**CHAPTER 1**

**INTRODUCTION**

Cryptocurrency is a digital, decentralized form of money based on block chain technology, which makes it the most secure method of making a transaction. There has been a huge increase in the number of cryptocurrencies in the past few years. Cryptocurrencies such as Bitcoin, Dogecoin and Ethereum have become an interesting subject of study in fields such as finance. In 2021, over 4,000 cryptocurrencies are already listed. There are many past studies that focus on predicting the price of cryptocurrencies using machine learning, but the majority of them only focused on Cryptocurrency. Moreover, the majority of the models implemented for price prediction only used the historical market prices, and do not utilize social signals related to the cryptocurrency. In the field of financial technology, many mathematical models are developed to forecast Dogecoin’s future price. These models can provide investment advice for quantitative investors.

* 1. **ARTIFICIAL INTELLINGENCE:**

Artificial intelligence (AI) is the ability of a computer program or a machine to think and learn. It is also a field of study which tries to make computers "smart". As machines become increasingly capable, mental facilities once thought to require intelligence are removed from the definition. AI is an area of computer sciences that emphasizes the creation of intelligent machines that work and reacts like humans. Some of the activities computers with artificial intelligence are designed for include: Face recognition, Learning, Planning, Decision making etc.,

Artificial intelligence is the use of computer science programming to imitate human thought and action by analysing data and surroundings, solving or anticipating problems and learning or self-teaching to adapt to a variety of tasks.

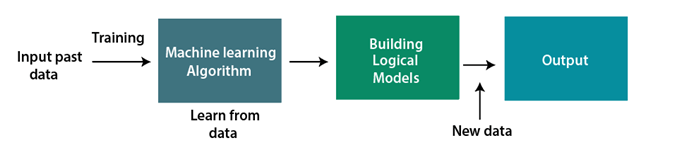
* 1. **MACHINE LEARNING**

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information.** Currently, it is being used for various tasks such as **image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system**, and many more.

Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by **Arthur Samuel**in **1959**. We can define it in a summarized way as: “Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed”.

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it.** The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:



* + 1. **Features of Machine Learning:**
* Machine learning uses data to detect various patterns in a given dataset.
* It can learn from past data and improve automatically.
* It is a data-driven technology.
* Machine learning is much similar to data mining as it also deals with the huge amount of the data.
  + 1. **Classification of Machine Learning**

At a broad level, machine learning can be classified into three types:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

### 1) Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Supervised learning can be grouped further in two categories of algorithms:

* **Classification**
* **Regression**

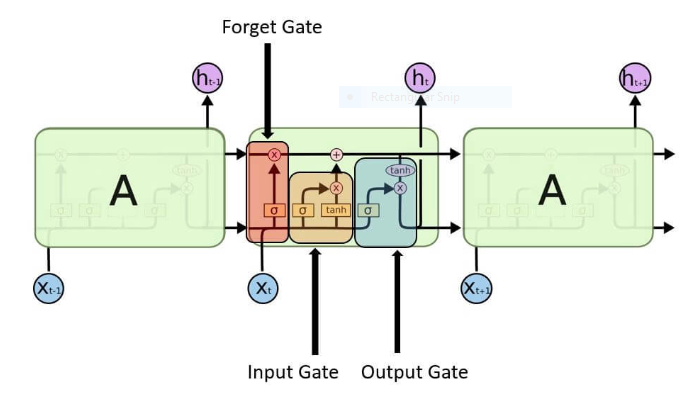
### 2)Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data.

It can be further classifieds into two categories of algorithms:

* **Clustering**
* **Association**
  1. **Long Short-Term Memory**(**LSTM)**
* Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. It was proposed in 1997 by **Sepp Hochreiter**and**Jurgenschmidhuber**. Unlike standard feed-forward neural networks, LSTM has feedback connections. It can process not only single data points (such as images) but also entire sequences of data (such as speech or video).
* A general **LSTM** unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and three gates regulate the flow of information into and out of the cell. LSTM is well-suited to classify, process, and predict the time series given of unknown duration.



* 1. **Root Mean Square Error (RMSE)**

In machine learning, it is extremely helpful to have a single number to judge a model's performance, whether it be during training, cross-validation, or monitoring after deployment. Root mean square error is one of the most widely used measures for this. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

RMSE is a common performance metric in model evaluation, and it is calculated as:



Where n is the number of instances, y is the actual value of the target feature and yˆ is the predicted value of it. Normalized RMSE (NRMSE) is used to compare the performance of different models predicting different target variables and it is calculated as:



Whereymax and yminare the maximum and minimum values of collected data respectively. In the deployment phase, the best model and features will be used to process unseen data and produce prediction results. The model performance is kept on check to validate its prediction results. Practically, and especially in changing environments, the process of training, testing, and deployment are periodically repeated to maintain high accuracy of results. Moreover, this iterative process can be performed to improve performance of the models as historical data become increasingly available.

* 1. **MONTE CARLO SIMULATION**

Monte Carlo simulation is a computerized mathematical technique to generate random sample data based on some known distribution for numerical experiments. This method is applied to risk quantitative analysis and decision making problems. This method is used by the professionals of various profiles such as finance, project management, energy, manufacturing, engineering, research & development, insurance, oil & gas, transportation, etc.

This method was first used by scientists working on the atom bomb in 1940. This method can be used in those situations where we need to make an estimate and uncertain decisions such as weather forecast predictions.

Following are the three important characteristics of Monte-Carlo method −

* Its output must generate random samples.
* Its input distribution must be known.
* Its result must be known while performing an experiment.
  1. **LITERATURE SURVEY**

**[1]Title:** Multi-Head Self-Attention Transformer for Dogecoin Price Prediction

**Authors:** Sashank Sridhar; Sowmya Sanagavarapu

**Description:**

Cryptocurrency market has witnessed a boom during the global pandemic and has proven as a strong investment with a wide institutional adoption. A time-series forecasting solution will play a vital role in analyzing the fluctuation of the bitcoin and altcoin markets. Dogecoin is one such altcoin that is a low-price, high-risk investment option garnering considerable interest this year. The variation of the price trend of this altcoin is studied using the multi-head attention mechanism implemented in a transformer, where the attention heads attend to the tokens that are relevant to each current token based on varying short-term and long-term dependencies. In this paper, a multi-head attention-based transformer encoder-decoder model is applied on the hourly data of the Dogecoin price for its prediction over time. The performance of the model has been evaluated using a number of evaluation metrics including MAE and predictive R-squared value. The model trained over the Dogecoin hourly price variation gave an impressive accuracy of 98.46% and R-squared value of 0.8616 comparable with the existing state-of-the-art cryptocurrency price forecasting models.

# **[2]Title:** Prediction of dogecoin price using deep learning and social media trends

# **Authors:** Basant Agarwal. Priyanka Harjule, Lakshit Chouhan, Upkar Saraswat

# **Description:**

Cryptocurrency is a digital, decentralized form of money based on blockchain technology, which makes it the most secure method of making a transaction. There has been a huge increase in the number of cryptocurrencies in the past few years. Cryptocurrencies such as Bitcoin and Ethereum have become an interesting subject of study in fields such as finance. In 2021, over 4,000 cryptocurrencies are already listed. There are many past studies that focus on predicting the price of cryptocurrencies using machine learning, but the majority of them only focused on Bitcoin. Moreover, the majority of the models implemented for price prediction only used the historical market prices, and do not utilize social signals related to the cryptocurrency.

**[3]Title:** Dogecoin price prediction – can be a determinism supposed?

**Authors:** Juraj Medzihorský

**Description**:

Pump-and-dump schemes are a phenomenon of many cryptocurrencies, including Dogecoin. We have identified repeating shapes on the graph of its price. It is obvious that the price development during last period replicates the development from the previous period. The aim of the article is to create a predictive model based on these replications. Using linear regression, we have developed the model with 87% accuracy on 3-month test set. Therefore, we claim that Dogecoin price development shows a significant degree of determinism. Our model also predicted the significant recent short-term rise (on 28 October 2021), despite the fact it was caused by Elon Musk’s tweet. We have shown that timing of this rise corresponds to the timing of another (also related to Elon Musk’s tweet) during a previous period. Thanks to the determinism in price development, we can predict the beginning of the next large Dogecoin pump most probably on 27 November 2021. This pump should be supported by Elon Musk’s tweet, as well.

**[4]Title:** Cryptocurrency Price Predictions Using High Performance Computing

**Authors:** Marissa Murphy, Naomi Rodriguez, Aaron Morgado, Mason Gawler.

**Description:**

Digital currency has recently gained popularity as it has become increasingly dependent on computers and the Internet. New forms of currency have been constantly evolving over the past few years, namely cryptocurrency. Virtual forms of currency have open new doors within the software industry in finance, data storage, and data collection. Cryptocurrency (crypto) is very volatile in terms of market value, which carries a host of unknowns that make it difficult to predict and analyze the future prices of crypto. However, cryptocurrency behaves similarly to stocks, which allows for the use of linear regression models to make predictions about price levels. With the ability to predict crypto prices, one can make a prediction for crypto stocks since the popular coin, Bitcoin, affects stock prices. This paper will discuss the use of two types of linear regression models, least squares and auto regression, as well as predictors such as social media and economic data to calculate the volatility of a given cryptocurrency and its prices. Using high performance computing techniques will allow regression models to predict relatively accurate crypto prices and past available cryptocurrency price data will be used to verify our results.

# **[5] Title:** Social Media Sentiment in Cryptocurrencies Markets:

# Application of granger causality and deep learning for price prediction

# **Authors:** Luan Fermino Pires

# **Description:**

Behavioural finance literature suggests the existence of bounded rationality within

investors state of mind, largely derived from emotional biases embedded in our human

nature. Consequently, this can affect the decision-making process. The question remains

on whether these emotional biases can affect society in general and their collective

decision making, and whether we are able to extract and test the predictivity that this

formed sentiment can have in financial markets. Most recently the rise of social media platforms has drawn attention as a valuable source of investor sentiment, more specifically along retail investors which are more prone to emotions biases, contrary to institutional investors, as suggested in previous studies. The purpose for this thesis research is on firstly testing whether investor sentiment derived from Twitter is significant in predicting cryptocurrency returns. We analyse the text using the Valence Aware Dictionary for Sentiment Reasoning (VADER) and classify tweets into positive, negative, and neutral polarities. Our results suggest that sentiment from the previous 3-days is found to be statistically significant predictor of daily BTC returns. Secondly using deep learning techniques, we create a neural network to classify text from Elon Musk tweets accordingly to the subsequent price movement of BTC and Dogecoin (DOGE) after the tweets posting. We find that the model is able to predict very short -term price movements of 1-minute and 30-minute, but unable for longer time-periods of 1-day.

We conclude that sentiment from Twitter presents itself as a powerful tool for the prediction of cryptocurrencies price movement as investors seem to be more driven by sentiment rather than fundamentals when trading cryptocurrencies, questioning the Efficient Market Hypothesis (EMH) for this market segment.