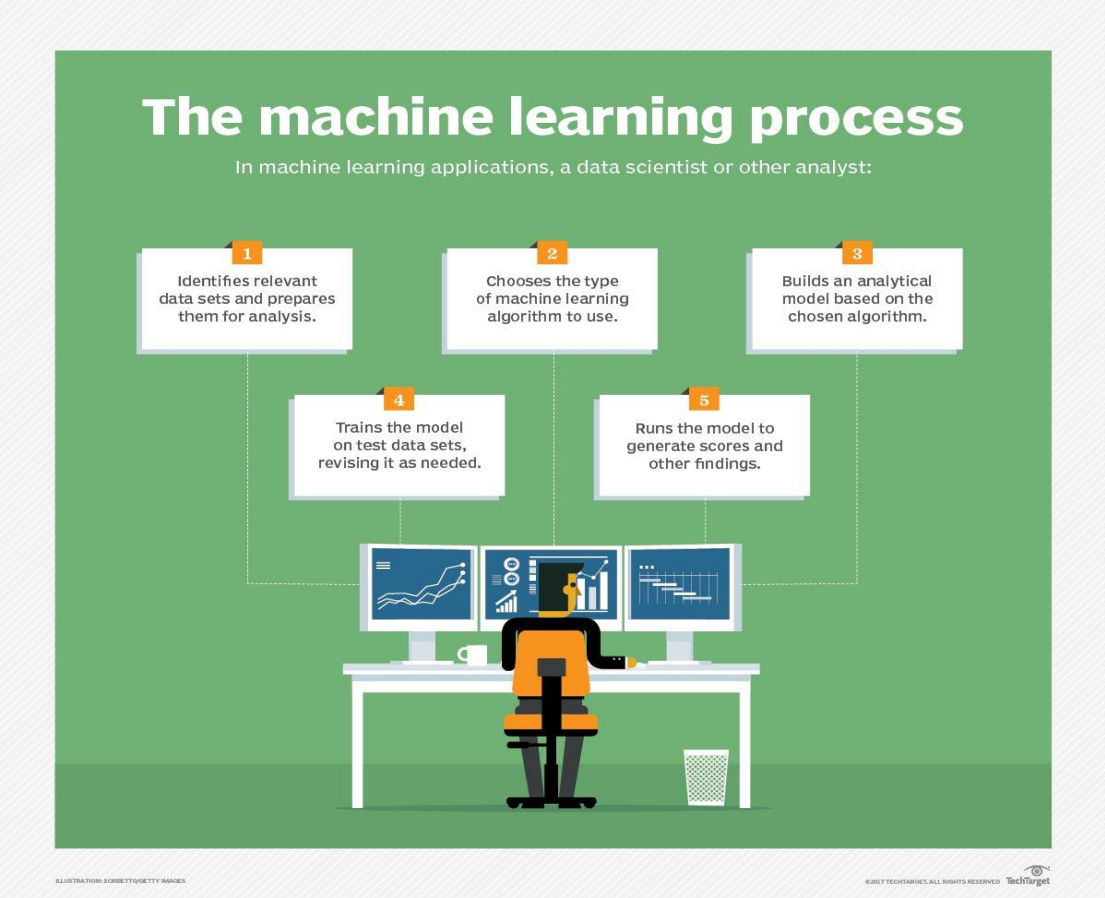

Figure 11: The processes of machine learning (Burns, 2022)

Machine learning starts with data such as numbers, photos, texts etc. Data is collected and prepared to be performed as training data, in other words, the information that the ML model will be trained on. SAS (2022) claimed that the more data assure the better performance. Malone, Rus & Laubacher (2020, pp. 6-7) described that the successful ML methods can be applied in various fields for example the function of ML system can be "Descriptive" such as explaining the event based on data, "Predictive" for instance to predict the event through using historical data and "Prescriptive" for example suggesting the event by using data or experience. Since the main objective of this research paper is to predict the future price of cryptocurrency therefore "Predictive" ML system has been adopted and will be executed throughout the forecasting processes.

The triumph exhibited by machine learning algorithms in forecasting financial instruments such as stocks, bonds, and gold implies a high likelihood of their efficacy in predicting cryptocurrency prices as well. Alessandretti, ElBahrawy, Aiello, and Baronchelli (2018, p. 2) have underscored that, up until now, the deployment of machine learning techniques in the cryptocurrency market has primarily centered around the analysis of Bitcoin and other cryptocurrencies like Litecoin, Ethereum, Ripple, etc. These analyses have leveraged an assortment of algorithms, including random forests, Bayesian neural networks, long short-term memory neural networks, and others. These machine learning methodologies have displayed varying degrees of success in anticipating fluctuations within the crypto market, culminating in the identification of optimal approaches (Alessandretti et al., 2018, p. 2).

However, it is noteworthy that Alessandretti et al. (2018, p. 2) have also acknowledged that the application of machine learning models to predict the prices of limited cryptocurrencies, as carried out by nonacademic sources, lacks benchmark comparisons of outcomes. Consequently, the present author is firmly persuaded to concentrate on an investigation that entails the exploration of the utilization of machine learning techniques in predicting cryptocurrency prices, sourced exclusively from academic literature. This approach aims to circumvent the limitations associated with nonacademic sources and thereby foster a more comprehensive understanding of the subject matter.

2.5.1 Types of machine learning

Machine learning has been categorized into three fundamental domains: supervised learning, unsupervised learning, and reinforcement learning, owing to its intricate nature (Sheikh, 2019). Each of these domains is characterized by distinct performance attributes and objectives. Notably, as asserted by Pryadharshini (2022), supervised learning encompasses approximately 70% of the machine learning landscape, with unsupervised learning accounting for anywhere between 10% to 20%. The remainder is allocated to reinforcement learning. In the ensuing subsections, the researcher will expound upon various types of machine learning techniques, delineating their pertinent application areas and associated advantages.

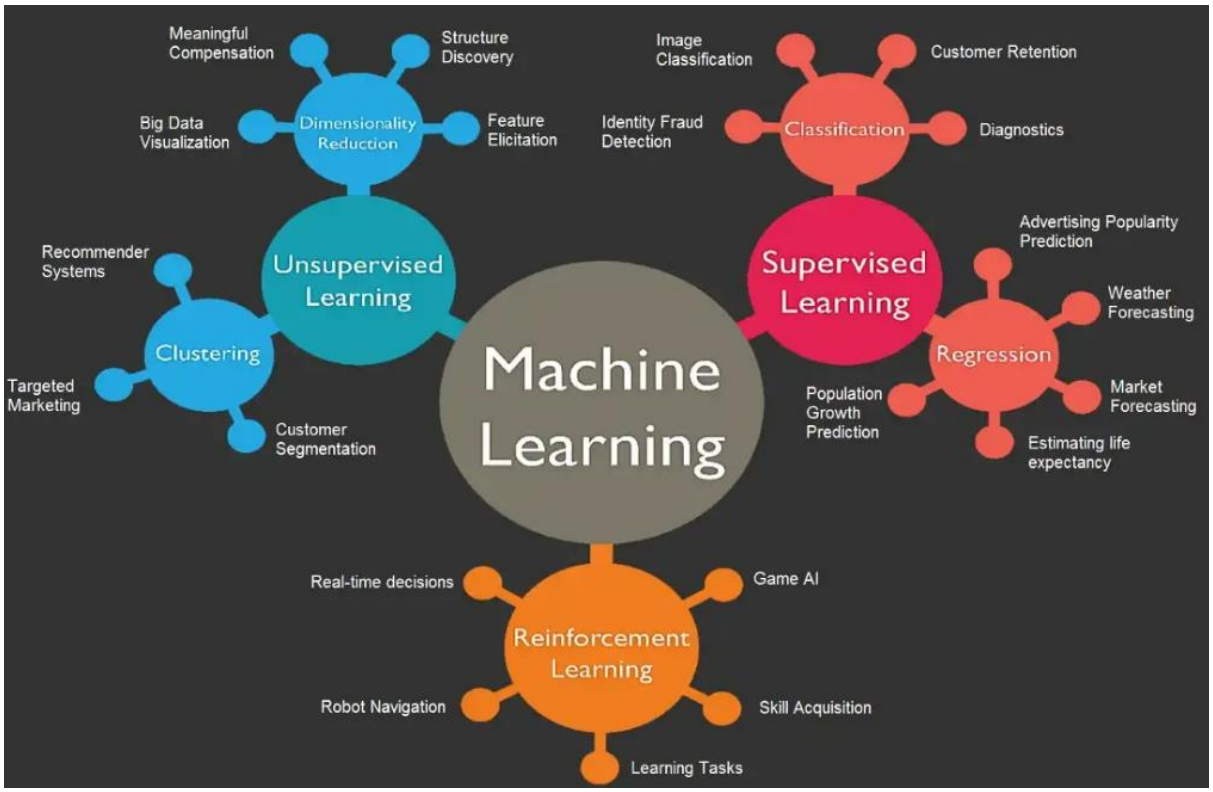

Figure 12: Commonly used Machine learning techniques (Savage, 2022)

2.5.1.1 Supervised machine learning

Supervised learning, a subfield encompassing both machine learning (ML) and artificial intelligence (AI), represents the most prevalent category of ML techniques, extensively employed in the realms of data science and data analysis. As elucidated by Chojecki (2021),

supervised learning is characterized by a paradigm of learning from examples, wherein the training dataset remains distinct from the test dataset. To illustrate, a machine learning model is trained using a collection of images containing various animals, including dogs, cats, pigs, and cows, all meticulously labeled by human annotators. Through this process, the machine learning algorithm autonomously assimilates the distinguishing features necessary to discern images of dogs (Brown, 2021).

According to IBM (2022), the input data is initially fed into the method, and subsequently, the machine learning technique adjusts its weights iteratively until a suitable fit is achieved. This adjustment takes place within the framework of cross-validation, a procedure designed to safeguard against overfitting or underfitting of the model. The primary objective underscoring supervised learning pertains to the accurate prediction of labels for novel, unseen data by leveraging discernible relationships among attributes. In practical applications, supervised learning serves as a pivotal tool for businesses and organizations to address a myriad of real-world intricacies at scale. For instance, it facilitates tasks like classifying spam emails into a separate folder, distinct from users' inbox content.

A repertoire of techniques is harnessed within the domain of supervised learning, encompassing neural networks, naïve Bayes, linear regression, logistic regression, random forest, and support vector machine (SVM), among others (Lang, 2022).

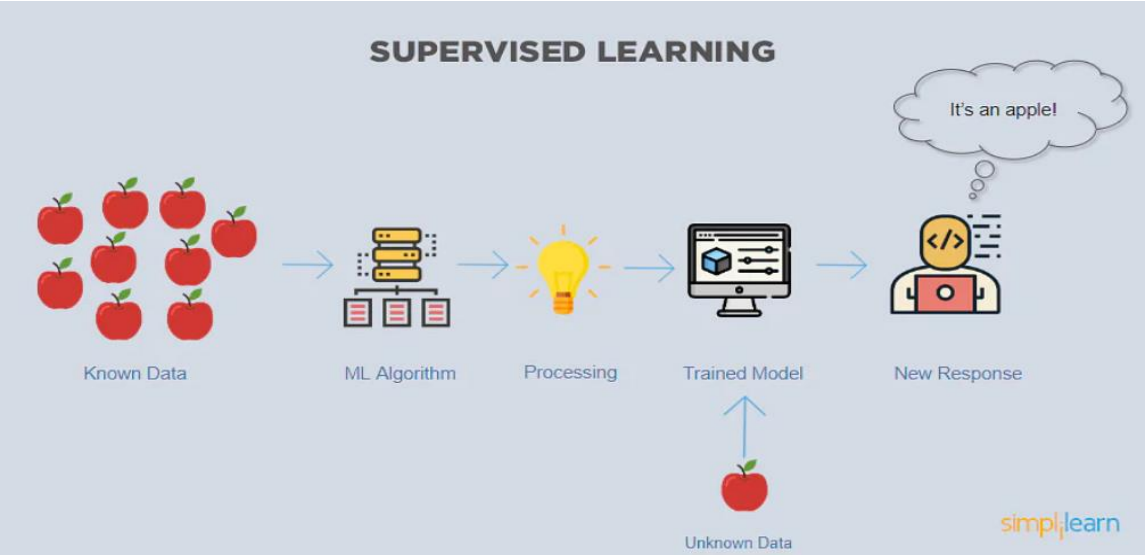

Figure 13: Supervised machine learning (Priyadharshini, 2022)

2.5.1.2 Unsupervised machine learning

Unsupervised learning, alternatively referred to as unsupervised machine learning, constitutes a subset of both artificial intelligence (AI) and machine learning (ML) disciplines. This domain encompasses a range of techniques geared toward uncovering latent patterns and inherent structures within data, devoid of human intervention or guidance (IBM, 2022). As highlighted by Brown (2021), this category of learning comes into play when datasets are bereft of labels or pre-defined categories, and programs are tasked with discerning patterns or trends that may not be immediately evident to human observers. To illustrate, consider an unsupervised machine learning algorithm perusing online sales data to discern and categorize various customer segments based on their purchasing behaviors.

Unsupervised machine learning algorithms excel in their capacity to discern both similarities and disparities inherent within data. As highlighted by IBM (2022), the realm of unsupervised learning extends its utility to feature reduction within methods via dimensionality reduction techniques. This approach proves particularly advantageous for purposes like exploratory data analysis, devising cross-selling strategies, segmenting consumers, and identifying trends within images.

Notably, one of the most prevalent techniques within unsupervised learning is clustering. This method is employed extensively in exploratory data analysis to unveil patterns or groupings concealed within datasets. The applications for cluster analysis are diverse, encompassing domains such as market research, object recognition, and gene sequence analysis (MathWorks, 2022). A plethora of techniques are commonly employed within the domain of unsupervised machine learning, including neural networks, k-means clustering, and probabilistic clustering algorithms, to name a few (IBM, 2022).

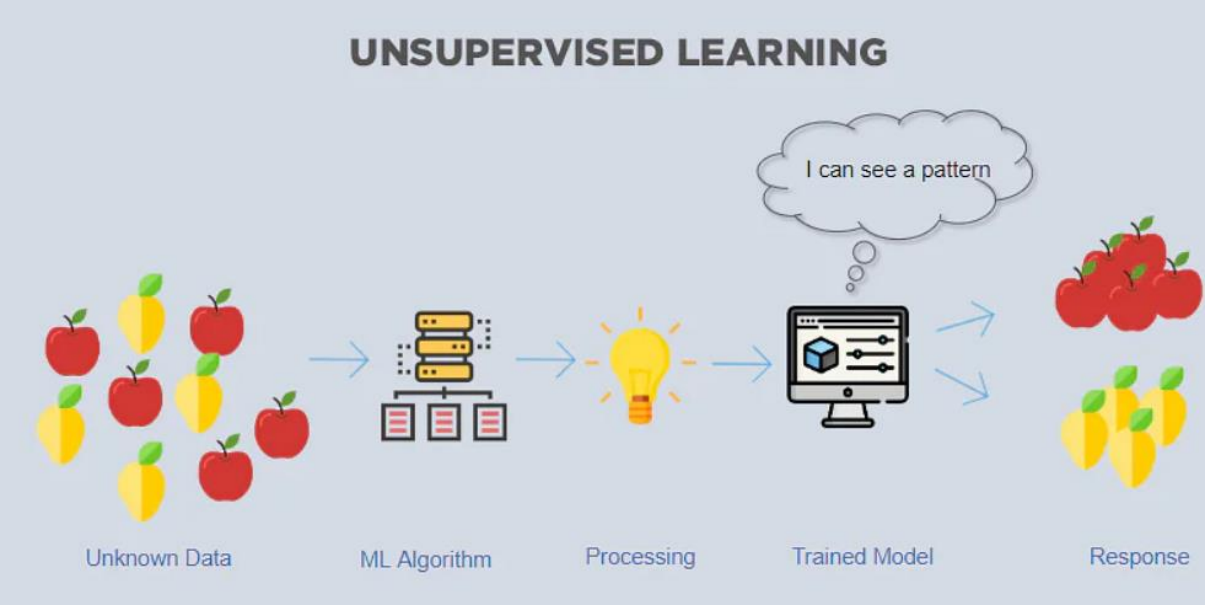

Figure 14: Unsupervised machine learning (Priyadharshini, 2022)

2.5.1.3 Reinforcement machine learning

The imperative for technological facilitation resonates strongly with the aim of streamlining daily existence, enhancing productivity, and fostering informed business choices. The realization of this objective hinges upon the creation of intelligent machines endowed with autonomous learning capabilities, enabling the execution of intricate tasks. Within this context, the significance of reinforcement machine learning emerges prominently, as underscored by Great Learning (2022).

Reinforcement learning has emerged as a promising frontier within the realm of machine learning, primarily addressing the complexities of sequential decision-making tasks that are often embedded in conditions of uncertainty. As elucidated by IBM Developer (2023), reinforcement learning stands as a machine learning algorithm analogous to supervised learning, albeit diverging in the manner it is trained. Unlike supervised learning that relies on labeled sample data, reinforcement learning adopts an approach of trial and error. It leverages a sequence of successful outcomes to reinforce the formulation of optimal recommendations for a given task. This methodology equips algorithms to master tasks such as game playing by making informed decisions. Furthermore, it extends its purview to the realm of instructing

autonomous vehicles and robots, facilitating self-driving capabilities through precise decision-making processes (Brown, 2021).

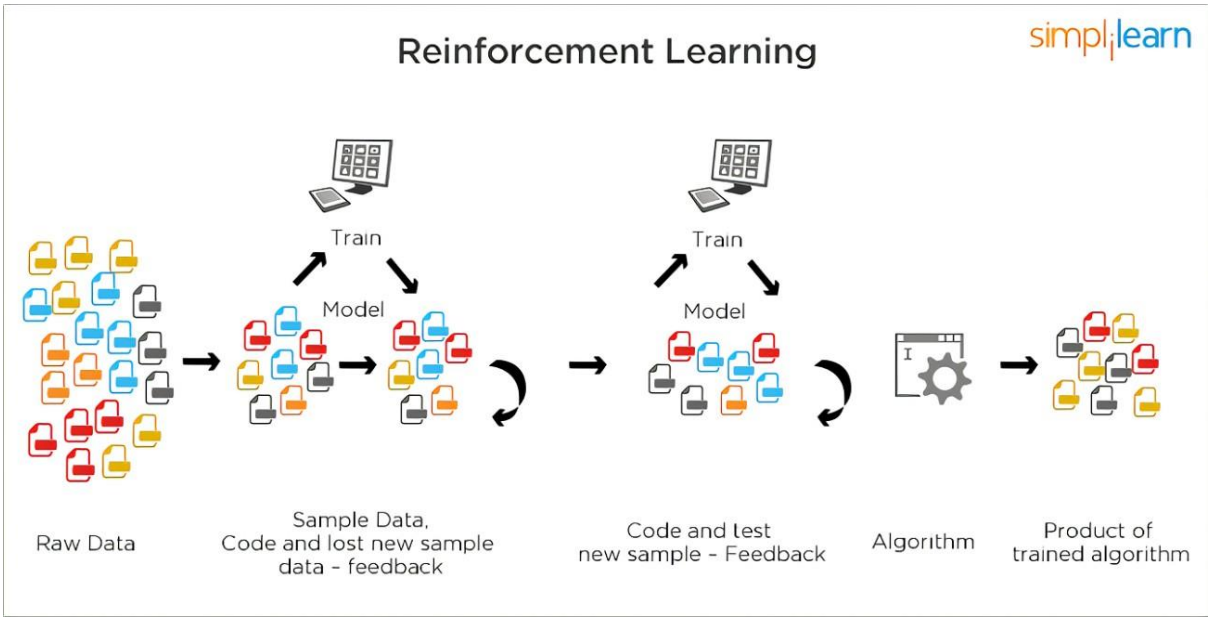

Figure 15: Reinforcement machine learning (Simplilearn, 2023)

2.6 The most relevant related research

Cryptocurrency stands as a novel digital asset within the finance domain, characterized by an exceedingly elevated level of volatility when juxtaposed with conventional financial instruments like stocks, bonds, and gold. The inherent instability and price oscillations within this domain have engendered a paucity of literature focusing on the prediction of high volatility assets, specifically cryptocurrencies. To the best of the researcher's knowledge, only a limited number of articles delve into this realm. This section briefly alludes to prior research endeavors that pertain to cryptocurrencies and their price prediction through the utilization of machine learning algorithms.

Within the realm of academia, several methodologies have been postulated to anticipate the price trajectories of diverse cryptocurrencies, including but not limited to Bitcoin and Ethereum. These methodologies have been subject to rigorous assessment, wherein their predictive accuracy has been contrasted against pre-existing methods, as elucidated in the accompanying table.

Greaves & Au (2015) have introduced Bitcoin transaction graph that contains every Bitcoin transaction data generated prior to 7th April 2013 (Greaves et al., 2015, p. 2) to forecast the Bitcoin prices. This study has proposed four different kind of classification methods such as baseline, logistic regression, SVM and neural network. The paper has achieved the accuracy applying these aforementioned models are 53.4 percent for baseline, 54.3 percent for logistic regression, 53.7 percent for SVM and 55.1 percent for neural network respectively (Greaves et al., 2015, p. 6). On the other hand, Bakar & rosbi (2017) have proposed an autoregressive integrated moving average (ARIMA) predicting model to explore the accuracy of Bitcoin exchange rates in a high volatility market. This paper opted to analyze monthly data concerning the exchange rate of Bitcoin, spanning from January 2013 to October 2107 (Bakar et al., 2017, p. 131). The investigation yielded absolute percentage errors of 1.4 percent for September 2017 and 9.3 percent for October 2017, respectively. The collective outcome was summarized by a mean absolute percentage error of 5.36 percent, reflecting the disparity between predicted and actual values (Bakar et al., 2017, p. 136).

In their study, Rathan et al. (2019) delved into an exploration of diverse machine learning algorithms, including decision trees and regression models, to ascertain the efficacy of predicting Bitcoin prices and to conduct a comparative analysis of their predictive accuracies. The results yielded a compelling insight: linear regression emerged as an efficient method for Bitcoin price prediction. Notably, the linear regression algorithm exhibited an impressive accuracy of approximately 97.5 percent in forecasting Bitcoin prices over a 5-day interval. In contrast, the decision tree algorithm achieved a slightly lower accuracy of 95.8 percent (Rathan et al., 2019, p. 193). The authors further asserted that their proposed model outperformed the accuracies achieved by existing methods, thereby demonstrating its heightened predictive capabilities.

Pang, Sundaraja, Ren (2019) explored cryptocurrency price through "Cryptocurrency price prediction using time series and social sentiment data" research paper. The chosen period of time such as October 2017 to March 2018 at an hourly interval, researchers have explored various types of ML methods towards predicting the price of cryptocurrency (Pang et al ,2019, p. 37). Scholars of this paper were concern and well informed upon the volatility of the market that challenge to forecast the price of cryptocurrency. However, their major motivation was

advance ML techniques that provide data-driven decision which help as well as motivate the investors to invest with minimal risk and maximum return. Pang et al (2019, pp. 41 - 42) claimed that buy and sell signals generated by simple moving average can generate 1800 dollars profit with an initial investment of 4329.69 dollars within a period of six months, exponential moving average can earn a profit up 6878.48 dollars from October 2017 to March 2018 with an initial investment 3888.74. Moreover, Pang et al (2019, p. 41 - 42) pointed out that decision tree models with sentiment data would suggest that investor would suffer 458.2 dollars lose based on the signals from October 2017 to March 2018. However, Pang et al (2019, p. 42) recommend that other supervised machine learning models such as random forest and XGboost can be applicable towards predicting the price movement of the cryptocurrency.

Kim, Kim, Kim, Im, Kim, Kang & Kim (2016) have studied on user comments in online cryptocurrency communities towards forecasting fluctuations of the cryptocurrencies' prices as well as in the number of transactions thereof. The Granger causality test was adopted for this paper, which is widely applied in research on the value of currencies as well as shares (Kim et al., 2016, p. 5). The study has obtained the accuracy of the forecasted fluctuation in Bitcoin price and in Bitcoin transaction are 49.462 percent and 45.161 percent for 12 days. Ethereum, on the other hand, the predicted accuracy of price fluctuation and transaction fluctuation are 50.286 percent as well as 54.286 percent respectively for 12 days. Moreover, in the case of Ripple, the price fluctuation accuracy is 53.157 percent for 12 days (Kim et al., 2016, pp. 11 – 12). Notably, Kim et al (2016) did not consider Ripple transaction fluctuations for the prediction.

McNally, Jason Roche & Simon Caton (2018) have applied recurrent neural network (RNN), long short time memory (LSTM) network as well as autoregressive integrated moving average (ARIMA) methods in order to forecast the price direction of Bitcoin in USD. RNN and LSTM are two deep learning pipelines, which outperformed the ARIMA predicting model. Root mean squared error (RMSE) has been introduced to compare as well as evaluate the regression accuracy and an 80/20 holdout validation strategy is applied to implement the validation of techniques. As a result, the accuracy and RMSE achieved applying ARIMA technique are 50.05 percent and 53.74 percent for the length of 100 days. For 100 days, RNN model gives the accuracy as well as RMSE 50.25 percent and 5.45 percent, respectively. In

contrast, LSTM method provides the accuracy and RMSE 52.78 percent as well as 6.87 percent for 100 days (McNally et al., 2018, p. 342).

Chih-Hung et al. (2018, p. 174) propose a new long short-term memory (LSTM) predicting approach towards forecasting bitcoin daily price with two various LSTM algorithms such as conventional LSTM method as well as LSTM with AR (2) method. The performances of the applied techniques are assessed using daily bitcoin price data during the period of 01/01/2018 to 28/07/2018 (Chih-Hung et al., 2018, p. 171). The study obtained Root mean square error (RMSE) for both LSTM and AR (2) 256.41 and 247.33, respectively. The researchers claimed that AR (2) method provides better forecasting accuracy in comparison to conventional LSTM model (Chih-Hung et al., 2018, p. 174). Moreover, Jain et al. (2018) investigated two different cryptocurrencies, Bitcoin and Litecoin, and attempted to forecast their future prices by applying a multi-linear regression model. The accuracy of their model was 44 percent for Litecoin and 59 percent for Bitcoin over a duration of four days (Jain et al., 2018, p. 4).

Learning model	**Cryptocurrency**	**Accuracy**	**Duration**
RNN, LSTM and ARIMA	Bitcoin	50.05%, 50.25% and 52.78%	100 days
Multi-linear regression	Bitcoin and Litecoin	59% for Bitcoin and 44% for Litecoin	4 days
Granger causality test	Bitcoin, Ethereum and Ripple	49.462%, 50.286% and 53.157%	12 days
Conventional LSTM method and LSTM with AR (2) method	Bitcoin	RMSE for both methods 256.41 and 247.33 respectively	Daily
Linear regression and decision tree	Bitcoin	97.5% and 95.8%	5 days
Baseline, logistic regression, SVM and neural network.	Bitcoin	53.4%, 54.3%, 53.7% and 55.1% respectively	Daily

Table 2: Relative comparison of existing methods for cryptocurrency price prediction

3. Methodology

This section of the research paper will explore the methodology employed, drawing upon a range of academic sources including empirical studies and review articles. Moreover, the specific methodology adopted for this thesis will be elaborated upon and clarified, in alignment with the research objectives.

3.1 Research design

Research is an endeavor aimed at garnering insights on a given phenomenon by applying scientific rigor and academic astuteness (Jennings, 2001, p. 13). Essentially, the research process encompasses three fundamental steps, including formulating questions, gathering responses to the queries, and conducting analysis to present the findings (Blankenship, 2022). Furthermore, Jansen & Warren (2020) elucidate that research methodology is a systematic strategy of conducting a study to ensure the validity and reliability of the findings in addressing the research objectives and aims.

Research design in information systems entails a meticulous planning and execution of studies aimed at generating new knowledge and insights into the respective field. Creswell & Creswell (2017, p. 4) define the research design as a "plan or blueprint for conducting a study that elucidates the principal elements of the research process, encompassing data type, sampling strategy, and data analysis methods". The choice of research design hinges on the research questions and objectives, as well as the nature of the phenomenon under scrutiny.

The experimental design is a prevalently used research design in information systems. It focuses on manipulating one or multiple variables and observing their impact on an outcome variable. This design is potent in examining causal relationships between variables, often utilized in studies investigating the effectiveness of technology interventions (Feldman & Lynch, 1988). Another type of research design is the survey design, which involves collecting data from a sample of participants via questionnaires or interviews. This design is instrumental in probing attitudes, beliefs, and perceptions of users concerning technology. The survey design

has found application in studies examining the adoption of e-commerce (Karahanna, Straub & Chervany, 1999).

Moreover, various other research designs are commonly adopted in academic research concerning information systems, including case studies, action research, and ethnography. Each research design offers its own unique strengths and limitations, and the choice of design should be guided by the research question and objectives (Creswell et al., 2017). Generally, academic research design in this field utilizes one of three types of methodologies: qualitative, quantitative, and mixed methods.

In any research project, researchers must have a clear vision and intention concerning expected outcomes and results (McCombes & George, 2022). This foundational principle is especially relevant to the current study. In light of the specific research objective, forecasting future prices of cryptocurrencies using various machine learning techniques and leveraging historical market data, the researcher has chosen to employ a quantitative research methodology for this thesis. This quantitative approach is particularly instrumental in this study for predicting future cryptocurrency prices. The complexity of the cryptocurrency market, influenced by factors such as market trends, news events, and social media activity, makes quantitative methods advantageous. These methods offer a more comprehensive understanding of these influencing factors through numerical data analysis, facilitated by the use of advanced machine learning algorithms.

3.1.1 Quantitative research method

Quantitative research generates insights derived from specific survey or observational data through statistical analysis (Veal, 2006, p. 40). It is generally more application-oriented in data collection and analysis, with the aim to describe, explain, predict, or control phenomena, compared to alternative research methods like qualitative research. The intricate nature of numerical data analysis requires a systematic approach (McCombes & George, 2022). Quantitative methods are crucial for shaping future policy directions in government and business organizations; they are widely employed in diverse fields such as marketing; sociology; economics; human development; political science; information systems; and community health (Godfrey & Clarke, 2000, p. 191).

Quantitative research in information systems involves the utilization of numerical and statistical techniques to investigate phenomena associated with the design, enhancement, and application of information systems. This research methodology aims to test hypotheses and establish cause-and-effect relationships by collecting and analyzing numerical data. Survey research, including data collection through self-report measures like questionnaires or interviews, is one of the most prevalent methods applied in quantitative research in information systems (Danesi & Ghelardi, 2018). Another common technique is experimental research, which concerns the manipulation of one or more independent variables to examine their impact on a dependent variable (Venkatesh, Brown, & Bala, 2013, p. 23). The findings of quantitative research in information systems are often scrutinized using statistical techniques, such as regression analysis or structural equation modelling (Byrne, 2016, p. 3).

The approach of the quantitative research method involves statistical analysis, which requires calculating measures such as mean, median, standard deviation, and variance. These calculations are typically performed using software or programming languages like SPSS, Python, or R (Study.com, 2021). The method relies on numerical evidence to achieve desired outcomes and draw conclusions. Reliability is a key consideration in academic research, as emphasized by Veal (2006, p. 1). To enhance the reliability of the results, it is often necessary for researchers to utilize relatively large datasets. In this context, programming languages like Python serve as crucial tools for their robust data analysis capabilities (Study.com, 2021). After collecting data from various sources, Python is specifically employed to analyze the data, thereby facilitating the achievement of the research objectives.

Quantitative research methods can be classified into three main categories, based on the nature of the data obtained (Veal, 2006, p. 10). The first category, questionnaire surveys, involves posing questions to target groups to collect standardized responses. The second category is observational methods, which may include either direct observation or participant observation, depending on the level of researcher involvement. The third category comprises the use of secondary data sources, where data is acquired from existing case studies or from the analysis of pre-existing texts or datasets (Veal, 2006, p. 20). This research paper will employ secondary data sources, such as historical trading data of cryptocurrencies, to gather the necessary information for predicting future prices.

3.2 Data collection methods

Data collection methods are an essential aspect of research in information systems. These methods serve to gather data and information that can subsequently be analyzed to yield insights into the behavior and performance of information systems. Data collection approaches can be broadly categorized into two groups. The first, known as primary data, is collected directly by the researcher. The second, referred to as secondary data, is gleaned from existing sources, such as articles or prior research projects (Veal, 2006, p. 147). Various techniques are commonly employed in the collection of data within the realm of information systems research, including but not limited to surveys, interviews, case studies, experiments, and secondary data analysis.

A variety of data collection approaches are utilized in information systems research to obtain information on subjects such as participant behavior, system performance, and technology adoption. Each method has its own set of advantages and disadvantages, requiring researchers to judiciously select the technique most aligned with their research questions. Surveys, interviews, case studies, experiments, and secondary data analysis are some of the most commonly employed data collection methods in this field. These techniques have proven effective in providing critical insights into the behavior and performance of information systems.

In this study, which aims to forecast the future prices of cryptocurrencies, the collection of historical price data is essential. However, it's essential to note that the data collection techniques used for predicting cryptocurrency prices should be carefully chosen and rigorously evaluated. This is to ensure that the data is both high-quality and directly relevant to the predictive tasks at hand. The quality of the data significantly impacts the accuracy of machine learning algorithm predictions. Given these considerations, a secondary data analysis method has been selected for this study. This approach will facilitate the collection of a large and diverse set of historical data, which is essential for accurate future price predictions. Various open data sources offer high-quality cryptocurrency datasets; these include platforms like investing.com, CoinMarketCap API, CoinMetrics, GitHub, Google BigQuery, and Kaggle. Such sources provide publicly accessible, high-quality datasets that can be effectively utilized for research and analysis.

Secondary data analysis is a critical technique in quantitative research, as it employs existing data to explore research questions that differ from those intended by the original data collectors. This method is particularly valuable in fields like social sciences, medicine, information systems, and public health. It provides access to large, representative datasets that would be impractical or cost-prohibitive to gather through primary data collection methods (Lee, 2008, p. 84). Compared to primary data collection, secondary data analysis often proves to be more efficient because the data is typically readily accessible, thus reducing the time and resources needed for data gathering (Ritchie & Lewis, 2003, p. 4).

However, one notable challenge of secondary data analysis lies in the validity and reliability of the data. Because researchers do not directly collect the data themselves, they may be unaware of potential errors or biases present in the original dataset (Bryman, 2012, p. 384). Additionally, secondary data may not encompass all variables of interest or might lack the granularity needed to address specific research questions (Gibson & Koziol, 2012, p. 98). Despite these limitations, thorough scrutiny of data sources, judicious choice of analytic techniques, and the inclusion of relevant covariates or moderators can enhance the validity and reliability of secondary data analysis. Consequently, this approach can offer valuable insights and make meaningful contributions to the field.

3.2.1 Algorithm building

Cryptocurrency price prediction presents a significant challenge due to the high volatility and complexity inherent in the market. Algorithm building is a critical phase in the development of appropriate machine learning models for cryptocurrency price prediction. According to a study by Hafid, Hafid & Makrakis (2022), the algorithm-building process for cryptocurrency price prediction involves several crucial steps. These include data collection, data exploration and processing, feature selection and engineering, as well as data normalization. This section will discuss these essential steps in detail to provide a comprehensive overview of the model-building process for cryptocurrency price prediction.

3.2.1.1 Data collection

Data collection serves as the initial phase in the machine learning algorithm-building process. It involves the accumulation of raw data from diverse sources, including cryptocurrency exchanges, social media platforms, and news outlets, to generate a holistic view of the market (Paudel and Kim, 2018). For this study, datasets for daily price prediction have been collected for four cryptocurrencies such as Bitcoin, Ethereum, Cardano, and Solana from Investing.com (Investing.com, 2023). Additionally, a dataset for hourly price prediction for Bitcoin has been sourced from Kaggle (Kaggle, 2023). Both platforms are reliable and high-quality data sources related to stocks, currencies, commodities, and cryptocurrencies, among others. The datasets gathered for this thesis contain all historical price data, including open, high, and low prices, for the selected cryptocurrencies. These datasets are ideal for exploratory data analysis, time-series analysis, and predictive modeling tasks. They can be utilized to study historical price trends, examine correlations among different cryptocurrencies, and identify seasonal fluctuations in price. Moreover, this data can be employed to build models that aim to forecast future prices for specific cryptocurrencies.

These datasets can serve a variety of stakeholders, including data analysts, data scientists, traders, investors, academics, and anyone keen to explore the dynamics of the cryptocurrency market. Their application is intended to facilitate research and analysis of market behaviors and the factors influencing cryptocurrency prices.

3.2.1.2 Data exploration and processing

Data exploration and processing are critical phases in the process of preparing data for machine learning applications. These steps are essential to ensure the accuracy of the chosen method. Data exploration and processing involve identifying and correcting inconsistencies, as well as addressing missing values in datasets to guarantee their quality and reliability. According to a study by Paudel et al. (2018), data processing encompasses the tasks of identifying and rectifying errors in the data, including outliers, missing values, and noise. The effectiveness of machine learning algorithms is heavily dependent on the quality of the input data. As demonstrated in research by Batista, Prarti, and Monard (2014), data processing serves to mitigate the adverse effects of data issues on model performance.

Data exploration and processing are essential techniques in machine learning for addressing various issues in datasets, including outlier detection, imputation, and feature scaling. These steps are also crucial for eliminating duplicate records, identifying and removing irrelevant or redundant variables, and properly encoding categorical data. Numerous academic studies underscore the importance of these preprocessing steps in enhancing the performance of machine learning models. For example, Brownlee (2017) demonstrated that optimizing data through methods like duplicate record removal and missing value handling is critical for improving model accuracy. Similarly, Kotsiantis, Zaharakis, and Pintelas (2006) found that data preprocessing techniques could significantly reduce the error rates in classification models.

Overall, data exploration and processing are vital steps in preparing data for machine learning and play a significant role in enhancing both the accuracy and reliability of models. In pursuit of these objectives, this paper will rigorously employ a range of data exploration and processing techniques. These will include the removal of duplicate records, identification and elimination of irrelevant variables, and the handling of missing values. By adhering to these best practices, the study aims to optimize the performance of the machine learning models under investigation.

3.2.1.3 Feature selection and engineering

Feature selection and feature engineering are crucial steps in the machine learning pipeline aimed at optimizing the performance of predictive models. Feature selection involves identifying and choosing the most relevant variables from a dataset, while feature engineering entails creating new variables through transformations or combinations of existing ones. These processes serve to reduce data dimensionality, enhance model accuracy, and mitigate the risk of overfitting. Various techniques have been proposed for feature selection and engineering, including genetic algorithms, principal component analysis, and clustering-based methods. Moreover, deep learning approaches have been employed for automated feature extraction, obviating the need for manual feature engineering (Schreck, 2018).

Research has shown that feature selection and engineering can significantly influence model performance. For example, a study focusing on breast cancer diagnosis using machine learning found that feature selection improved the model's effectiveness by reducing the number of variables and preventing overfitting (Fernández-Delgado, Cernadas, Barro & Amorim, 2014). Similarly, a study on stock price forecasting demonstrated that feature engineering increased model accuracy by creating new variables based on technical analysis indicators (Xie, Zhang, You, Cao, & Liu, 2019).

Feature selection and engineering are crucial techniques for improving the accuracy of cryptocurrency price prediction models. Feature selection involves choosing a subset of the most relevant features from the dataset, while discarding redundant or less informative ones, as noted by Chen, Xu, Liu, Wang, & Liu (2019). Research has demonstrated that these processes can significantly enhance the accuracy of models used for cryptocurrency price prediction. For example, a study by Chen et al. (2019) employed feature selection techniques like correlation-based feature selection and mutual information feature selection to identify the most pertinent features for accurate cryptocurrency price forecasting.

Feature engineering involves the creation of new features by altering or amalgamating existing ones. In the domain of financial instruments, including stocks, bonds, and cryptocurrencies, feature engineering can incorporate the development of technical indicators such as moving averages, relative strength indices, and stochastic oscillators, as detailed by Xie et al. (2019). These engineered features can uncover hidden patterns and trends in financial price movements, thereby improving the robustness of predictive models. For instance, research by Khoo, Lim, & Tan (2021) leveraged techniques like wavelet transformations and principal component analysis to generate new features for cryptocurrency price prediction.

In summary, feature selection and engineering are pivotal techniques for enhancing the precision of cryptocurrency price forecasting algorithms. The cornerstone of a model's success lies in the features used during its training phase. Incorporating multiple, independent features that individually correlate well with the target variable increases the likelihood of effective learning. Consequently, this study will focus on identifying the most salient features, as well as generating new ones, allowing machine learning algorithms to better grasp the intricate relationships among variables and thereby optimize predictive accuracy.

3.2.1.4 Data normalization

Normalization is a critical phase in data pre-processing for machine learning, as several algorithms are sensitive to the scale of the input features. The primary objective of normalization is to adjust feature values to a common scale, ensuring that each feature contributes equally to the learning process. Kotsiantis, Kanellopoulos, and Pintelas (2006) identify two commonly used normalization techniques: Min-Max normalization and Z-score normalization. Expanding upon this, Hastie, Tibshirani, and Friedman (2009, pp. 57-59) explain that Min-Max normalization scales feature values to a fixed range, typically between 0 and 1, by subtracting the minimum value of each feature and dividing by its range.

Z-score normalization scales feature values to have a zero mean and unit variance by subtracting the feature's mean and dividing by its standard deviation. This form of normalization enhances the convergence of gradient-based methods and the generalization performance of the resulting models (Hastie, Tibshirani & Friedman, 2009, pp. 57-59). It also increases the model's robustness to outliers and improves interpretability. Nevertheless, Goodfellow, Bengio, and Courville (2016, pp. 204-206) caution that it is imperative to apply normalization after splitting the data into training and validation sets to prevent the introduction of bias into the model.

Kotsiantis et al. (2006) further discuss the significance of both discretization and normalization in machine learning. Discretization involves reducing the number of possible values for a continuous feature by categorizing them into bins. This is undertaken to circumvent slow and inefficient learning that can arise from having an excessive range of possible values. In contrast, normalization involves scaling down features to ensure that all features contribute equally to the model, irrespective of their original scale. This is particularly important because the scales of different features can vary substantially.

3.2.1.5 Algorithm selection and parameter tuning

The processed data is now primed for the subsequent phase: algorithm selection. According to Shalev-Shwartz and Ben-David (2014, p. 144), this step involves choosing the most

appropriate learning algorithm along with its optimal parameters for a specific task. This process is also known as the selection of the model class and its hyperparameters, a concept elaborated by VanderPlas (2017, p. 348). The parameters necessary for the selected learning algorithms are contingent upon various factors such as the type, quantity, and adaptability of these parameters, which will be elaborated upon in Section 3.4.

Algorithm selection is a critical and iterative process in machine learning. It entails conducting a series of trials over multiple iterations, during which parameters are fine-tuned to identify an optimal set of configurations. Bennett and Parrado-Hernández (2006, p. 1266) have posited that a model is trained through solving an optimization problem, which refines the algorithm parameters according to a specified loss function and potentially a regularization function. However, Murphy (2012, p. 24) has emphasized that no single method is universally superior; a set of assumptions that performs well in one domain may not be equally effective in another. This notion is commonly referred to as the "No Free Lunch Theorem".

3.2.1.6 Algorithm validation and resampling methods

After the algorithm selection phase, the subsequent step is to validate the chosen algorithms. The primary aim of validation is to ascertain that the parameters were effectively optimized during the preceding algorithm selection phase, as articulated by VanderPlas (2017, p. 35). The validation process is also iterative in nature, as each iteration from the algorithm selection stage requires validation through multiple techniques.

The primary objective of validation is to accurately assess the algorithm's capacity to generalize to new, unseen data. According to Hastie et al. (2009, p. 219), the results of the validation process serve to estimate the algorithm's error rate, which is indicative of its predictive efficacy on independent test data and its genuine generalization performance. Validation constitutes a critical stage in the process of algorithm development, as it assists in both selecting the most robust model and critically evaluating the quality of that selection. A variety of techniques for algorithm validation are available, among which the validation set approach, K-fold cross-validation, and bootstrapping are the most prevalent. This thesis will employ the K-fold cross-validation method to validate the chosen algorithms.

3.3 Performance metrics for algorithm evaluation

Upon successful completion of both algorithm selection and validation, the model's generalization performance can be rigorously assessed using a variety of quantitative evaluation metrics. These metrics serve as indicators of the algorithm's error rate and overall efficacy. The subsequent sub-section will provide a comprehensive discussion of these quantitative evaluation metrics, including classification accuracy, the confusion matrix, precision, recall, and the F1 score.

3.3.1 Classification accuracy

In line with the research conducted by Marsland (2015, p. 23), classification accuracy stands as the most prevalently utilized metric for evaluation. It is computed by taking the ratio of correctly classified instances to the total number of instances in the dataset, as elucidated in Figure 17. Despite its widespread use, this metric has limitations, especially in scenarios characterized by class imbalance, situations in which one class is disproportionately overrepresented compared to others. Such imbalance can result in distorted findings; for instance, an algorithm that systematically classifies all instances into the majority class would yield a high classification accuracy but might, in reality, be an ineffective model. Additionally, in multiclass settings, classification accuracy fails to provide insight into the specifics of misclassification. Consequently, it is advisable to employ additional metrics alongside classification accuracy for a more nuanced evaluation.

3.3.2 Confusion matrix

The confusion matrix is an alternative technique to assess the progress of a classification model alongside the classification accuracy. As Harrington (2012, p. 143) described, the confusion matrix shows the misclassification mistakes that occur when a model tries to classify instances into several classes. The matrix is usually displayed as a square grid, with the actual output label displayed on the x-axis, and the model's forecasted output on the y-axis. If all of the off-diagonal components in the confusion matrix are zero, then a perfect classifier with a 100% classification accuracy has been obtained.

In binary classification problems, the confusion matrix is a 2x2 matrix that explains the possible outcomes "1" or "-1", as shown in figure 18. For example, a true positive (TP) occurs when an instance is appropriately forecasted as belonging to the positive class ("1"), while a false negative (FN) occurs when an instance is inappropriately forecasted as belonging to the negative class ("-1"). Likewise, a true negative (TN) occurs when an instance is correctly forecasted as belonging to the negative class ("-1"), while a false positive (FP) occurs when an instance is inappropriately predicted as belonging to the positive class ("1") (Harrington, 2012, p. 144).

		Predicted	
		+1	-1
Actual	+1	True Positive (TP)	False Negative (FN)
	-1	False Positive (FP)	True Negative (TN)

Figure 16: A confusion matrix for a binary classification problem. The possible outputs for the two categorical output labels ("1" and "-1") are displayed in statistical terms (Harrington, 2012, p. 144)

Overall, the confusion matrix serves as an indispensable instrument for evaluating the performance of a classification model. Not only does it furnish detailed insights into the model's accuracy, but it also elucidates the specific classes that the algorithm may be misclassifying, thereby indicating areas where further optimization is warranted.

3.3.3 precision, recall, and F1

Although the confusion matrix effectively identifies which target labels have been conflated, it does not inherently account for imbalances among classes. Nevertheless, Marsland elucidates in his book 'Machine Learning: An Algorithmic Perspective' (2015, p. 23) that there exist metrics capable of addressing such imbalances. These metrics, derived directly from the values within the confusion matrix, include precision, recall, and the F1 score. Moreover, the classification consistency can be computed as the sum of true positives and true negatives, divided by the overall number of instances in the dataset.

Geron (2019, p. 92) assorted that precision is typically significant when the cost of false positives is high. For instance, in a medical diagnosis task, a false positive outcome may lead to unnecessary medical processes, which can be costly and possibly harmful to the patient. In such cases, a high precision is desirable to reduce false positives. Recall, on the other end, is important when the cost of false negatives is high, explained Marsland (2015, p. 23). For example, in a fraud detection task, a false negative outcome may lead to significant financial loss for a business. In such cases, a high recall is desirable to optimize false negatives. Both precision and recall can be derived from the formula shown in figure 19.

F1 score is often exercised to assess the overall execution of a classification technique because it takes into consideration both precision and recall. F1 score is particularly effective when there is an imbalance in the data set, where one class is significantly smaller compared to the others. According to Marsland (2015, p. 23), the F1 score can provide a more appropriate scale of a model's performance compared to accuracy, especially in such imbalanced data sets. F1 can be derived from the formula presented in Figure 17.

In summary, precision, recall, and the F1 score serve as integral metrics for evaluating the efficacy of a classification algorithm. The selection of a specific metric should be dictated by the unique constraints and objectives of the research problem in question. High precision becomes a priority when the cost associated with false positives is substantial, whereas high recall gains prominence when mitigating the cost of false negatives is paramount. The F1 score frequently serves as a balanced measure of an algorithm's performance, especially in scenarios characterized by data set imbalances.

$$\text{Accuracy} = \frac{\#TP + \#TN}{\#TP + \#FP + \#TN + \#FN}$$

$$\text{Precision} = \frac{\#TP}{\#TP + \#FP}$$

$$\text{Recall} = \frac{\#TP}{\#TP + \#FN}$$

$$F_1 = 2\,\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Figure 17: Formulas for how the four numeric performance metrics can be derived from a confusion matrix (Marsland, 2015, p. 23)

3.4 Data analysis methods

Data analysis methods involve applying statistical and computational algorithms to transform and explain raw data into meaningful insights. These methods are commonly exercised in various fields such as scientific study, social sciences, information systems, business, and healthcare, to discover patterns, correlations, and trends in large as well as complex data sets.

Data analysis techniques are essential for scholars to modify raw data into meaningful insights. Several data analysis methods use in academic research, each with its strengths and weaknesses. According to a study by Dang, Zhang, Huang & Zheng (2018) demonstrated that descriptive statistics, involving measures such as mean, median, and standard deviation, address a comprehensive understanding of the central tendency, variability, and distribution of data. Inferential statistics, including hypothesis testing and regression analysis, are applied to make generalizations and predictions regarding a population based on a sample, another study demonstrated by Murray, Mills & Johnson (2017). Machine learning algorithms leverage statistical models and computational techniques to discern patterns and relationships within data for predictive purposes. These algorithms are especially efficacious in handling vast and

intricate datasets, as emphasized by Alpaydin (2010). Moreover, data visualization techniques, ranging from graphs to charts, are instrumental for providing a succinct yet comprehensive representation of data. Such visual aids facilitate rapid identification of trends and patterns, a notion supported by Wickham (2016).

3.4.1 Learning algorithms

Learning algorithms, also familiar as machine learning algorithms, are a subdivision of artificial intelligence that enable computers to learn from and make predictions or decisions based on data without being explicitly programmed (Mitchell, 1997). These algorithms are based on mathematical and statistical principles, including probability theory, linear algebra, and optimization techniques, to process and analyze data (Bishop, 2006). The following section will provide an intuitive description regarding learning algorithms as well as the reasons for selecting them for this study.

3.4.1.1 Logistic regression

Logistic regression (LR) is a statistical approach commonly used in machine learning for binary classification tasks, such as forecasting whether an event will occur or not (IBM, 2023). According to Jaquart, Kopke, and Weinhardt (2022, p. 338), the logistic regression technique serves as a straightforward and simply trainable benchmark model, against which more complex models are compared. LR is equivalent to simple linear regression but is specifically utilized for binary response variables, such as in classification problems. It models the probability of a binary event occurring based on a linear combination of predictors. Unlike standard linear regression, logistic regression does not have a closed-form solution, but the global optimum can be efficiently explore using numerical techniques due to the convexity of the loss function. It is essential to notice that the LR technique is the only technique used for inference without creating an ensemble of individual algorithms trained with multiple seeds, as it has a unique solution and is not subject to a stochastic optimization process (Greaves et al., 2015, p. 5).

In the context of cryptocurrency price prediction, logistic regression has obtained significant importance due to its ability to analyze and forecast market trends. Several studies have

highlighted the impact of logistic regression in predicting cryptocurrency prices, making it a crucial tool for traders and investors. One key aspect of logistic regression that makes it relevant in cryptocurrency price prediction is its ability to model the relationship between multiple variables and the likelihood of an event occurring, such as whether the price of a cryptocurrency will increase or decrease, demonstrated a study by Akyildirim, Goncu & Sensoy (2021, p. 15). Jaquart et al. (2022, p. 338) elaborated that LR considers multiple features, such as cryptocurrencies' historical price data, trading volumes, market sentiment, and technical indicators, to create a probabilistic forecasting of the direction of price movement. By utilizing these variables, LR can capture complex patterns and trends in the cryptocurrency market, allowing for more accurate price predictions.

Furthermore, Akyildirim et al. (2021, p. 15) explained that logistic regression offers interpretability, which is crucial in the cryptocurrency market, where understanding the underlying issues driving price movements is critical for informed decision-making. Traders as well as investors can analyze the coefficients of the LR technique to detect the most influential factors affecting cryptocurrency prices. This provides valuable insights into the dynamics of the market, enabling better risk management and investment strategies. Moreover, Jaquart et al. (2022, p. 338) claimed that LR can be integrated with other machine learning techniques, such as feature selection, ensemble methods, and time-series analysis, to optimize the accuracy and robustness of cryptocurrency price prediction models. These combined methods can help overcome the challenges of cryptocurrency market volatility, data noise, and non-linearity, resulting in more reliable predictions.

3.4.1.2 Decision tree

A decision tree is a type of supervised learning algorithm that is used for both classification and regression tasks. It is a non-parametric approach that employs a hierarchical, tree-like structure containing a root node, branches, internal nodes, and leaf nodes, as illustrated in Figure 18 (IBM, 2023). In each decision tree, the first node is known as the root node and serves as the initial point for splitting the data. VanderPlas (2017, pp. 421-422) describes that the root node uses a function to determine the best feature for dividing the instances in the dataset. Subsequent split nodes then ask sequential questions to further narrow down the options, often taking the form of axis-aligned splits utilizing cutoff values within features. This

results in two categories or branches from each split node, each leading to its own subsequent split node.

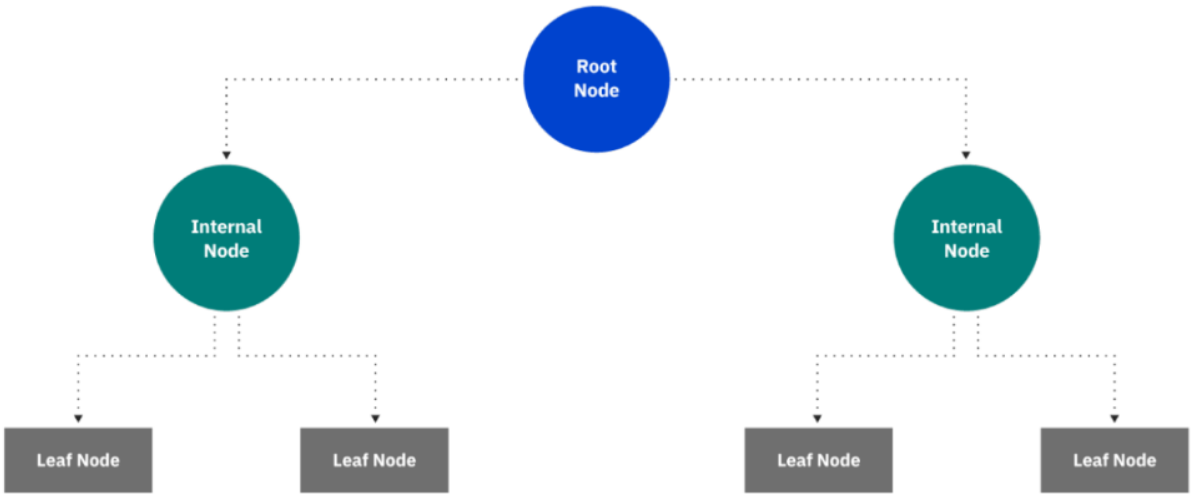

Figure 18: Decision tree (IBM, 2023)

Decision trees are an essential technique in cryptocurrency price forecasting due to their capability to capture complex patterns and relationships within the data (IBM, 2023). As a non-parametric supervised learning algorithm, decision trees can effectively manage nonlinear and non-monotonic relationships that often characterize cryptocurrency price data. Decision trees allow for the identification of optimal splitting or root points in the data, which assist in determining crucial features and their respective cutoff values for making predictions. Furthermore, Rathan et al. (2019) assorted that decision trees are capable to handle both categorical and continuous features, making them versatile for analyzing several types of cryptocurrency data.

3.4.1.3 Random forest

Random Forest is a critical ensemble learning technique that integrates several decision trees to generate a robust and appropriate prediction model (Breiman, 2001). Each decision tree in a Random Forest is trained on a random segment of the data with replacement (i.e., bootstrapped), and a random subset of features is responsible for splitting at each node, leading

to various and independent trees, as depicted in Figure 20. The final prediction of a Random Forest is achieved by averaging or taking a majority vote of the predictions of all the single trees (Liaw & Wiener, 2002).

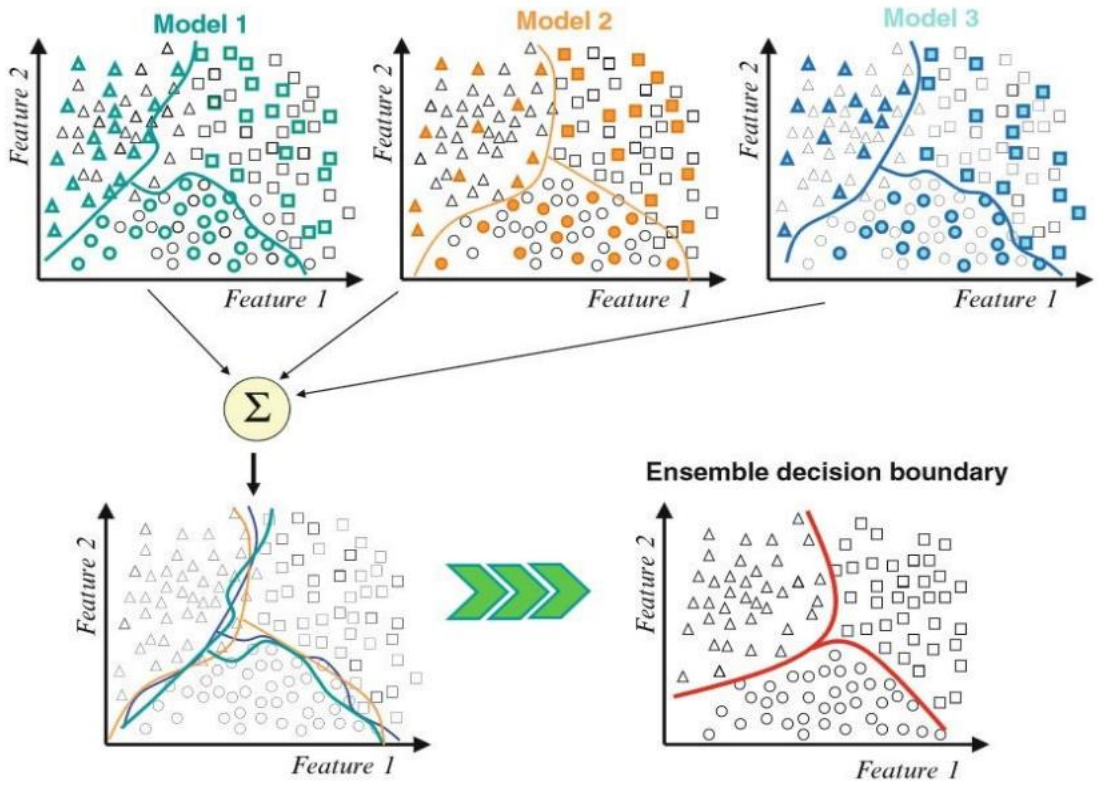

Figure 19: An illustration of a random forest in a two-dimensional space with three target labels. The votes from three decision trees are combined into a single model. For each tree, a coloured data point depicts a bootstrapped input vector (Polikar, 2012, p. 3)

In addition, Breiman claimed (2001) that Random Forest has been testified to be impactful in handling overfitting, a common complexity in ML, by averaging the predictions of several trees and considering only a subdivision of features for splitting. This helps to minimize the challenge of overfitting as well as optimizes the generalization execution of the model in unseen data, making it credible for forecasting cryptocurrency prices in various market situations, described in a study conducted by Giudici, Milne & Vinogradov (2020).

However, Random Forest has been acknowledged as an essential technique for cryptocurrency price forecasting due to its inherent strengths. Firstly, Random Forest can

manage high-dimensional data with difficult interactions among features, making it appropriate for analyzing cryptocurrency data that often includes multiple variables and intricate relationships (Makridakis, Spiliotis & Assimakopoulos, 2018). Secondly, Random Forest is resilient to noisy and missing data, which are very usual in cryptocurrency markets that are characterized by high volatility and limited availability of historical data (Breiman, 2001). Lastly, Random Forest is capable to capture non-linear patterns and identifying outliers, which are crucial considerations in cryptocurrency price forecasting as these markets are well-known for their non-linear dynamics and occasional extreme price movements (Makridakis et al., 2018).

3.4.1.4 SVM

Support vector machines (SVM's) are a powerful and flexible class of supervised learning algorithms for both classification and regression problems (Javatpoint, 2021). According to Kelley (2023), Support Vector Machines (SVMs) play a crucial role in the prediction of cryptocurrency prices, due to their inherent abilities in handling complex datasets. SVM is a supervised learning model that is primarily utilized in regression and classification problems. The algorithm is capable of constructing a hyperplane in an N-dimensional space (where N is the number of features) which uniquely categorizes the data points (Burges, 1998).

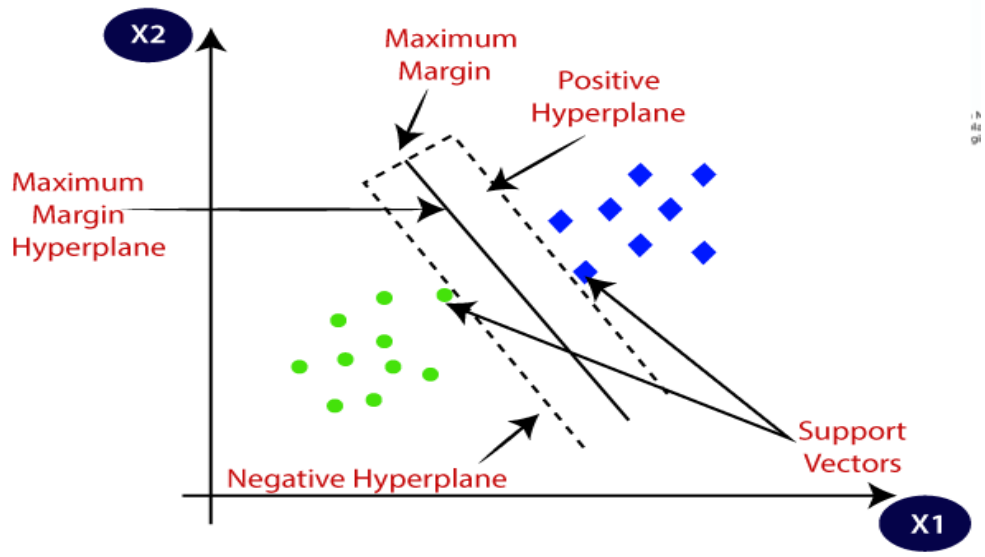

Figure 20: Description of SVM (Javatpoint, 2021)

The key advantage of SVM in the prediction of cryptocurrency prices lies in its proficiency in managing high dimensional data and its ability to utilize the kernel trick. By using different kernel functions, SVM can handle both linear and non-linear data, providing an opportunity to create optimal decision boundaries. This ability becomes exceedingly useful in highly volatile markets, such as those seen in cryptocurrencies (MathWorks, 2023).

Moreover, SVM has shown its resilience against overfitting, especially in situations where the number of dimensions is greater than the number of samples. This scenario is commonly seen in financial time series data like sequences of cryptocurrency prices (Kelley, 2023b). The model's resistance to overfitting significantly boosts the accuracy and reliability of its predictions, making SVM a valuable tool in cryptocurrency price forecasting. Another study conducted by Shah et al. (2018) provided a striking demonstration of SVM's effectiveness in predicting cryptocurrency prices. The SVM model outperformed multiple other machine learning models in predicting Bitcoin prices.

In summary, the SVM's capability to manage high dimensional data, its resilience to overfitting, and its proven success in prior studies highlight its importance in the domain of cryptocurrency price prediction. However, despite these benefits, it remains essential to continuously evaluate its performance across different cryptocurrencies and time resolutions, as the ideal model may change depending on these variables.

3.4.2 Data analysis tools

Data analysis tools are crucial for extracting meaningful insights from large and complex datasets in the field of machine learning. These tools enable scholars as well as practitioners to process, analyze, and visualize data in order to discover patterns, trends, and relationships that can inform ML algorithms and models. Various data analysis tools are available, such as Python libraries like NumPy, Pandas, and Scikit-learn, as well as R packages like dplyr as well as caret, which provide extensive functionality for data manipulation, transformation, and statistical analysis (Kelley, 2023).

Kelley (2023) continued that data analysis tools allow for tasks such as data cleaning, feature engineering, exploratory data analysis, and model evaluation, which are critical steps in the

machine learning workflow. Furthermore, Islam (2020, p. 12) added that visualization tools such as Matplotlib, seaborn and ggplot2 enable the graphical presentation of data, aiding in the interpretation and communication of findings. Overall, data analysis tools for machine learning play a critical role in the procedure and analysis of data, enabling users to derive meaningful insights and make informed decisions in the development and deployment of ML models (Islam, 2020, pp. 13 – 14).

In the present study, Python 3.9 was employed for data acquisition, processing, and analysis. The NumPy and Pandas libraries were utilized for data manipulation and feature engineering, respectively. Machine learning algorithms were developed and trained using the Scikit-learn library. All computations were performed on a laptop running Windows 10 and equipped with an Intel Core i5 processor.

4. Empirical Results

When it comes to making predictions using machine learning algorithms, it is imperative to ensure that the predictive power remains consistent across diverse products and time periods. Moreover, examining the algorithm's capability to generalize to new data is crucial for ensuring robustness (Akyildirim et al., 2021). To verify the reliability of the findings, this research employed a range of statistical tests and developed four distinct machine learning algorithms, each tailored for a different cryptocurrency. These algorithms were equally weighted for comparative analysis. The selected datasets (cryptocurrencies) enabled the researcher of this thesis to assess the predictive capabilities of various machine learning models and to verify their ability to generalize to new data. The subsequent chapter will analyze the findings from these four models. Additionally, this chapter will address the remaining research questions of this study, namely, what is the most effective machine learning model, and whether the same model can be generalized for predicting prices across different cryptocurrencies.

4.1 Logistic regression result analysis.

A Logistic Regression (LR) algorithm was developed to analyze the model's performance across various cryptocurrencies. The subsequent sub-chapter will examine the algorithm's findings and its efficacy in predicting the prices of four distinct cryptocurrencies.

4.1.1 LR model analysis for daily Bitcoin price prediction

The hyperparameter tuning process returned the optimal parameters for the LR algorithm, which were a regularization strength (C) of 0.001, a class weight of "balanced" (meaning classes are automatically adjusted to be equal), and using the "l2" penalty (which indicates Ridge regularization). These parameters achieved an average cross-validated accuracy of 82% on the training set.

The results of the analysis indicate an overall balanced performance of the logistic regression algorithm with an accuracy, precision, recall, and F1 score of 0.86. This suggests that the model was able to correctly classify 86% of the samples in the test set, while only

misclassifying 14% of them. The confusion matrix shows that out of the 146 total samples in the test set, 67 were true positives, 59 were true negatives, 10 were false positives, and 10 were false negatives. The model performed slightly better in identifying true positives and true negatives than in identifying false positives and false negatives.

Based on the results obtained, it can be concluded that the logistic regression model has a good performance in classifying the samples in the test set. However, the slightly lower performance in identifying false positives and false negatives suggests that further improvement in the model could be achieved. Therefore, it is recommended that additional data be collected to improve the model's training and that alternative classification algorithms be explored to determine whether better performance can be achieved. Overall, the current model provides a good basis for further research and can be used to classify similar samples with a high level of accuracy.

	Date	Price	Open	High	Low	Vol.	Change %
0	Jan 01, 2022	47,738.0	46,217.5	47,917.6	46,217.5	31.24K	3.29%
1	Dec 31, 2021	46,219.5	47,123.3	48,553.9	45,693.6	58.18K	-1.92%
2	Dec 30, 2021	47,123.3	46,470.7	47,901.4	46,003.0	60.96K	1.42%
3	Dec 29, 2021	46,461.7	47,548.4	48,121.7	46,127.8	63.92K	-2.28%
4	Dec 28, 2021	47,545.2	50,703.4	50,703.8	47,345.7	74.39K	-6.18%

```
Date        object
Price       object
Open        object
High        object
Low         object
Vol.        object
Change %    object
dtype: object
```

Figure 21: Bitcoin dataset overview

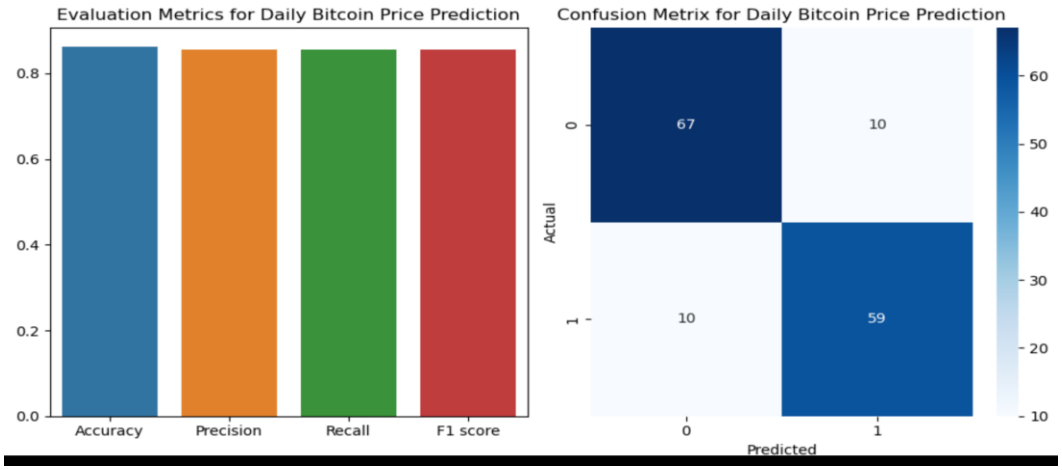

Figure 22: LR model evaluation and confusion metrics for daily Bitcoin price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Bitcoin	LR	0.86	0.86	0.86	0.86	0.82

Table 3: Daily Bitcoin price prediction report for logistic regression

4.1.2 LR model analysis for hourly Bitcoin price prediction

The model selection process was carried out using GridSearchCV, a method that performs exhaustive search over specified parameter values for an estimator. The hyperparameters "C", "penalty", and "class_weight" were tuned in this process. The average cross-validation accuracy achieved during this hyperparameter tuning phase was approximately 0.6, which implies that on average, the model was correct 60% of the time during the cross-validation phase.

The optimal hyperparameters found were "C" = 0.001, "class_weight" = None, and "penalty" = "l2". "C" is the inverse of regularization strength, with smaller values specifying stronger regularization. "Class_weight" = None indicates that all classes are given equal importance during model training, and "penalty" = "l2" implies the use of L2 or Ridge regularization in the logistic regression model.

The Logistic Regression model for hourly Bitcoin price prediction achieved a high accuracy score of 0.98, indicating that the model was able to classify 98% of the test instances correctly. The high precision score of 0.99 suggests that the model correctly identified 99% of the positive instances, while the recall score of 0.97 indicates that the model was able to identify 97% of the actual positive instances in the test set. The F1 score of 0.98, which is the harmonic mean of precision and recall, indicates that the model's overall performance was excellent. On the other hand, confusion matrix shows that the model correctly classified 3313 instances as negative and 3218 instances as positive. It generated 23 false negative instances and 97 false positive instances. These results suggest that the Logistic Regression model has high accuracy, precision, recall, and F1 score for hourly Bitcoin price prediction.

In conclusion, the results suggest that the Logistic Regression model is highly effective for hourly Bitcoin price prediction. The model achieved high accuracy, precision, recall, and F1 score, indicating that it is capable of accurately identifying positive instances while minimizing false positive and false negative instances. These results have important implications for decision-making in the cryptocurrency market and can be used to inform trading strategies and investment decisions. However, it is important to continue monitoring the model's performance over time and to consider the limitations and potential biases of the model in the specific context in which it will be used.

```
         unix                 date   symbol      open      high       low  \
0  1646092800  2022-03-01 00:00:00  BTC/USD  43221.71  43626.49  43185.48
1  1646089200  2022-02-28 23:00:00  BTC/USD  43085.30  43364.81  42892.37
2  1646085600  2022-02-28 22:00:00  BTC/USD  41657.23  44256.08  41650.29
3  1646082000  2022-02-28 21:00:00  BTC/USD  41917.09  41917.09  41542.60
4  1646078400  2022-02-28 20:00:00  BTC/USD  41361.99  41971.00  41284.11

      close  Volume BTC    Volume USD
0  43312.27   52.056320  2.254677e+06
1  43178.98  106.816103  4.612210e+06
2  42907.32  527.540571  2.263535e+07
3  41659.53   69.751680  2.905822e+06
4  41914.97  247.151654  1.035935e+07
unix            int64
date           object
symbol         object
open          float64
high          float64
low           float64
close         float64
Volume BTC    float64
Volume USD    float64
dtype: object
```

Figure 23: BTC-Hourly dataset overview

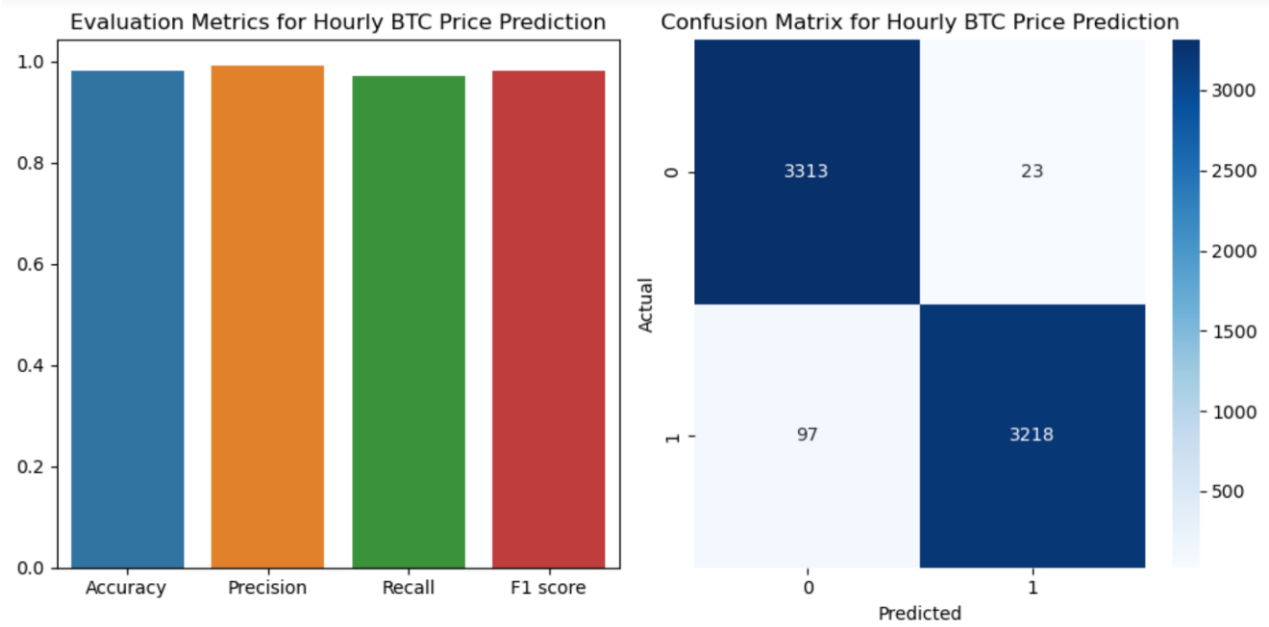

Figure 24: LR model evaluation and confusion metrics for hourly Bitcoin price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Bitcoin	LR	0.99	0.97	0.98	0.98	0.60

Table 4: Hourly Bitcoin price prediction report for logistic regression

4.1.3 LR model analysis for daily Ethereum price prediction

In this research study, a logistic regression algorithm has been employed to forecast the Ethereum price movement direction using the historical price and trading volume data. Prior to model building, the issue of class imbalance in the target variable was addressed by oversampling the minority class. The model has been further optimized by tuning its hyperparameters through grid search and cross-validation. The grid search revealed the optimal hyperparameters for the logistic regression model to be a "C" value, which controls the inverse of regularization strength, of approximately 3792.69 and a penalty term of "l2". The "l2" penalty, also known as Ridge Regularization, reduces the squared magnitude of the coefficients, thereby assisting to prevent overfitting of the model.

In terms of the model's predictive performance, it obtained an average cross-validated accuracy of 0.77 during the training phase. This performance metric indicates that the model correctly predicted the direction of Ethereum price movement for 77% of the instances in the validation sets during cross-validation. However, when the final model was tested on unseen data, it yielded an accuracy score of 0.85, implying that it correctly classified the price movement direction for 85% of the test instances. The precision score, which measures the proportion of true positives out of all positive predictions, was found to be 0.77. This indicates that 77% of the instances that the model forecasted as price increases were indeed price increases.

The model exhibited an exceptional recall score of 0.99, meaning that it successfully identified 99% of all actual price increases. The F1 score, which provides a balanced measure of the model's precision and recall, was computed to be 0.86, demonstrating a harmonious balance between the two aforementioned performance metrics. The confusion matrix, on the other hand, further provides insights into the model's performance. Of all the test instances, the model correctly identified 55 price decreases and 69 price increases, while incorrectly predicting 21 price decreases as price increases and 1 price increase as a price decrease.

In summary, the logistic regression model, with its tuned hyperparameters, presented a reasonably high degree of predictive accuracy for Ethereum price movements. The model's strong recall score further assures its reliability in capturing potential price surges, making it a potentially beneficial tool for traders and investors. Nonetheless, the precision score points to a certain level of false alarms (i.e., predicting price increases that did not occur), which should be taken into consideration when applying the model's predictions for decision-making purposes.

```
          Date     Price      Open      High       Low     Vol. Change %
0  Jan 01, 2022  3,765.67  3,677.69  3,775.20  3,675.75  239.54K    2.39%
1  Dec 31, 2021  3,677.85  3,709.38  3,812.01  3,623.74  405.65K   -0.85%
2  Dec 30, 2021  3,709.57  3,629.33  3,767.93  3,589.64  355.72K    2.25%
3  Dec 29, 2021  3,627.93  3,792.95  3,825.97  3,607.20  456.03K   -4.32%
4  Dec 28, 2021  3,791.69  4,036.86  4,036.86  3,760.86  527.23K   -6.06%
Date         object
Price        object
Open         object
High         object
Low          object
Vol.         object
Change %     object
dtype: object
```

Figure 25: Ethereum dataset overview

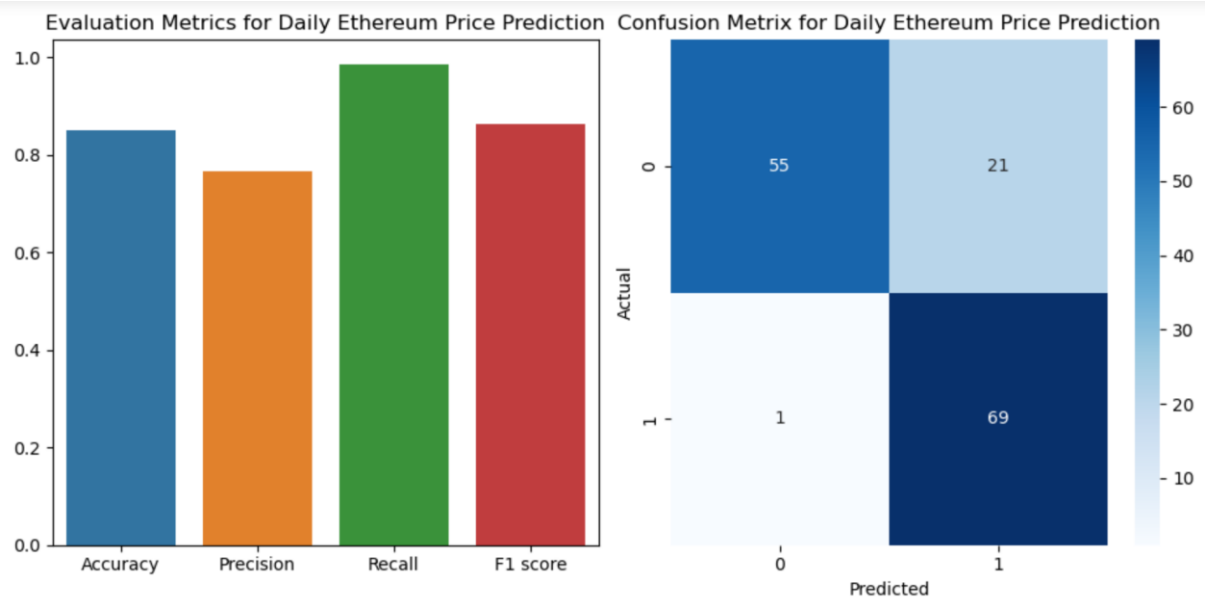

Figure 26: LR model evaluation and confusion metrics for daily Ethereum price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Ethereum	LR	0.77	0.99	0.86	0.85	0.77

Table 5: Daily Ethereum price prediction report for logistic regression

4.1.4 LR model analysis for daily Cardano price prediction

The algorithm selection procedure was carried out employing GridSearchCV, a technique that performs exhaustive search over specified parameter values for an estimator. The hyperparameters "C", and "penalty" were tuned in this process. The average cross-validation accuracy achieved during this hyperparameter tuning phase was approximately 0.51, which implies that on average, the model was correct 51% of the time during the cross-validation phase.

The results of the analysis show a moderate performance of the model with an accuracy score of 0.49, indicating that the model correctly classified only 49% of the samples in the test set. The precision score of 0.54 indicates that 54% of the samples classified as positive by the model were truly positive. Similarly, the recall score of 0.69 indicates that 69% of the true positive samples in the test set were correctly identified by the model. The F1 score of 0.61, which is the harmonic mean of the precision and recall scores, suggests that the model's performance is suboptimal. The confusion matrix shows that out of the 111 total samples in the test set, 10 were true positives, 44 were true negatives, 20 were false positives, and 37 were false negatives. The model performed better in identifying true negatives than in identifying true positives, with a higher number of false negatives than false positives.

Based on the results obtained, it can be concluded that the current LR algorithm for Cardano dataset has a moderate performance in classifying the samples in the test set. However, the relatively low precision and recall scores recommend that the model could be optimized by retraining on a larger and more balanced dataset. Additionally, alternative classification algorithms and feature engineering techniques should be explored to enhance the model's performance. Overall, the current model suggests a starting point for further research and improvement, but caution should be exercised in its application for classification purposes.

```
           Date   Price    Open    High     Low     Vol. Change %
0  Jan 01, 2022  1.3789  1.3088  1.3789  1.3067  104.96M    5.34%
1  Dec 31, 2021  1.3090  1.3580  1.3819  1.2819  184.18M   -3.67%
2  Dec 30, 2021  1.3589  1.3307  1.3778  1.2992  195.11M    2.14%
3  Dec 29, 2021  1.3304  1.3998  1.4368  1.3247  229.60M   -4.85%
4  Dec 28, 2021  1.3982  1.5142  1.5381  1.3792  347.26M   -7.65%
Date          object
Price        float64
Open         float64
High         float64
Low          float64
Vol.          object
Change %      object
dtype: object
```

Figure 27: Cardano dataset overview for daily price prediction

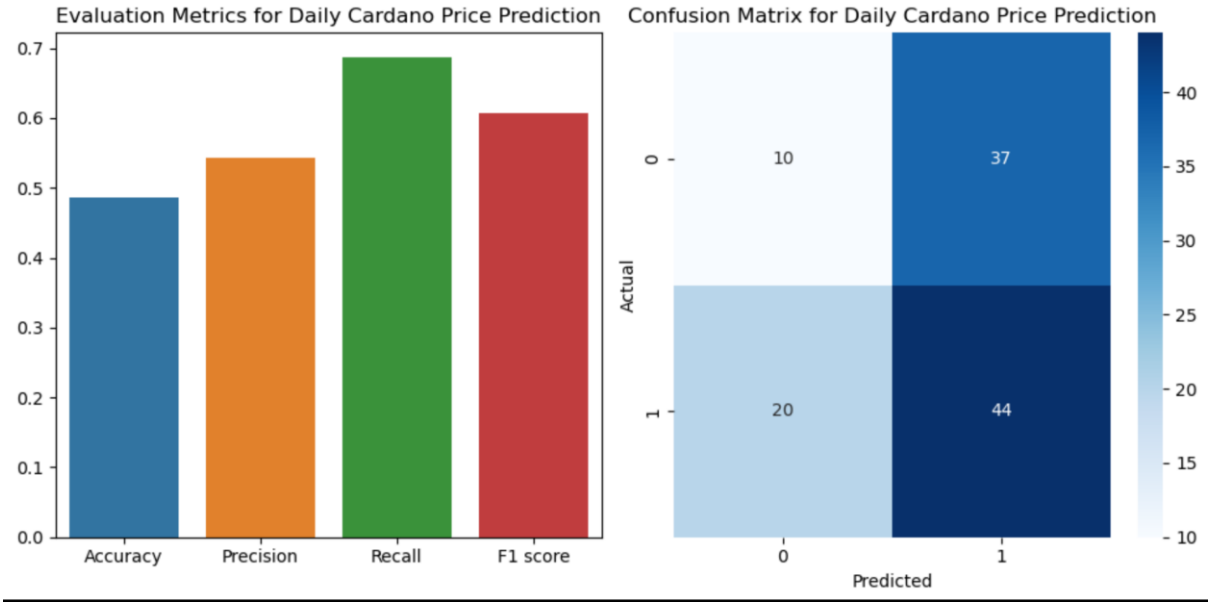

Figure 28: LR model evaluation and confusion metrics for daily Cardano price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Cardano	LR	0.54	0.69	0.61	0.49	0.51

Table 6: Daily Cardano price prediction report for logistic regression

4.1.5 LR Model analysis for daily Solana price prediction

In the present investigation, researcher sought to develop a method for forecasting the daily movements in the price of Solana cryptocurrency using logistic regression. For the model optimization, GridSearchCV was deployed for hyperparameter tuning.

The cross-validated accuracy of the model was found to be approximately 0.47. This indicates that the model correctly predicts the daily price change direction about 47% of the time when assessed across various subsets of the data. This level of performance, while not extremely high, is notable given the inherent unpredictability of cryptocurrency markets. The grid search recognized the best hyperparameters for the logistic regression model as a C parameter (inverse of regularization strength) of 0.0001 and a penalty of "l2". An "l2" penalty refers to L2 regularization, which can help prevent overfitting by encouraging smaller weights.

Applying these parameters, the model yielded an accuracy of 0.50 on the test data. This signifies that the model correctly forecasted the direction of price change in 50% of the cases in the test set. The precision of the model, which measures the proportion of true positive predictions among all positive predictions, was recorded as 0.47. This indicates that when the model predicted a positive price change, it was correct 47% of the time. Similarly, the recall score of 0.90 indicates that the model was able to correctly identify 90% of the true positive samples, while misclassifying 10% of them as false negatives. The F1 score of 0.62, which is the harmonic mean of the precision and recall scores, recommends that the model's overall performance is suboptimal. The confusion matrix shows that out of the 22 total samples in the test set, 2 were true positives, 9 were true negatives, 10 were false positives, and 1 was a false negative. The model performed better in identifying true negatives than in identifying true positives, with a higher number of false positives than false negatives.

In conclusion, while the model demonstrates certain strengths, such as a high recall, there is evident room for improvement, particularly in its precision and overall accuracy. It is advisable to explore other modelling techniques or additional features that could enhance the prediction capability for this task.

```
           Date    Price     Open     High      Low   Vol. Change %
0  Jan 01, 2022  179.068  169.985  179.176  169.985  1.30M    5.34%
1  Dec 31, 2021  169.985  172.509  177.604  167.773  2.05M   -1.45%
2  Dec 30, 2021  172.482  170.602  175.697  168.382  1.86M    1.09%
3  Dec 29, 2021  170.626  177.191  180.581  170.434  2.59M   -3.68%
4  Dec 28, 2021  177.151  195.602  195.602  176.793  3.64M   -9.45%
Date          object
Price        float64
Open         float64
High         float64
Low          float64
Vol.          object
Change %      object
dtype: object
```

Figure 29: Solana dataset overview for daily price prediction

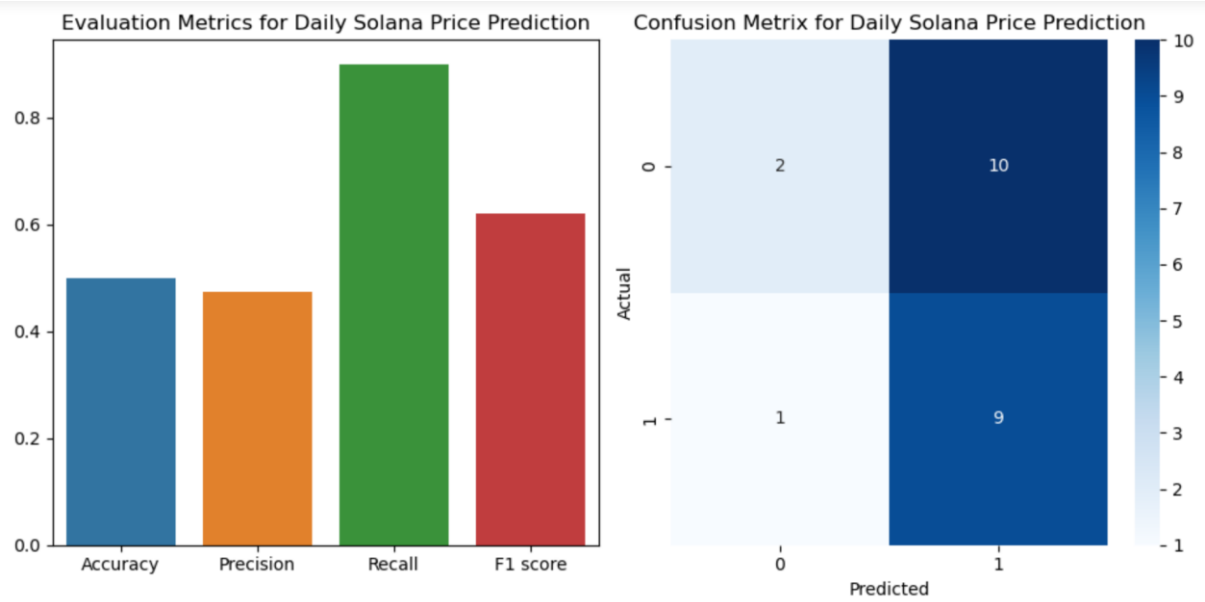

Figure 30: LR model evaluation and confusion metrics for daily Solana price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Solana	LR	0.47	0.90	0.62	0.50	0.47

Table 7: Daily Solana price prediction report for logistic regression

4.2 Decision tree result analysis

The following sub-chapter will discuss the findings and performance of the decision tree model in predicting price movements for four distinct cryptocurrencies: Bitcoin, Ethereum, Cardano, and Solana.

4.2.1 Decision tree model analysis for daily Bitcoin price prediction.

During this research, a Decision Tree Classifier was employed to predict daily fluctuations in the Bitcoin price, specifically, whether it would increase or decrease, based on a range of features including the opening price, highest price, lowest price, and transaction volume. Prior to training the classifier, class imbalance in the target variable was mitigated using the RandomOverSampler technique.

Hyperparameter tuning was carried out applying GridSearchCV, which implements a fit and score method, with the best parameters selected based on a cross-validation method. The parameters under investigation included the maximum depth of the tree (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf). The grid search across these parameters yielded an optimal model with a max_depth of 20, min_samples_leaf of 1, and min_samples_split of 5. This configuration resulted in the highest average cross-validated accuracy score of 0.59 during the training phase.

The model's performance was further evaluated on the test set, revealing an overall accuracy of 0.53. Precision, a measure of the model's capability to correctly predict a price increase when it occurs, stood at 0.5. Recall, which assesses the model's capacity to identify all actual price increases, reached a higher score of 0.83. This indicates that the model was relatively more successful at capturing all instances of price increase, albeit at the cost of misclassifying some non-increase instances as increases (as evinced by the lower precision). The F1 Score, a harmonic mean of precision and recall, was recorded as 0.63.

The confusion matrix presented a more detailed view of the model's performance, indicating the count of true positives (57), true negatives (21), false positives (56), and false negatives

(12). This further corroborates the interpretation that the model tends to predict more instances of price increase, including both correct and incorrect predictions, thereby leading to higher recall and lower precision.

Given these outcomes, there is room for further optimization of the model, and other predictive algorithms may be considered in future to seek improved performance.

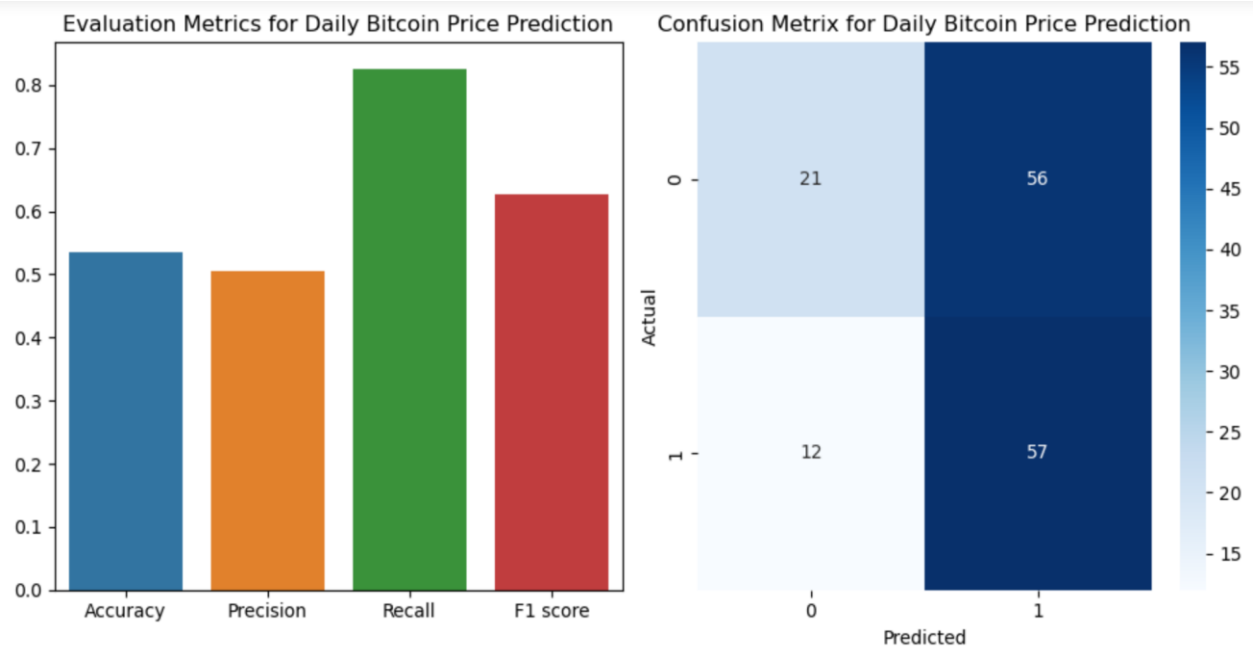

Figure 31: Decision tree model evaluation and confusion metrics for daily Bitcoin price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Bitcoin	Decision tree	0.50	0.83	0.63	0.53	0.59

Table 8: Daily Bitcoin price prediction report for Decision tree

4.2.2 Decision tree model analysis for hourly Bitcoin price prediction.

The computational analysis of the model's performance through a quintuple (five-fold) cross-validation process yielded an average validation score of approximately 0.755. The

individual fold scores were observed to be 0.770, 0.778, 0.765, 0.739, and 0.724. This indicates a moderate level of consistency in the algorithm's performance across various partitions of the training set. The generalizability of the model on unseen data is presented as an accuracy of 0.71, implying that the method correctly identifies the directional shift in Bitcoin's hourly price about 71% of the time. The precision of the algorithm, reflecting its ability to correctly identify true positive outcomes relative to the total predicted positives, is found to be 0.77, which signifies that a positive prediction of price increase by the model can be deemed reliable about 77% of the time.

However, the recall or sensitivity of the model, which represents the proportion of actual positive outcomes that were correctly identified, stands at 0.58. This means that the decision tree approach correctly detects 58% of all instances where the Bitcoin price increases. The F1 score, which is a harmonic mean of precision and recall and thus provides a singular measure of model performance, is reported as 0.66. This score suggests a reasonable balance between precision and recall in the model's performance.

The confusion matrix for the algorithm, which is a structured summary of prediction results on a classification problem, is presented as follows: [[2769 567], [1384 1931]]. In the context of this matrix, the method has correctly classified 2769 instances of price decrease (true negatives) and 1931 instances of price increase (true positives). Conversely, the model incorrectly classified 567 instances of price decrease as increases (false positives), and 1384 instances of price increase as decreases (false negatives).

Considering these performance metrics, it can be inferred that while the algorithm has demonstrated a fair degree of predictive accuracy, there exists significant potential for enhancement, specifically in terms of improving the recall and reducing the number of false negatives. Suggested avenues for improvement might include the exploration and incorporation of additional or different predictive variables, the application of more sophisticated or diverse algorithm architectures (e.g., ensemble methods or deep learning models), and the potential usage of advanced techniques to address the class imbalance issue evident in the dataset, such as oversampling, undersampling, or Synthetic Minority Over-sampling Technique (SMOTE). Moreover, an extensive exploration of the model hyperparameter space could potentially lead to optimized model performance.

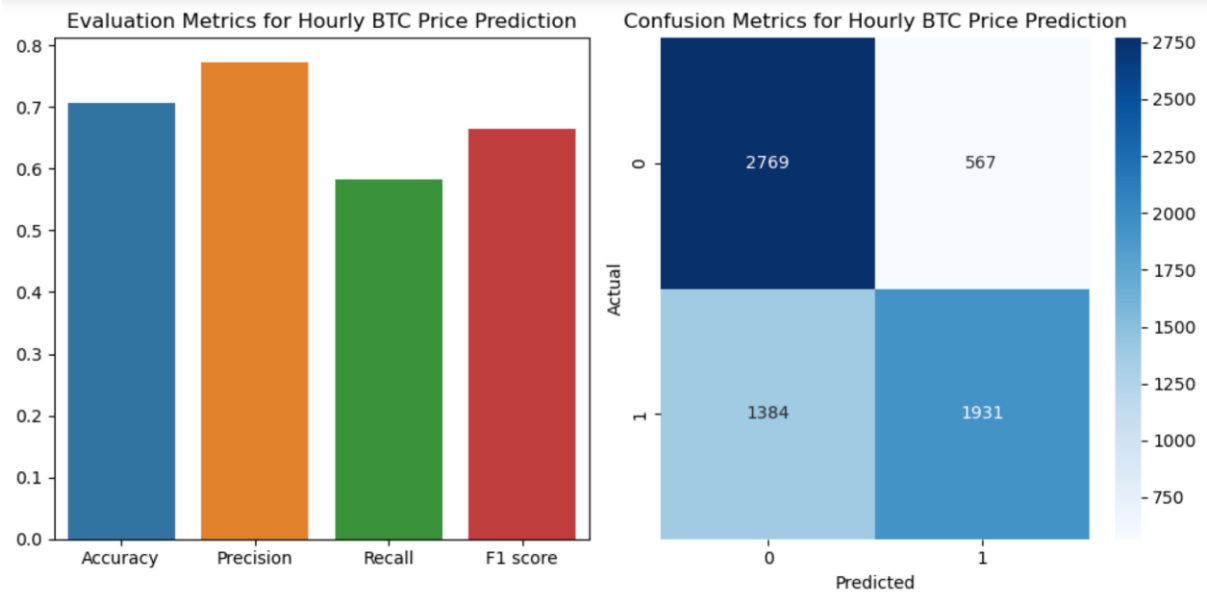

Figure 32: Decision tree model evaluation and confusion metrics for hourly Bitcoin price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Bitcoin	Decision tree	0.77	0.58	0.66	0.71	0.755

Table 9: Hourly Bitcoin price prediction report for DT (Decision Tree)

4.2.3 Decision tree model analysis for daily Ethereum price prediction.

The evaluation of predictive method, a decision tree classifier with hyperparameters fine-tuned via grid search, reveals several pertinent findings. The optimal hyperparameters identified were a maximum depth of 16, a minimum samples per leaf of 1, and a minimum samples to split of 2. An average cross-validated accuracy of 0.55 was achieved in the model's training phase, indicating moderate performance during the model's validation against unseen subsets of the training data. Upon applying this model to the independent test dataset, 0.50 accuracy was obtained. This suggests that the model correctly predicted the direction of Ethereum's price movement, either increase or decrease, for half of the instances in the test set.

Precision, which assesses the capability of the proposed model to appropriately identify positive instances among all instances it labeled as positive, is measured at 0.48. This figure endorses that the algorithm correctly identified a price increase for nearly 48% of all instances it labeled as price increases. The algorithm's recall or sensitivity is 0.67. This denotes its capacity to recognize true positive instances from all actual positive instances. Thus, it managed to detect 67% of all actual price increases. The F1 score, a harmonic mean of precision and recall, stands at 0.56, showing that a balance between precision and recall has been achieved in our predictive solution. On the other hand, the confusion matrix, a tool for visualizing the model's performance, shows that the algorithm forecasted 26 true negatives and 47 true positives correctly. However, it also mistakenly identified 23 instances as price increases (false positives) when they were not and classified 50 instances as price decreases (false negatives) when they were not.

Given these results, there are several recommendations for future steps. Firstly, although the decision tree model is an interpretable and straightforward algorithm, it may not be the most appropriate for predicting Ethereum's price direction given its performance metrics. Thus, other more advanced machine learning algorithms such as Random Forests or Gradient Boosting could be considered. Secondly, feature engineering and selection could also be further explored to identify the most predictive features for Ethereum's price movement. For example, technical indicators from financial analysis, past price patterns, and other market data such as trading volume could be incorporated into the model. Finally, the imbalanced nature of the target variable could be addressed further. The model's relatively lower precision could be attributed to the imbalance in the classes of the target variable. Therefore, other ML algorithms for dealing with imbalances such as SMOTE or ADASYN could be explored.

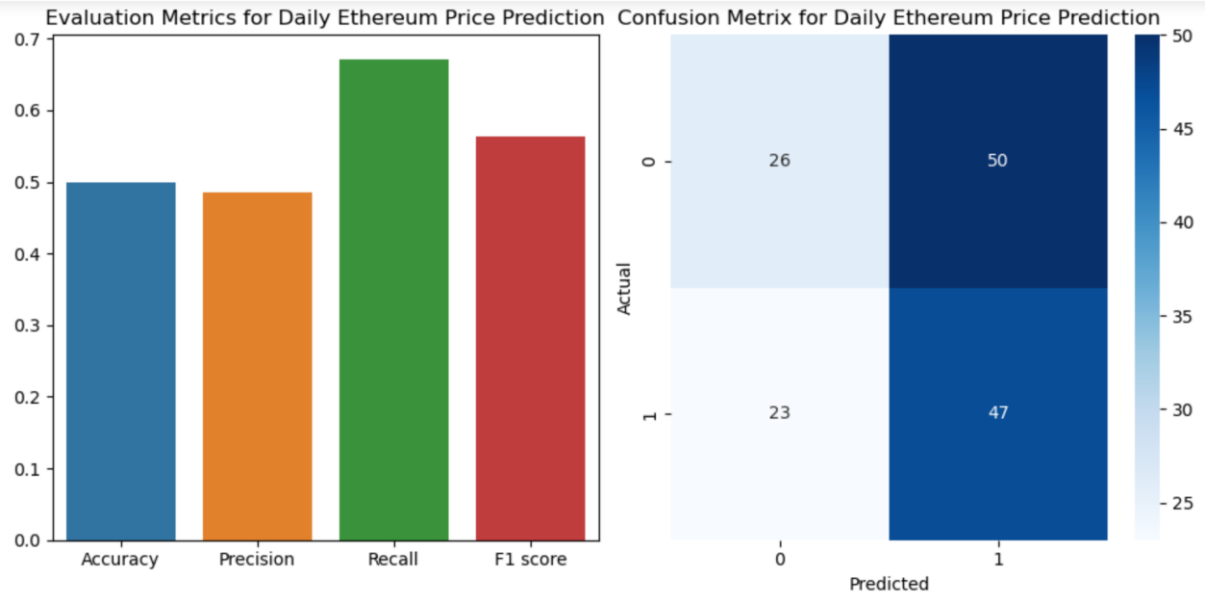

Figure 33: Decision tree model evaluation and confusion metrics for daily Ethereum price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Ethereum	Decision tree	0.48	0.67	0.56	0.50	0.55

Table 10: Daily Ethereum price prediction report for DT (Decision Tree)

4.2.4 Decision tree model analysis for daily Cardano price prediction.

An average cross-validated accuracy of 0.61 was obtained in the algorithm's training phase, indicating moderate performance during the model's validation against unseen subsets of the training data. Upon applying this technique to the independent test dataset, 0.60 accuracy was achieved. The model also had a precision of 0.6, indicating that 60% of the predicted positive values were correct. The recall of 0.94 implies that the model correctly identified 94% of the actual positive values. The F1 score, which is the harmonic mean of precision and recall, was 0.73, indicating a balance between precision and recall. According to the confusion matrix, the model correctly predicted 80 out of 85 positive values and correctly identified 7 out of 61 negative values. However, it misclassified 54 negative values as positive and 5 positive values as negative. These results suggest that the model has a relatively high recall, but low precision,

indicating that it is better at identifying the actual positive values but is prone to a high false positive rate.

Overall, the decision tree technique has shown moderate performance in predicting daily Cardano price movements. While it had a high recall rate, its precision rate was relatively low, indicating that the model has room for improvement. Therefore, further refinement of the algorithm parameters and feature selection may be required to enhance its performance.

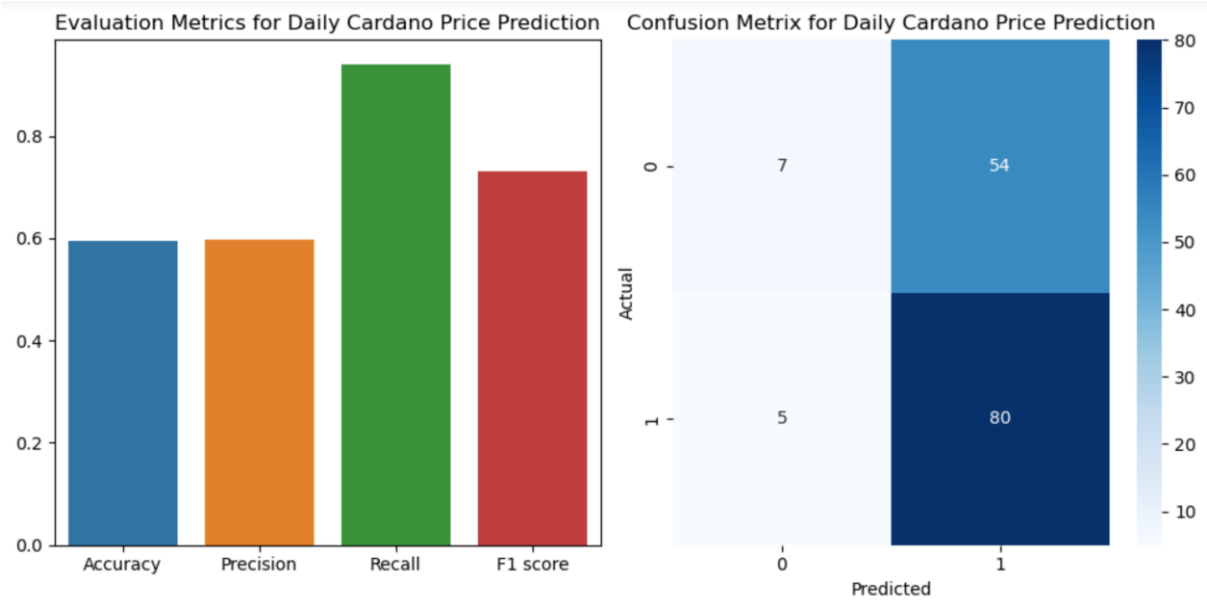

Figure 34: Decision tree model evaluation and confusion metrics for daily Cardano price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validated accuracy
Cardano	Decision tree	0.60	0.94	0.73	0.60	0.61

Table 11: Daily Cardano price prediction report for DT (Decision Tree)

4.2.5 Decision tree model analysis for daily Solana price prediction.

This research study explored the effectiveness of a decision tree model to predict price movements in Solana, a digital asset. An examination of the average cross-validated accuracy

shows an average value of 0.58, providing some evidence for the model's effectiveness at appropriately identifying price changes in the asset.

The hyperparameters for the algorithm that offered optimal outcomes were a max depth of 25, minimum samples per leaf of 3, and minimum samples per split of 2. Once these parameters were determined and the model was trained, it yielded an accuracy of 0.58 on the test data. This recommends that the technique was correct in its predictions of price increases or decreases approximately 58% of the time.

When considering the decision tree algorithm's precision, measured at 0.56, it can be determined that when the model forecasted a price increase, it was correct in 56% of cases. Besides, the model's recall, or sensitivity, was observed to be 0.78. This denotes the method's capability to correctly recognize true price increases out of all actual price increases, indicating that the model successfully detected 78% of actual price increases. Furthermore, the F1 score, a measure that encapsulates both precision and recall into a single metric, stands at 0.65. This suggests a relatively balanced performance between precision and recall, indicating that the model performed reasonably well on both fronts. On the other hand, the confusion matrix offers a more detailed look at the algorithm's performance. It shows that there were 19 true negatives and 40 true positives. Conversely, the decision tree model made 32 type I errors (false positives) and 11 type II errors (false negatives).

In terms of recommendations, despite reasonable performance in certain metrics, there is significant room for improvement in the algorithm's precision and overall accuracy. To this end, the use of more complex techniques, such as ensemble methods or neural networks, could be examined. Additionally, the implementation of more advanced feature engineering or the use of a larger, more comprehensive dataset may improve the algorithm's predictive abilities. Lastly, given the dynamic and volatile nature of digital asset markets, a focus on achieving robust out-of-sample performance is crucial for the practical utility of such predictive models.

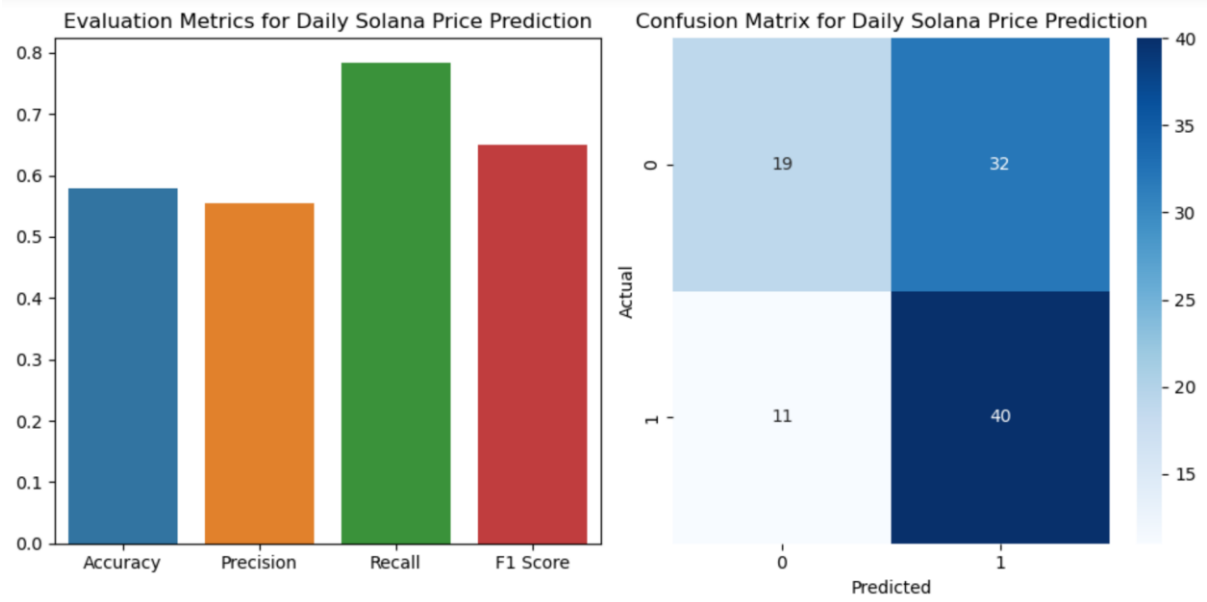

Figure 35: Decision tree model evaluation and confusion metrics for daily Solana price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Solana	Decision Tree	0.56	0.78	0.65	0.58	0.58

Table 12: Daily Solana price prediction report for DT (Decision Tree)

4.3 Random forest result analysis

The subsequent sub-section will discuss the results and performance of the Random Forest (RF) model in predicting the price movements of four distinct cryptocurrencies: Bitcoin, Ethereum, Cardano, and Solana.

4.3.1 RF model analysis for daily Bitcoin price prediction.

The research employs a Random Forest Classifier to generate predictive algorithms for Bitcoin price changes. The model's performance is critically assessed via several metrics including accuracy, precision, recall, and the F1 score. An added layer of robustness is provided

by implementing a cross-validation strategy. Cross-validation outcomes, based on five partitions or folds, indicate a range of scores: 0.603, 0.504, 0.536, 0.272, and 0.488. The average cross-validation score, an overall measure of the algorithm's performance across all folds, is 0.48. This suggests a moderate level of predictive accuracy; a perfect model would achieve a score of 1.

The RF method's accuracy for Bitcoin price forecasting, the proportion of correct predictions out of total predictions, stands at 0.58. While this signifies a correct prediction rate of more than half, it also demonstrates that a significant number of instances are incorrectly classified. Precision, a measure quantifying the number of true positive outcomes out of all positive predictions, is calculated as 0.53. This implies that when the algorithm predicts an increase in Bitcoin price, the forecast is correct 53% of the time. The method's recall, indicating the proportion of true positive results out of all actual positive instances, is 0.88. This recommends that the model is only capable to correctly identify 88% of all instances where Bitcoin price actually increased. The F1 score, or the harmonic mean of precision and recall, is 0.66. A high F1 score is indicative of a high-performing model. In this case, the model presents a relatively low F1 score. The confusion matrix, on the other end, reveals that the algorithm correctly forecasts 24 true negatives and 60 true positives, while misclassifying 8 instances as false negatives and 53 as false positives.

Overall, the performance of the algorithm signals the potential for enhancement. The foremost step could be hyperparameter tuning to refine the Random Forest Classifier's performance, involving adjustments to parameters such as the number of trees, maximum depth of trees, or splitting criteria. Secondly, the feature engineering process could be revisited to investigate whether the inclusion of additional features or modification of lag periods could better capture the patterns in Bitcoin price movements. Furthermore, the application of alternate oversampling techniques or different classification algorithms could provide improved outcomes. Pursuing these avenues is recommended to develop an optimized model.

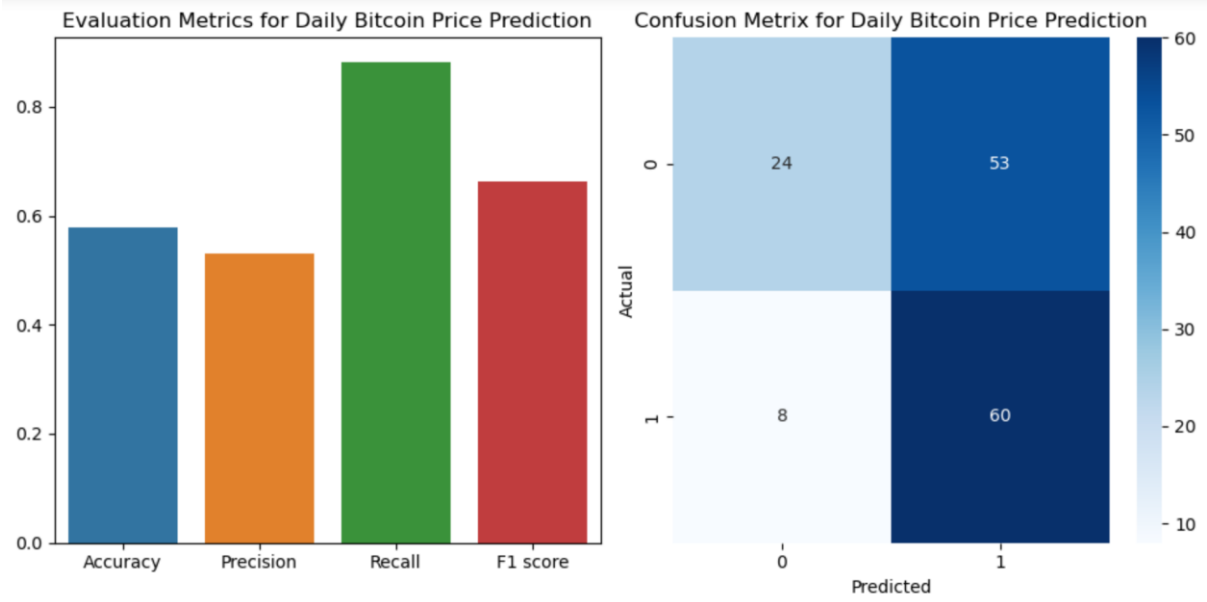

Figure 36: RF model evaluation and confusion metrics for daily Bitcoin price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Bitcoin	RF	0.53	0.88	0.66	0.58	0.48

Table 13: Daily Bitcoin price prediction report for RF

4.3.2 RF model analysis for hourly Bitcoin price prediction.

The empirical results analysis included an examination of Bitcoin hourly price movements, as modeled by a Random Forest Classifier. The robustness of this algorithm was evaluated through 5-fold cross-validation. Cross-validated scores, which provide a measure of the model's predictive performance across various subsets of the data, ranged from 0.66 to 0.74, with a mean of 0.68. This average cross-validation score suggests that the method, on average, accurately predicts hourly price changes about 68% of the time across multiple data partitions. It's an acceptable performance given the high volatility and unpredictability inherent in Bitcoin prices.

Furthermore, the algorithm's performance was evaluated on a testing set that had not been used during the training process. The model's accuracy, precision, recall, and F1 score were 0.63, 0.66, 0.52, and 0.58, respectively. These metrics indicate that the model correctly predicted the direction of the Bitcoin price change in 63% of cases in the testing dataset. However, of all positive forecasts made by the RF method, only 66% were actually positive, as evidenced by the precision score. The recall score reveals that the model was able to identify 52% of all actual positive instances. The F1 score, a balanced measure of precision and recall, stands at 0.58, showing a decent balance between precision and recall. The confusion matrix, a table that explains the performance of the model on the testing data, reveals that the algorithm correctly predicted 2460 negative instances and 1713 positive instances. However, the RF technique incorrectly predicted 876 actual negative instances as positive, and 1602 actual positive instances as negative.

The aforementioned outcomes imply that while the algorithm displays a satisfactory predictive performance, there is still room for improvement, particularly regarding recall. Future efforts should focus on enhancing this aspect to improve the overall model performance. This could be obtained by tuning the hyperparameters of the Random Forest technique, incorporating more relevant features, or considering the application of different machine learning methods. It is also suggested to consider other metrics, such as AUC-ROC, which are less sensitive to class imbalance and might provide additional insights into the model performance.

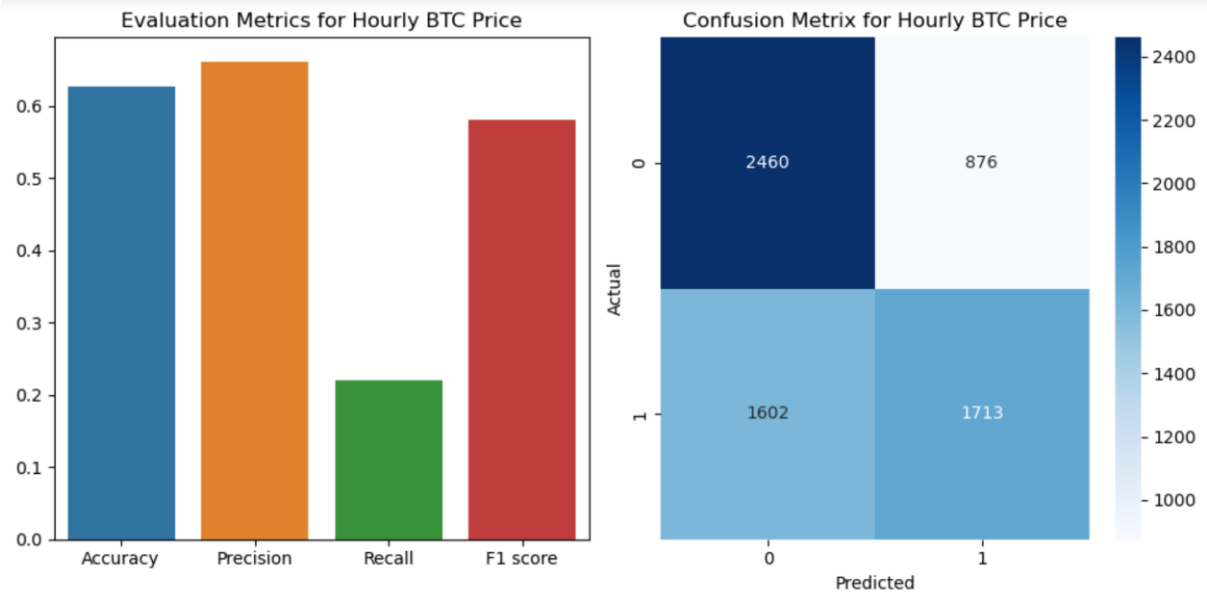

Figure 37: RF model evaluation and confusion metrics for hourly Bitcoin price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Bitcoin	RF	0.66	0.52	0.58	0.63	0.685

Table 14: Hourly Bitcoin price prediction report for RF

4.3.3 RF model analysis for daily Ethereum price prediction.

The outcomes derived from the RF algorithm underscore the variability and subtleties inherent in the predictive modelling process. Following are the inferences drawn from the evaluation metrics.

A set of cross-validation scores was achieved from the model: [0.48461538, 0.46923077, 0.4, 0.3255814, 0.54263566]. The average of these scores, which is approximately 0.444, provides an indication of the model's generalized performance across various subsets of the data. While this score is somewhat lower than ideal, it gives a more realistic evaluation of the algorithm's capacity to generalize unseen data compared to using the accuracy metric alone. The accuracy metric, which is the ratio of correct forecasts to total predictions, stands at 0.52.

This signifies that approximately 52% of the total predictions made by the algorithm are accurate. While this metric offers a straightforward interpretation, it may not provide a comprehensive view of the model's performance, especially in the case of imbalanced datasets.

The precision of the model, which gauges the accuracy of positive predictions, is at 0.5. This implies that when the technique forecasts a price increase, it is correct about 50% of the time. Besides, the algorithm exhibits a high recall (sensitivity) of 0.86, indicating that it correctly identifies 86% of the actual price increases. However, a high recall often comes at the cost of precision, and indeed, the precision is relatively low in this instance. Thus, the model shows a strong tendency to favor sensitivity over precision. Furthermore, the F1 score, which balances precision and recall, stands at 0.63. This metric provides a more holistic view of the technique's performance when dealing with imbalanced classes. Despite the relatively low precision, the high recall bolsters the F1 score. The confusion matrix, on the other hand, a specific table layout impactful for understanding the performance of the classification algorithm, represents a more detailed figure of the model's performance. The matrix reveals a higher number of false positives (60), suggesting that the model often incorrectly predicts a price increase.

To sum up, while the RF algorithm for Ethereum price prediction shows strong sensitivity, it lacks precision, resulting in many false positives. To optimize the predictive power and applicability of this algoritm, one might consider employing techniques such as feature selection or engineering, hyperparameter tuning, or leveraging a different, potentially more sophisticated, machine learning model. These adjustments could potentially lead to more robust and accurate predictions.

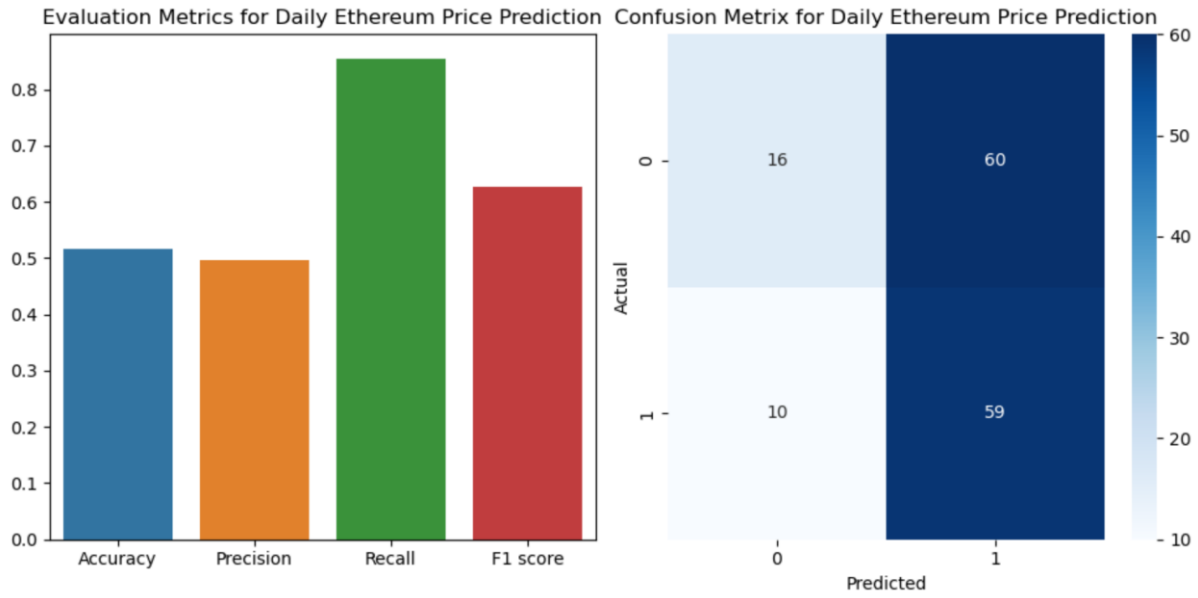

Figure 38: RF model evaluation and confusion metrics for daily Ethereum price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Ethereum	RF	0.50	0.86	0.63	0.52	0.444

Table 15: Daily Ethereum price prediction report for RF

4.3.4 RF model analysis for daily Cardano price prediction.

The cross-validation scores obtained from this algorithm are [0.47692308, 0.42307692, 0.41538462, 0.31007752, 0.55813953]. These values represent the algorithm's capability to forecast unseen data for five different subsets of the dataset. The mean of these scores, approximately 0.437, is used as a representative value to gauge the model's overall performance. While this mean score is not exceedingly high, it does exhibit a reliable insight into how well the model might perform on new, unseen data.

A key evaluation metric, accuracy, comes in at 0.58. This metric indicates that the RF algorithm correctly predicted the direction of Ethereum price changes around 58% of the time. Nonetheless, accuracy can be a misleading measure of model performance, particularly in

instances where the dataset might be imbalanced. Precision, which assesses the exactness of positive predictions made by the algorithm, is valued at 0.6. This means that when the model predicts a price increase, it is correct approximately 60% of the time. The model's recall score is 0.84. Recall measures the completeness of the positive predictions, indicating that the model correctly identified 84% of all actual price increases. It's worth noting that while a high recall score is typically desirable, it may also be indicative of an increased number of false positives, especially when coupled with a precision score that is comparatively lower. The model's F1 score, a balanced measure of precision and recall, is calculated at 0.7. This relatively high score can be attributed to the above-average recall score and suggests a more balanced performance of the model in terms of precision and recall. Further insights can be drawn from the confusion matrix, which displays a more granular view of the algorithm's performance. The matrix shows that the model obtained 14 true negatives and 71 true positives. Conversely, the algorithm also misclassified 47 false positives and 14 false negatives. This indicates a moderate level of both type I and type II errors, contributing to the precision and recall scores observed.

In summary, while this algorithm presents an acceptable level of performance, there are evident areas for improvement. The moderate precision score implies a number of false positives, i.e., the model may frequently predict price increases that do not occur. For applications where the cost of false positives is high, this model may prove less than optimal. Potential strategies for optimization could include tuning hyperparameters to improve precision, engineering new features that may contribute to the prediction power, or even exploring different machine learning algorithms that may offer improved performance. These enhancements could lead to the development of a more robust and accurate prediction model.

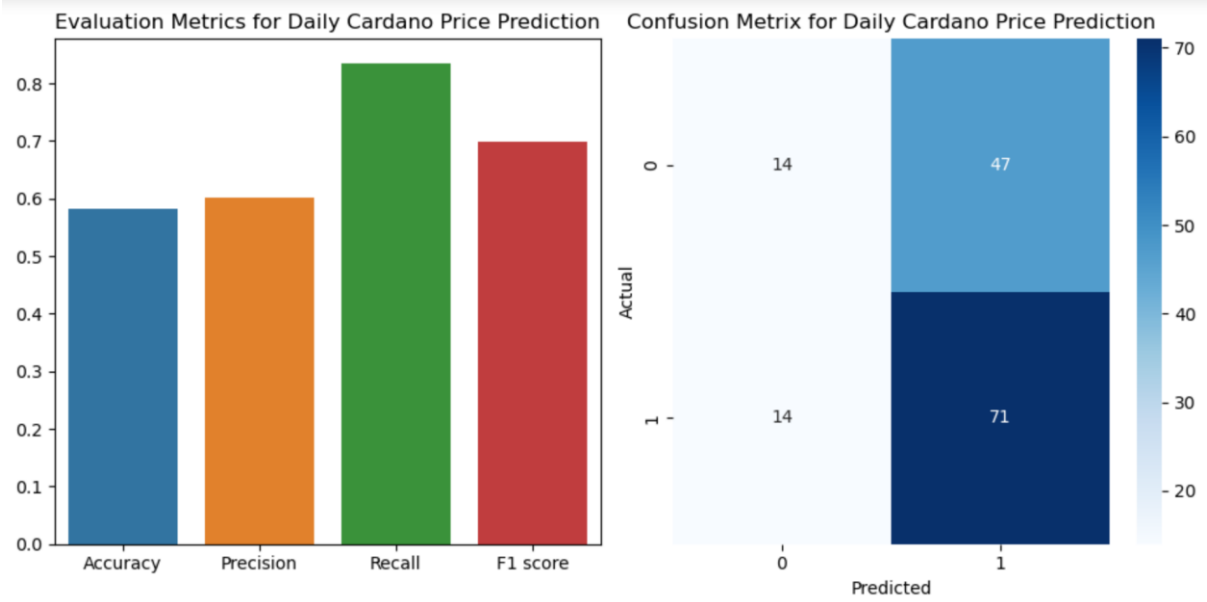

Figure 39: RF model evaluation and confusion metrics for daily Cardano price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Cardano	RF	0.6	0.84	0.70	0.58	0.437

Table 16: Daily Cardano price prediction report for RF

4.3.5 RF model analysis for daily Solana price prediction.

In this algorithm evaluation, the collection of cross-validation scores are [0.5308642, 0.56790123, 0.66666667, 0.56790123, 0.48148148]. These values originate from an assortment of partitions of the dataset, expressing the model's proficiency with various segments of the data. An average cross-validation score approximating 0.563 translates to a modestly effective method in terms of its generalization capacity on unseen data.

With an accuracy rate of 0.64, the algorithm correctly forecasts outcomes 64% of the time. Although an insightful metric, accuracy doesn't provide a complete picture as it lacks information about the distribution of true and false predictions. Precision, calculated as 0.60, quantifies the proportion of positive identifications that were indeed correct. In this case, 60%

of the model's positive predictions were accurate, suggesting a decent level of precision. Recall, also known as sensitivity or true positive rate, measures the fraction of the total amount of relevant instances that were actually retrieved. With a recall score of 0.84, it indicates that the algorithm successfully identified 84% of all true positives, which is a strong result. The F1 score, sitting at 0.7, can be understood as the harmonic mean of precision and recall. It strives to find the balance between these two metrics. A higher F1 score is desired, as it indicates higher precision and recall. Moreover, the confusion matrix further elucidates the model's performance. It reveals the number of true positives (43), true negatives (22), false positives (29), and false negatives (8). It gives a snapshot of how well the algorithm has performed in terms of binary classification.

Given the above observations, there is room for improvement in the algorithm's performance, especially in terms of precision and the cross-validation score. It would be advantageous to explore other predictive algorithms or tune the parameters of the existing one. Additionally, consideration could be given to refining the features in the model or addressing class imbalance if it exists. The ultimate aim should be to enhance the model's ability to generalize unseen data and make reliable predictions.

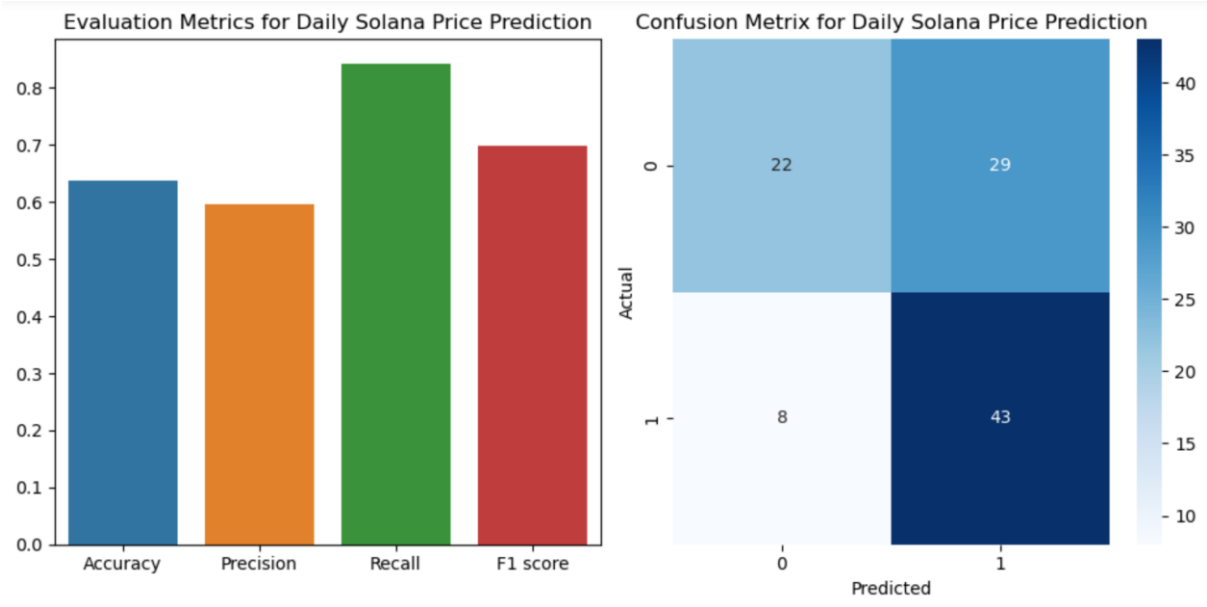

Figure 40: RF model evaluation and confusion metrics for daily Solana price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Solana	RF	0.60	0.84	0.70	0.64	0.563

Table 17: Daily Solana price prediction report for RF

4.4 SVM result analysis

This forthcoming subsection will provide an analysis of the performance of the Support Vector Machine (SVM) algorithm in forecasting the prices of multiple cryptocurrencies. The examination will include a detailed discussion of the model's results and its predictive efficacy.

4.4.1 SVM model analysis for daily Bitcoin price prediction.

The analysis of the Bitcoin price prediction SVM (support vector machine) algorithm reveals various outcomes. The performance of the model was evaluated using a 5-fold cross-validation, resulting in scores ranging from 0.5603 to 0.9655. The mean of these cross-validation scores was computed to be 0.8483, indicating that the model was, on average, about 84.83% accurate on the validation folds. However, for the test data, the algorithm presented an accuracy of 0.53, which recommends that it was capable to correctly classify 53% of the cases. This is somewhat lower than the mean cross-validation score and implies that the model's performance on unseen data was not as strong as it was during cross-validation.

In terms of precision, the SVM method scored 0.5, which indicates that when the model predicted a price increase, it was correct half of the time. The recall score of 0.18 is relatively low, indicating that the model was only able to correctly identify 18% of the actual price increases. This is further corroborated by the F1 score (a harmonic mean of precision and recall) of 0.26, which is quite low and indicates a poor balance between precision and recall. The confusion matrix, on the other hand, provides a more detailed view of the SVM algorithm's performance. Out of 77 instances where the price did not increase, the model correctly predicted 65 and misclassified 12. Conversely, out of 68 instances where the price did increase, the method only correctly predicted 12 and misclassified 56.

Given these outcomes, it is recommended to explore alternative algorithms such as decision trees, random forests, or gradient boosting models, which could better handle the complexity of the data. Additionally, tuning the parameters of the existing model, such as the regularization strength and kernel of the SVM, may yield better results. Lastly, reconsidering the feature selection and engineering could help capture the underlying patterns in the data more effectively. Using a combination of these strategies could potentially lead to substantial improvements in the model's predictive performance.

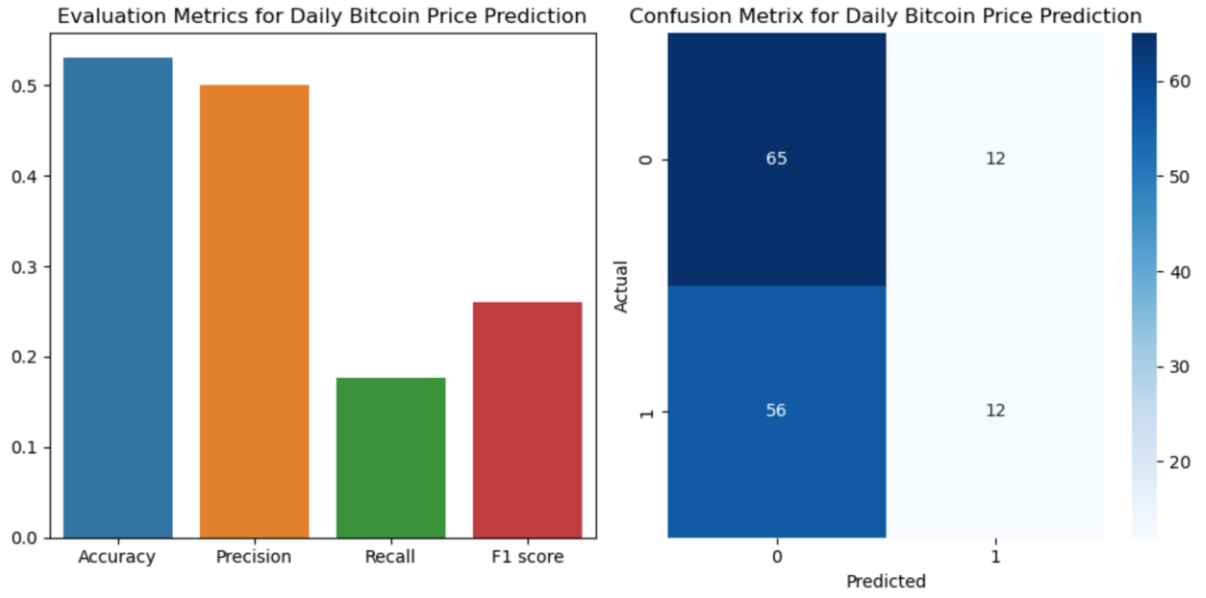

Figure 41: SVM model evaluation and confusion metrics for Daily Bitcoin price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Bitcoin	SVM	0.50	0.18	0.26	0.53	0.848

Table 18: Daily Bitcoin price prediction report for SVM

4.4.2 SVM model analysis for hourly Bitcoin price prediction.

The evaluation of the SVM prediction algorithm exhibits an assortment of notable outcomes. The model's performance was ascertained using a 5-fold cross-validation, yielding scores from 0.8481 to 0.9786. The mean of these cross-validation scores was calculated to be

approximately 0.9069, demonstrating an overall 90.69% accuracy in the validation folds, which suggests a strong performance. However, concerning the algorithm's performance on the test data, an accuracy of 0.8 was recorded, meaning the model correctly predicted 80% of the cases. This level of accuracy, although slightly lower than the cross-validation average, indicates a robust model capable of generalizing well to unseen data.

The SVM method obtained a precision score of 1.0, which suggests that every instance that the model predicted as positive was indeed positive. However, a recall score of 0.59 reveals that the SVM technique was only capable to correctly identify 59% of the actual positive cases. The F1 score, being a harmonic mean of precision and recall, is 0.74, which indicates a balance between precision and recall, skewed towards precision due to its perfect score. Nonetheless, the confusion matrix provides a more detailed illustration of the model's performance. It shows that the algorithm correctly predicted 3336 instances of the negative class without any false positives. However, it identified only 1953 out of 3315 instances of the positive class, with 1362 instances being false negatives.

While the overall performance of the model is quite robust, some improvements can still be made, especially in terms of recall. Techniques such as oversampling the minority class, undersampling the majority class, or a combination (SMOTE) could be employed to balance the dataset and potentially improve the recall. Adjusting the decision threshold of the algorithm might also help in optimizing the recall at the expense of precision. Furthermore, additional features could be engineered or existing ones transformed to enhance the model's capability to capture the underlying patterns in the data. Lastly, advanced ensemble techniques, like gradient boosting or stacking multiple algorithms, could be explored to further enhance the model's performance.

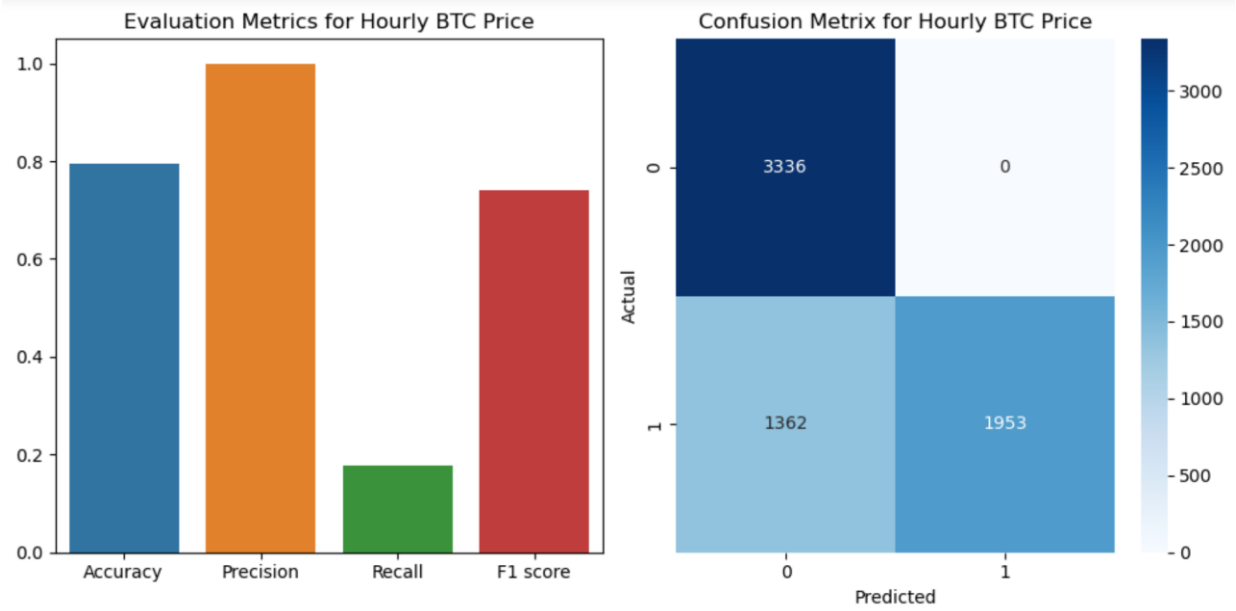

Figure 42: SVM model evaluation and confusion metrics for hourly Bitcoin price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Bitcoin	SVM	1.00	0.59	0.74	0.80	0.907

Table 19: Hourly Bitcoin price prediction report for SVM

4.4.3 SVM model analysis for daily Ethereum price prediction.

In assessing the effectiveness of the SVM predictive algorithm, several remarkable findings have been identified. Employing a 5-fold cross-validation approach, the model's performance scores ranged between 0.6983 and 0.9569. The cross-validation scores recommend a robust model performance across most folds with only the last two folds depicting a drop. The average cross-validation score was calculated as approximately 0.8569, signifying that the algorithm correctly classified 85.69% of the data across the validation folds on average. Examining the model's outcomes on the test data, however, it produced an accuracy score of 0.55. This score reveals that the SVM algorithm accurately predicted 55% of the instances. Though a little lower than the cross-validation average, it still represents a fair degree of prediction power.

The model achieved a precision score of 0.62, indicating that when the SVM method predicted an instance as positive, it was correct 62% of the time. However, the algorithm's recall score was a mere 0.14, which suggests that the model correctly identified only 14% of the actual positive cases. The resulting F1 score, which presents the harmonic mean of precision and recall, is 0.24, showing a significant imbalance between precision and recall, biased towards precision. Moreover, the confusion matrix offers further insight into the method's performance, indicating that out of 76 actual negatives, the model correctly identified 70, and falsely identified 6 as positives. In contrast, for 69 actual positive instances, the algorithm correctly predicted only 10, while misclassifying 59 as negatives.

Given these findings, there are several areas for potential algorithm optimization. Given the low recall score, it is apparent that the model is struggling to identify positive instances. Hence, strategies such as changing the decision threshold or utilizing cost-sensitive learning might improve recall. Furthermore, since the model's performance varied across different folds, data preprocessing steps, like outlier removal or feature scaling, could be revisited and modified for consistency. Feature selection and engineering techniques might also be beneficial for enhancing predictive power. Lastly, given the imbalance depicted in the confusion matrix, techniques like oversampling the minority class or undersampling the majority class could be considered to help improve the model's performance on imbalanced data.

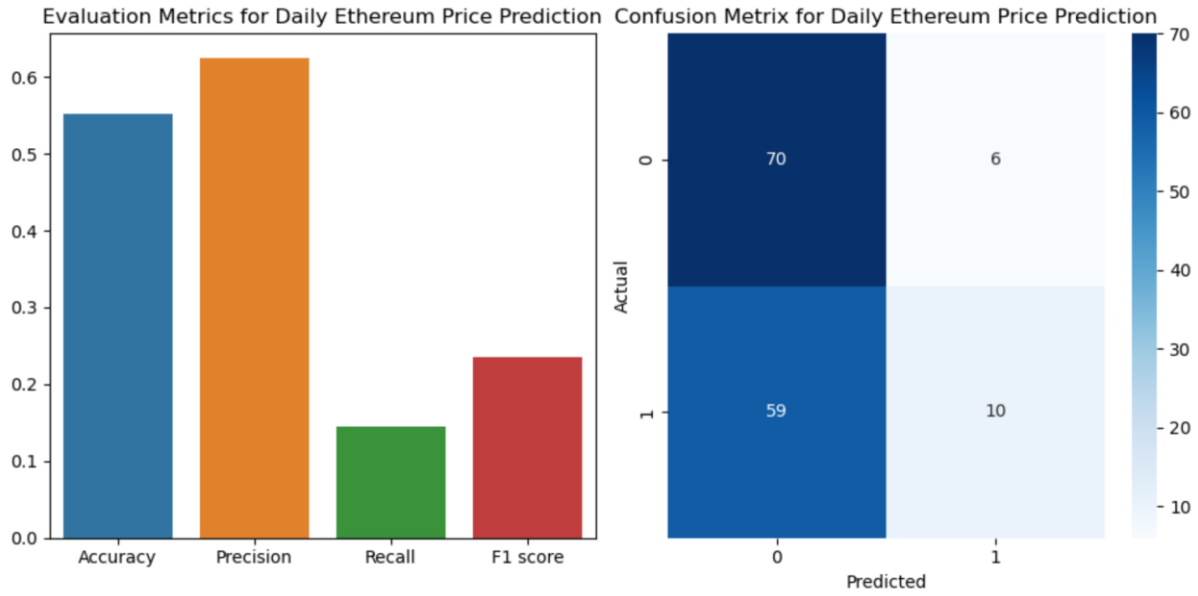

Figure 43: SVM model evaluation and confusion metrics for daily Ethereum price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Ethereum	SVM	0.62	0.14	0.24	0.55	0.857

Table 20: Daily Ethereum price prediction report for SVM

4.4.4 SVM model analysis for daily Cardano price prediction.

This academic study employed a Support Vector Machine (SVM) algorithm to predict daily Cardano prices. The evaluation of predictive algorithm's effectiveness has yielded several notable insights. The cross-validation approach adopted was 5-fold cross-validation. The model's performance, as gauged by the scores across these folds, varied between 0.7863 and 0.9658. While the scores for the most part were quite robust, there was a noticeable drop in the fourth fold. Nevertheless, the mean cross-validation score came out to be approximately 0.9196, indicating an average correct classification rate of 91.96% across all folds, which is highly satisfactory. Further evaluation on the test set provided an encouraging accuracy score of 0.90. This suggests that the algorithm made correct predictions for 90% of the instances in the test set, demonstrating strong predictive power.

In terms of precision, the model achieved a score of 0.86, which denotes that when it predicted a positive class, it was correct 86% of the time. Remarkably, the method obtained a perfect recall score of 1.0, implying that it successfully identified all actual positive cases in the dataset. The resulting F1 score was 0.92, a score that demonstrates an excellent balance between precision and recall. Besides, the confusion matrix further clarifies the performance of the model. It correctly predicted 47 out of 61 negative instances, and falsely identified 14 negatives as positive. For positive instances, the algorithm correctly identified all 85, demonstrating its remarkable sensitivity.

Despite the generally excellent performance of the SVM algorithm, certain areas offer potential for refinement. The disparity in cross-validation scores across the folds suggests some inconsistency in the model's handling of different subsets of the data. This could be mitigated through more uniform preprocessing or through application of stratified cross-validation.

Furthermore, although the recall was perfect, the precision score left some room for improvement, indicating a degree of false positive predictions. To address this, it may be beneficial to revise the decision threshold or to implement a cost-sensitive learning approach. Lastly, given that the confusion matrix showed a number of false positives, it may be useful to further examine these instances to identify any shared characteristics or patterns that could help refine the model's predictive capabilities.

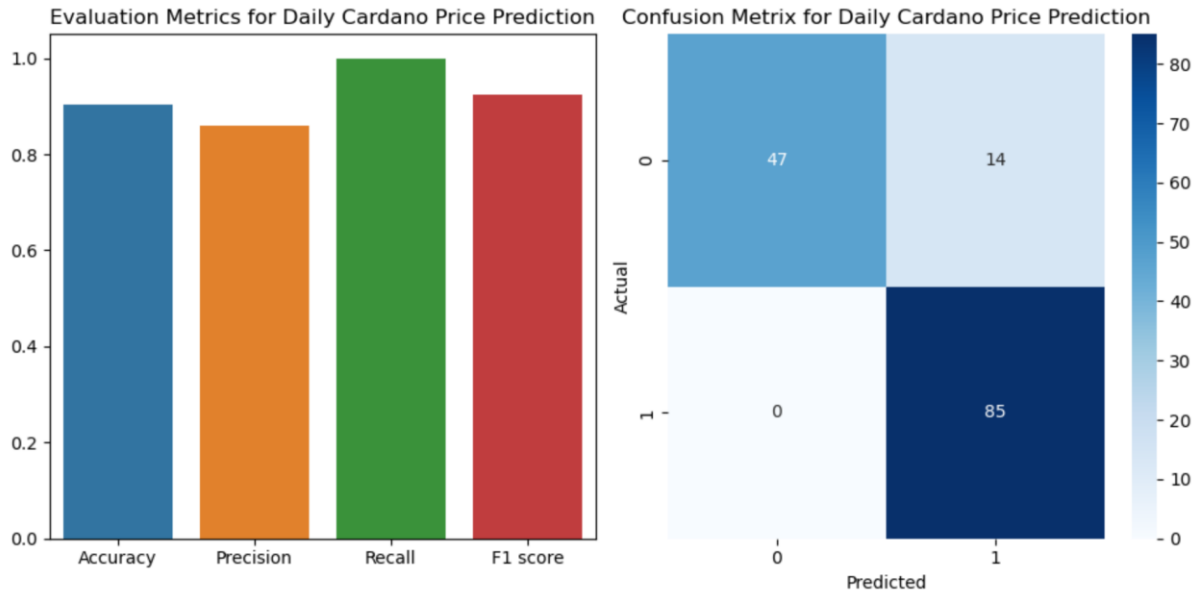

Figure 44: SVM model evaluation and confusion metrics for daily Cardano price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Cardano	SVM	0.86	1.00	0.92	0.90	0.919

Table 21: Daily Cardano price prediction report for SVM

4.4.5 SVM model analysis for daily Solana price prediction.

In this academic study, a Support Vector Machine (SVM) algorithm was utilized for daily Solana price prediction. The application of a five-fold cross-validation procedure to assess our predictive method's performance provided notable results. The model's scores across the

different folds ranged from 0.9383 to a perfect 1.0, a variation that indicates minor inconsistency in the algorithm's behavior with differing subsets of data. Nevertheless, the mean cross-validation score emerged as a robust 0.9704, translating to an average successful classification rate of 97.04% across the five folds. This high score indicates strong reliability and effectiveness of the model. Further the method's performance was evaluated employing a separate test dataset, yielding an impressive accuracy score of 0.97. This suggests that the model was able to correctly predict the classification of 97% of the instances in the test set. This is a strong indicator of the algorithm's predictive power.

In the evaluation of precision and recall, the model delivered excellent outcomes. With a precision score of 0.96, the model was correct 96% of the time when predicting a positive class. The recall score of 0.98 indicates that it was able to identify 98% of the actual positive cases in the dataset. The resulting F1 score, which provides a balanced measure of precision and recall, was 0.97, a further endorsement of the model's proficiency. Furthermore, the SVM algorithm's performance can also be evaluated via the confusion matrix. It correctly predicted 49 out of 51 negative instances, falsely predicting 2 as positive. In the case of positive instances, the model correctly identified 50 out of 51, showcasing its sensitivity.

The outcomes, while impressive, also highlight potential areas of improvement. Although the variation in cross-validation scores was minor, it does suggest that the model may not handle different subsets of data with complete uniformity. To mitigate this, more uniform preprocessing methods or the application of stratified cross-validation could be beneficial. Moreover, given that the model's precision was marginally lower than recall, suggesting a small proportion of false positives, adjusting the method's decision threshold or applying a cost-sensitive learning technique could help to enhance its precision. Finally, as the confusion matrix identified a few instances of misclassification, a deeper examination of these cases could help to uncover specific patterns or characteristics that might further refine the model's predictive ability.

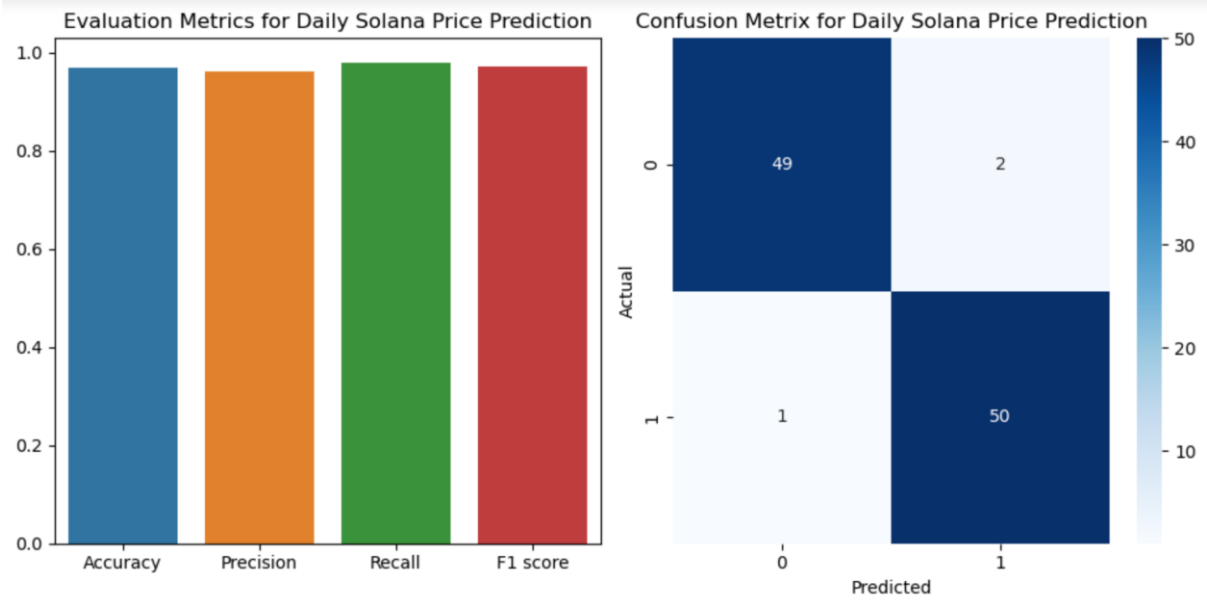

Figure 45: SVM model evaluation and confusion metrics for daily Solana price prediction

Cryptocurrency	Algorithm	Precision	Recall	F1	Accuracy	Mean cross-validation score
Solana	SVM	0.96	0.98	0.97	0.97	0.970

Table 22: Daily Solana price prediction report for SVM

5. Discussion

The recent technological advancements in the underlying technology of cryptocurrency markets have garnered significant attention, positioning them as an alternative investment option. As a result, these markets have become a crucial asset class for both researchers and traders. In the academic discourse surrounding the efficacy of machine learning algorithms for predicting cryptocurrency prices, a critical consideration is the temporal resolution of these predictions.

The focus of the study pertains to the application of diverse machine learning algorithms for predicting the daily and hourly prices of cryptocurrencies, specifically Bitcoin, Ethereum, Cardano, and Solana. The research findings reveal a spectrum of outcomes based on the application of several ML algorithms and the specific cryptocurrency under scrutiny. The application of the Logistic Regression (LR) algorithm demonstrated superior accuracy when predicting the prices of Bitcoin and Ethereum, while the Support Vector Machine (SVM) model outperformed others when applied to Cardano and Solana.

Contrasting this with previous studies, researcher discovers interesting differences. Notably, previous research that leveraged linear regression and decision tree models yielded high predictive accuracy rates for Bitcoin, scoring 97.5% and 95.8% respectively over a span of 5 days. The difference between findings of this research and the previous studies may be partially due to the variations in the duration of prediction. This paper focuses on daily predictions differs from the previous study's 5-day prediction interval, which could have a significant impact on the model's accuracy.

In the realm of financial markets, particularly cryptocurrency markets, short-term predictions are often more accurate due to their reduced susceptibility to accumulated error. Consequently, the higher accuracy rates achieved by the regression-based models in the previous study could be attributed to their emphasis on short-term, 5-day forecasts. On the other hand, ML models of this study were designed for daily predictions, which, despite being short-term, are exposed to a higher risk of inaccuracy due to the volatility and unpredictability of cryptocurrencies. Furthermore, the discrepancy between this paper and the previous

researchers also indicates that while regression-based models are proficient at short-term predictions, they may not perform as efficiently over longer durations. By contrast, SVM, an algorithm that thrives in handling complex, higher-dimensional datasets, might be more capable of discerning the intricate patterns inherent in cryptocurrency prices, particularly for Cardano and Solana.

Delving into the hourly predictions for Bitcoin, the LR model showed the highest accuracy, closely followed by the SVM model. Comparatively, prior research that incorporated the use of RNN, LSTM, and ARIMA models for Bitcoin price prediction over a timeframe of 100 days reported accuracy levels of approximately 50%. This stark discrepancy can be, again, attributed to the prediction duration; shorter prediction intervals likely yield higher accuracy due to the reduction in accumulated errors over time. Moreover, the comparison highlights the prowess of the LR and SVM models in deciphering short-term patterns in Bitcoin's price changes. Conversely, the longer-term methods, such as RNN, LSTM, and ARIMA, while powerful, may be less efficient for short-term forecasts due to their emphasis on capturing long-term dependencies and trends.

Another interesting observation from this research was the relatively superior performance of the Random Forest (RF) algorithm in predicting Solana prices. In contrast, a previous study reported accuracy levels of 59% for Bitcoin and 44% for Litecoin over a 4-day period using a multi-linear regression model. The increased accuracy of the RF model for Solana indicates that ensemble methods, such as RF, may be more suited to capture the intricacies and volatility of certain cryptocurrencies.

Comparing findings in this study with the results of a Granger causality test on Bitcoin, Ethereum, and Ripple, the author observes similar accuracy levels across all three cryptocurrencies. This outcome underscores the idea that different cryptocurrencies may share similar time-series patterns, which certain statistical tests like the Granger causality can potentially capture.

In summary, the insights gleaned from this research contributes to an evolving body of knowledge on cryptocurrency price prediction. This thesis marks an improvement over conventional methods such as LSTM and ARIMA for daily Bitcoin price prediction. However,

the variability in accuracy rates for various cryptocurrencies, particularly when compared with previous studies, signals the need for a tailored, context-specific approach to selecting prediction algorithms.

A clear understanding of each cryptocurrency's behavioral patterns, coupled with the right choice of prediction duration and potential amalgamation of models, is key to optimizing prediction accuracy. This research throws light on the promising yet challenging field of cryptocurrency price prediction, thereby presenting opportunities for future research. Through continuous exploration, author of this thesis hopes to uncover deeper insights that can help advance the scientific community's understanding of cryptocurrency price dynamics.

Cryptocurrency	Algorithm	Accuracy	Duration
Bitcoin, Ethereum, Cardano and Solana	LR	0.86, 0.85, 0.49, 0.50	1 day
Bitcoin, Ethereum, Cardano and Solana	Decision Tree	0.53, 0.50, 0.58, 0.58	1 day
Bitcoin, Ethereum, Cardano and Solana	RF	0.58, 0.52, 0.58, 0.64	1 day
Bitcoin, Ethereum, Cardano and Solana	SVM	0.53, 0.55, 0.90, 0.97	1 day

Table 23: Model accuracy achieved for daily cryptocurrencies price prediction in current study

Algorithm	Cryptocurrency	Accuracy	Duration
LR, Decision Tree, RF and SVM	Bitcoin	0.98, 0.71, 0.63, 0,80	Hourly

Table 24: Model accuracy obtained for hourly Bitcoin price prediction in current research

Learning model	**Cryptocurrency**	**Accuracy**	**Duration**
RNN, LSTM and ARIMA	Bitcoin	50.05%, 50.25% and 52.78%	100 days
Multi-linear regression	Bitcoin and Litecoin	59% for Bitcoin and 44% for Litecoin	4 days
Granger causality test	Bitcoin, Ethereum and Ripple	49.462%, 50.286% and 53.157%	12 days
Conventional LSTM method and LSTM with AR (2) method	Bitcoin	RMSE for both methods 256.41 and 247.33 respectively	Daily
Linear regression and decision tree	Bitcoin	97.5% and 95.8%	5 days
Baseline, logistic regression, SVM and neural network.	Bitcoin	53.4%, 54.3%, 53.7% and 55.1% respectively	Daily

Table 25: Algorithm's accuracies from previous studies for cryptocurrency price prediction

6. Conclusion

This thesis has explored several key areas pertinent to cryptocurrencies and machine learning, which were guided by the research questions posited at the beginning of this study. The research questions for this study were:

1. How does cryptocurrency differ from fiat currency?
2. Which machine learning methods have been used to predict the prices of cryptocurrency in academic research?
3. What is the most effective machine learning method for predicting the price of cryptocurrency?
4. Can the same machine learning model be applied to other cryptocurrencies' price prediction?

The first research question highlighted the fundamental differences between cryptocurrencies and fiat currencies. Cryptocurrencies, unlike government-issued and regulated fiat currencies, are decentralized digital assets that operate using cryptographic and blockchain technology. This fundamental difference contributes to their unique market behavior and price volatility, which often presents a significant challenge for price forecasting.

The second research question led to an examination of a variety of machine learning techniques employed in academic research for predicting cryptocurrency prices. The landscape of applied algorithms is diverse, including linear regression, decision trees, random forests, support vector machines, and several others. However, the effectiveness of these methods tends to vary based on the specific cryptocurrency and prediction horizon, as was explored in response to the third research question.

Answering the third question, the research findings indicate that Logistic Regression and Support Vector Machine algorithms exhibited considerable effectiveness in forecasting daily prices for Bitcoin, Ethereum, Cardano, and Solana in this study context. However, it is crucial to note that a model's success hinges on several factors, including the nature of the data, algorithm parameters, and the duration of the prediction.

Addressing the fourth research question, author found that while a machine learning technique may demonstrate effectiveness for one cryptocurrency, it may not necessarily replicate the same level of accuracy for another. This inconsistency stems from the unique features and market behaviors associated with different cryptocurrencies. Therefore, a tailored approach that aligns the algorithm with the characteristics of each specific cryptocurrency is recommended for optimal prediction accuracy.

Despite the promising results, this research is not without its limitations. The volatile nature of cryptocurrencies, alongside potential external factors and latent variables influencing their prices, can introduce prediction inaccuracies. Future research directions should therefore explore the development of hybrid algorithms or ensemble methods capable of encapsulating the complexities and volatile nature of cryptocurrency markets. Furthermore, investigating the role of external influencing factors and incorporating them into predictive models will be an exciting and valuable future research direction.

Ultimately, this research unveils the multifaceted nature of predicting cryptocurrency prices, navigating through the intricate interplay of machine learning algorithms and individual cryptocurrency characteristics. It underscores the evolving dynamism of this emerging field, reflecting both its potential and its challenges. The outcomes presented are not merely an academic exercise but a substantive advancement that contributes to the broader understanding of financial technology's role in our increasingly interconnected global economy. As the frontier of cryptocurrency continues to expand, the insights drawn from this study are a beacon, guiding future research, policy considerations, and innovative practices. The path illuminated here invites further exploration, promising to unlock new horizons and foster a deeper, more nuanced understanding of cryptocurrency's influence on the financial landscape.

Reference

Abrol, A. (2022). What are blockchain nodes? Detailed guide. Retrieve: What are Blockchain nodes? Detailed Guide - (blockchain-council.org)

Alessandretti, L., ElBahrawy, A., Aiello, L. M. and Baronchelli, A. ORCID: 0000- 0002-0255-0829 (2018). Machine Learning the Cryptocurrency Market. Complexity, 2018. Doi: 10.1155/2018/8983590

Alpaydin, E. (2010). Introduction to machine learning. Cambridge, MA: MIT Press.

Akyildirim, E., Goncu, A. & Sensoy, A. (2021). Prediction of cryptocurrency returns using machine learning. Annals of Operations Research, 297, 3–36. https://doi-org.ezproxy.vasa.abo.fi/10.1007/s10479-020-03575-y

Azure. (2022). What is machine learning? Retrieve: What is machine learning? | Microsoft Azure

Bakar, A., N. & rosbi, S. (2017). Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Cryptocurrency Exchange Rate in High Volatility Environment: A New Insight of Bitcoin Transaction. Retrieve: https://dx.doi.org/10.22161/ijaers.4.11.20

Batista, G. E., Prati, R. C., & Monard, M. C. (2014). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20-29.

Becher, B. (2022). What are blockchain nodes and do they work? Retrieve: What Are Blockchain Nodes and How Do They Work? | Built In

Bennett, K. P., & Parrado-Hernández, E. (2006). The interplay of optimization and machine learning research. IEEE Intelligent Systems, 21(3), 26-35.

Benzinga. (2022). Best cryptocurrencies to hedge against inflation. Retrieve: Best Cryptocurrencies To Hedge Against Inflation – Benzinga

Best, R. D. (2022). Estimate of the number of cryptocurrency users worldwide 2016-2021. Retrieve: https://www.statista.com/statistics/1202503/global-cryptocurrency-user-base/

Best, R. D. (2022). Quantity of cryptocurrencies as of February 3, 2022. Retrieve: https://www.statista.com/statistics/863917/number-crypto-coins-tokens/

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Blankenship, D. (2022). Steps of the research process. Retrieve: http://www.humankinetics.com/excerpts/excerpts/steps-of-the-research-process

Blockgenic. (2018). Different blockchain consensus mechanisms. Retrieve: Different Blockchain Consensus Mechanisms | HackerNoon

Böhme, R., Christin, N., Edelman, B., & Moore, T. (2015). Bitcoin: Economics, technology, and governance. Journal of Economic Perspectives, 29(2), 213-238.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Britwise technologies. (2019). Benefits & Drawbacks of cryptocurrency. Retrieve: Advantages and Disadvantages of Cryptocurrency (britwise.com)

Brown, S. (2021). Machine learning, explained. Retrieve: Machine learning, explained | MIT Sloan

Brownlee, J. (2017). Data Cleaning in Machine Learning. Retrieved https://machinelearningmastery.com/data-cleaning-necessary-for-accurate-machine-learning-models/

Bryman, A. (2012). Social research methods (4th ed.). Oxford University Press.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2), 121–167.

Burns, E. (2022). Machine learning. Retrieve: What Is Machine Learning and Why Is It Important? (techtarget.com)

Bybit. (2020). Explained: what is hashing in blockchain? Retrieve: Explained: What Is Hashing in Blockchain? | Bybit Learn

Byrne, B. M. (2016). Structural equation modeling with AMOS: Basic concepts, applications, and programming. Routledge. DOI: https://doi org.ezproxy.vasa.abo.fi/10.4324/9781315757421

CFI. (2022). Fiat money. Retrieve: Fiat Money - Overview, History, How It Works, Pros and Cons (corporatefinanceinstitute.com)

Chen, J. (2022). Encryption. Retrieve: What Is Encryption? (investopedia.com)

Chih-Hung, W., Yu-Feng, M., Chih-Chiang, L., Ruei-Shan, L. (2018). A New Forecasting Framework for Bitcoin Price with LSTM. DOI: 10.1109/ICDMW.2018.00032

Chen, S., Xu, X., Liu, Y., Wang, C., & Liu, X. (2019). Cryptocurrency price prediction using multiple features and artificial neural network. Journal of Intelligent & Fuzzy Systems, 36(6), 5951-5961.

Chojecki, P. (2021). What is supervised machine learning and how does it relate to unsupervised machine learning? Retrieve: What is supervised machine learning and how does it relate to unsupervised machine learning? | by Przemek Chojecki | Artificial Intelligence in Plain English

Coinbase. (2022). What is a blockchain? Retrieve: What is a blockchain? | Coinbase

Coin cloud. (2021). Cryptocurrency vs Fiat. Retrieve: Cryptocurrency vs Fiat. Just How Different is Crypto from Fiat? | by Coin Cloud | Medium

Creswell, J. W., & Creswell, J. D. (2017). Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications.

Criddle, C. (2021). Bitcoin consumes "more electricity than Argentina". Retrieve: Bitcoin consumes 'more electricity than Argentina' – BBC News

Cryptopedia. (2021). Types of nodes: light nodes, full nodes, and masternodes. Retrieve: Masternodes, Light Nodes, and Full Nodes | Gemini

Cryptosecure. (2022). Understanding crypto secure. Retrieve: Cryptosecure

Danesi, M., & Ghelardi, D. (2018). A Guide to Surveys and Questionnaires for Information Systems Research. In The Oxford Handbook of Survey Methodology (pp. 447-470). Oxford University Press.

Dang, Y., Zhang, L., Huang, X., & Zheng, X. (2018). Descriptive statistics: Concepts, definitions, and applications in clinical research. Journal of the American Osteopathic Association, 118(8), 520-526.

DeMatteo, M. (2022). Bitcoin Doubled Its Value in 2021, Then Nearly Lost It All In the First Month of 2022. Here's a Look at Its Price Over the Years. Retrieve: https://time.com/nextadvisor/investing/cryptocurrency/bitcoin-price-history/

Dutta, B. (2021). 3 types of block in a blockchain network. Retrieve: 3 Types of Block in a Blockchain Network | Analytics Steps

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. Journal of Applied Psychology, 73(3), 421 - 435. DOI: 10.1037/0021-9010.73.3.421

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. Journal of Machine Learning Research, 15(1), 3133-3181.

Founders Guide. (2021). What are the advantages and disadvantages of fiat money? Retrieve: What Are The Advantages and Disadvantages of Fiat Money? (foundersguide.com)

Frankenfield, J. (2022). Currency: what it is, how it works, and how it relates to money. Retrieve: Currency: What It Is, How It Works, and How It Relates to Money (investopedia.com)

Frankenfield, J. (2021). Cryptocurrency. Available: https://www.investopedia.com/terms/c/cryptocurrency.asp.

Frankenfield, J. (2022). Crypto Tokens Definition. Retrieve: Crypto Tokens Definition (investopedia.com)

Frankenfield, J. (2022). Cryptocurrency Explained with Pros and Cons for Investment. Retrieve: Cryptocurrency Explained With Pros and Cons for Investment (investopedia.com)

Geeksforgeeks.org. (2022). Advantages and disadvantages of cryptocurrency in 2020 . Retrieve: Advantages and Disadvantages of Cryptocurrency in 2020 – GeeksforGeeks

Geeksforgeeks.org. (2022). Cryptography in blockchain. Retrieve: Cryptography in Blockchain - GeeksforGeeks

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media, Inc.

Gibson, K., & Koziol, N. (2012). The strengths of secondary data analysis. American Journal of Maternal/Child Nursing, 37(2), 97-101.

Ginez, F. (2019). Bitcoin versus traditional payment systems: is one more effective than other? Retrieve: market-insight-bitcoin-vs-traditional-payment.pdf (wisdomtree.eu)

Godfrey & Clarke. (2000). The Tourism Development Handbook: A Practical Approach to Planning and Marketing. 370 Lexington Avenue, New York, USA.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Great Learning. (2022). Reinforcement learning. Retrieve: Reinforcement Machine Learning-An Introduction to the Basics (mygreatlearning.com)

Greaves, A., & Au, B. (2015). Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin. Retrieve: Using_the_Bitcoin_Transaction_Graph_to_Predict_the_Price_of_Bitcoin.pdf (stanford.edu)

Giudici, G., Milne, A. & Vinogradov, D. (2020). Cryptocurrencies: market analysis and perspectives. J. Ind. Bus. Econ. 47, 1–18. DOI: 10.1007/s40812-019-00138-6

Hadif, A. Hadif, S., A., & Makrakis, D. (20022). Bitcoin Price Prediction using Machine Learning and Technical Indicators. DOI:10.20944/preprints202212.0188.v1

Handscomb, P. (2022). What is the role of cryptography in blockchain? Retrieve: What is the Role of Cryptography in Blockchain? - TechSling Weblog

Harrington, P. (2012). Machine Learning in Action. Manning Publications Co. Greenwich, CT, USA

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer-Verlag New York.

Hayes, A. (2022). Blockchain facts: what is it, how is it works, how it can be used. Retrieve: Blockchain Facts: What Is It, How It Works, and How It Can Be Used (investopedia.com)

Hwang, I. (2022). Fiat currencies: defined, explained, compared to cryptocurrencies. Retrieve: What Is Fiat Currency? Does It Differ From Crypto? | SoFi

IBM. (2022). What is blockchain technology? Retrieve: What is Blockchain Technology? - IBM Blockchain | IBM

IBM. (2023). What is decision tree? Retrieve: What is a Decision Tree | IBM

IBM. (2013). What is logistic regression. Retrieve: What is Logistic regression? | IBM

IBM. (2022). What is Machine learning. Retrieve: What is Machine Learning? | IBM

IBM Developer. (2023). What is reinforce learning? Retrieve: What is reinforcement learning? - IBM Developer

Investing.com. (2023). All cryptocurrencies. Retrieve: All Cryptocurrencies - Investing.com

Iredale, G. (2021). Blockchain cryptography: everything you need to know. Retrieve: Blockchain Cryptography: Everything You Need to Know - 101 Blockchains

Islam, M. (2020). Data Analysis: Types, Process, Methods, Techniques and Tools. International Journal on Data Science and Technology, (6)1, pp. 10-15. DOI: 10.11648/j.ijdst.20200601.12

Jansen, D & Warren, K. (2020). What (exactly) is research methodology? Retrieve: What Is Research Methodology? Definition + Examples - Grad Coach

Jain, A., Tripathi, S., Dwivedi, D. H., Saxena, P. (2018). Forecasting Price of Cryptocurrencies Using Tweets Sentiment Analysis. DOI: 10.1109/IC3.2018.8530659

Javatpoint. (2021). Support vector machine algorithm. Retrieve: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

Jennings. (2001). Tourism Research. Johan Wiley & Sons Ltd, Australia.

Kaggle.com. (2023). Bitcoin historical dataset. Retrieve: https://www.kaggle.com/datasets/prasoonkottarathil/btcinusd?resource=download&select=BTC-Hourly.csv

Karahanna, E., Straub, D. W., & Chervany, N. L. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. MIS Quarterly, 23(2), 183-213. DOI: 10.2307/249751

Kaspersky.com. (2022). What is cryptocurrency and how does it work? Retrieve: https://www.kaspersky.com/resource-center/definitions/what-is-cryptocurrency.

Kelley, D. (2023a). An Introduction to Support Vector Machines (SVM). MonkeyLearn Blog. Retrieved https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

Kelley, K. (2023). What is Data Analysis? Methods, Process and Types Explained. Retrieve: What is Data Analysis? Process, Types, Methods and Techniques (simplilearn.com)

Khan, S. A., Augustine, P. (2019). Predictive Analytics in Cryptocurrency Using Neural networks: A Comparative Study. International Journal of Recent Technology and Engineering (IJRTE).

Khedr, M. A., Arif, I. Raj. P., El-Bannany, M., Alhasmi, M. S. & Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. Doi: 10.1002/isaf.1488.

Khoo, H. L., Lim, K. L., & Tan, C. W. (2021). An ensemble learning framework for cryptocurrency price prediction. Neural Computing and Applications, 33(13), 5975-5990.

Kim, B., Y., Kim, G., J., Kim, W., Im, H., J., Kim, H., T., Kang, J., S., Kim, H., C. (2016). Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies. Retrieve: https://doi.org/10.1371/journal.pone.0161197

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190.

Kotsiantis, S.B., Kanellopoulos, D.N., and Pintelas, P.E. (2006). Data preprocessing for supervised learning. International Journal of Computer Science, 1(2), 111-117.

Lang, N. (2022). What is supervised learning? Retrieve: What Is Supervised Learning?. Get to know the Applications and… | by Niklas Lang | Towards Data Science

Lai, V. (2018). Introduction to cryptography in blockchain technology. Retrieve: Introduction to Cryptography in Blockchain Technology - Crush Crypto

Laycock, R. (2022). Finder's Cryptocurrency Adoption Index. The definitive ranking of the most popular cryptocurrencies across 27 countries.
Retrieve: https://www.finder.com/finder-cryptocurrency-adoption-index

Lee, R. M. (2008). Doing research on sensitive topics. Sage Publications.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. Plos one, 13. DOI:
10.1371/journal.pone.0194889 M

Malone, W. T., Rus, D., & Laubacher, R. (2020). Artificial intelligence and the future of work. Retrieve: 2020-Research-Brief-Malone-Rus-Laubacher2.pdf (mit.edu)

Marsland, S. (2015). Machine Learning: An Algorithmic Perspective (Second Edition). Boca Raton, FL: Taylor & Francis Group.

MathWorks. (2022). What is machine learning? Retrieve: What is Machine Learning? | How it Works, Tutorials, and Examples - MATLAB & Simulink (mathworks.com)

MathWorks (2023). Support Vector Machines for Binary Classification. Retrieved https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

McCombes, S. & George, T. (2022). What is research methodology? steps & tips. retrieve: What Is a Research Methodology? | Steps & Tips (scribbr.com)

McDonald, T. (2022). What is fiat currency? History, how it works and more. Retrieve: What Is Fiat Currency? History, How It Works, and More (askmoney.com)

McNally, S., Jason Roche, J., Simon Caton, S. (2018). Predicting the Price of Bitcoin Using Machine Learning. DOI: 10.1109/PDP2018.2018.00060

Mitchell, T. M. (1997). Machine learning. McGraw Hill.

Morris, A. (2022). What is a block in blockchain? Retrieve: What Is a Block in Blockchain? - Coinformant Australia

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Murray, J. A., Mills, J. S., & Johnson, R. D. (2017). Inferential statistics: An introduction. Journal of the American Association of Nurse Practitioners, 29(11), 728-733.

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. Retrieve: https://bitcoin. org/bitcoin.pdf

Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S. (2016). Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction. Princeton University Press.

Mominul Islam: Cryptocurrencies' Price Discovery Through Machine Learning Algorithms: Bitcoin and Beyond

npr.org (2021). El Salvador Just Became The First Country To Accept Bitcoin As Legal Tender. Retrieve: https://www.npr.org/2021/09/07/1034838909/bitcoin-el-salvador-legal-tender-official-currency-cryptocurrency?t=1649497332112

Pandey, R. (2019). What is a block in blockchain? Retrieve: What is a Block in BlockChain? | SAP Blogs

Pande, S. (2021). What is hash in blockchain technology? Retrieve: What is Hash in Blockchain Technology? (blockchainshiksha.com)

Pang, Y., Sundaraja, G., Ren, J. (2019). Cryptocurrency price prediction using time series and social sentiment data. DOI: 10.1145/3365109.3368767

Patel, M. M., Tanwar, S., Gupta, R., & Kumar, N. (2020). A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions, 55, 102583. DOI: https://doi.org/10.1016/j.jisa.2020.102583

Polikar, R. (2012). Ensemble Learning. In: Zhang C., Ma Y. (eds) Ensemble Machine Learning. Springer, Boston, MA.

Pryadharshini. (2022). What is machine learning and how does it work? Retrieve: What Is Machine Learning and Types of Machine Learning [Updated] (simplilearn.com)

Ray, S. (2017). Cryptographic hashing. Retrieve: Cryptographic Hashing | HackerNoon

Rathan, K., Sai, V. S., Manikanta, S. T. (2019). Cryptocurrency price prediction using Decision Tree and Regression techniques. DOI:10.1109/ICOEI.2019.8862585

Ritchie, J., & Lewis, J. (2003). Qualitative research practice: A guide for social science students and researchers. Sage Publications.

Rosen, A. (2022). What is fiat money, and how does it differ from cryptocurrency? Retrieve: What Is Fiat Money, and How Does it Differ from Cryptocurrency? - NerdWallet

Sahu, M. (2020). Centralized vs Decentralized Cryptocurrency: Difference Between Centralized vs Decentralized Cryptocurrency. Retrieve: (3) New Messages! (upgrad.com)

Sanchez, F., J. (2018). Centralized Cryptocurrencies Explained. Retrieve: Centralized Cryptocurrencies explained. - Coinnounce

SAP. (2022). What is machine learning? Retrieve: What is machine learning? | Definition, types, and examples | SAP Insights

SAS. (2022). Machine learning: what it is and why it matters? Retrieve: Machine Learning: What it is and why it matters | SAS

Savage, C. (2022). What is supervised learning? Retrieve: What is Supervised Learning? | Medium

Schreck, B. (2018). Feature engineering vs feature selection. Retrieve: Feature Engineering vs Feature Selection (alteryx.com)

Selena. (2022). What is a hash in blockchain? Retrieve: What Is A Hash In Blockchain - Pearl Lemon Group

Shah, D., Qureshi, A., Rizwan, M., & Kamal, S. (2018). A comparative study of machine learning models for predicting cryptocurrency prices - A case of Bitcoin. 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET).

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge University Press.

Sheikh, S. (2019). Battery Health Monitoring Using Machine Learning. Retrieve: 1: Commonly used Machine learning Hierarchy. | Download Scientific Diagram (researchgate.net)

Simplilearn. (2023). Machine learning: what it is and why it matters? Retrieve: https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article

Song, J. (2018). Why blockchain is hard. Retrieve: Why Blockchain is Hard. The hype around blockchain is massive… | by Jimmy Song | Medium

Sovbetov, Y. (2018). Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. Journal of Economics and Financial Analysis, 2(2):1–27. DOI:10.1991/jefa.v2i2.a16.

Study.com. (2021). What is research methodology? retrieve: Research Methodology | Examples, Approaches & Techniques - Video & Lesson Transcript | Study.com

Surbhi, S. (2019). Difference between fiat currency and cryptocurrency. Retrieve: Difference Between Fiat Currency and Cryptocurrency (with Comparison Chart) - Key Differences

Surbhi, S. (2019). Difference between fiat currency and cryptocurrency. Retrieve: Difference Between Fiat Currency and Cryptocurrency (with Comparison Chart) - Key Differences

Swan, M. (2015). Blockchain: Blueprint for a new economy. O'Reilly Media, Inc.

Thomas, S. (2020). Fundamentals of blockchain. Retrieve: Fundamentals of Blockchain. Have you ever wondered if we can… | by Sterin Thomas | DataDrivenInvestor

Tutorialspoint.com. (2022). What are the nodes in the blockchain? Retrieve: What are nodes in the Blockchain? (tutorialspoint.com)

University of Cambridge. (2022). Cambridge Bitcoin Energy Consumption Index. Retrieve: Cambridge Bitcoin Electricity Consumption Index (CBECI) (ccaf.io)

VanderPlas, J. (2017). Python data science handbook: Essential tools for working with data. O'Reilly Media.

Veal, A. J. (2006). Research method for leisure and tourism. Pearson Education Ltd.

Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. MIS quarterly, 37(1), 21-54.

Vidrih, M. (2018). What is a block in blockchain? Retrieve: What Is a Block in the Blockchain? | by Marko Vidrih | DataDrivenInvestor

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. New York, NY: Springer.

Worldcoin.org. (2022). What's a blockchain node? Retrieve: What's a Blockchain Node? | Worldcoin

Yan, P., Ganeshkumar, S., Jiewen, R. (2019). Cryptocurrency Price Prediction Using Time Series and Social Sentiment Data. Available: https://doi.org/10.1145/3365109.3368767