GOOGLE PLAYSTORE APPLICATION ANALYSIS AND PREDICTION

INDEX

ABSTRACT

ABSTRACT

Application distribution platform, for example, Google play store gets overwhelmed with few thousands of new applications regularly with a lot progressively a huge number of designers working freely or on the other hand in a group to make them successful. With the enormous challenge from everywhere throughout the globe, it is basic for a developer to know whether he is continuing the correct way. Dissimilar to making movies where the nearness of famous heroes raise the likelihood of accomplishment even before the movies are coming into the picture, it isn’t the situation with creating applications. Since most play store applications are free, the income model is very obscure and in accessible regarding how the in-applications buys, in-application adverts and memberships add to the achievement of an application.

In this way, an application’s prosperity is normally dictated by the quality of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. so in this project, I have tried to perform analysis and prediction into the Google play store application dataset that I have collected from kaggle.com. Using big data techniques such as Machine learning I have tried to discover the relationships among various attributes present in my data set such as which application is free or paid, about the user reviews, rating of the application. And using Deep learning I have tried to make a prediction about the user reviews that which review is positive or negative.

INTRODUCTION

INTRODUCTION

1.1 INTRODUCTION

Big Data likewise information yet with a gigantic size. Big Data is a term used to portray a gathering of information that is big in size but then developing exponentially with time. In short, such information is so substantial and complex that none of the customary information the board devices can store it or produce it proficiently.

We can define a data is a big data with the help of these 5v's:

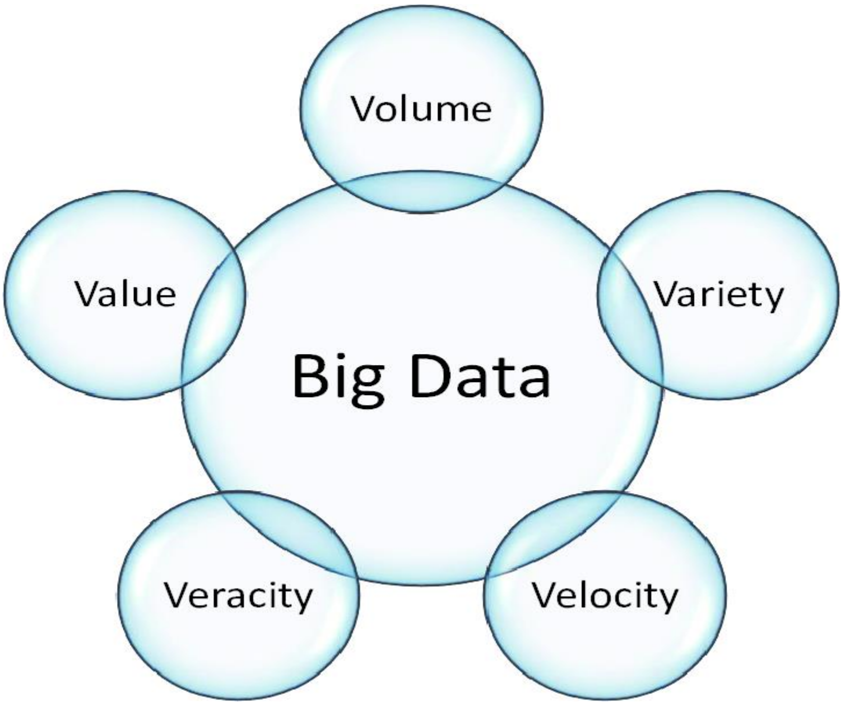


Figure 1:5v's of BIG DATA

* Volume: Volume means a huge amount of data is generated in every second from any social media, cars, bank, from flights etc.

* Velocity: It means speed at which the data is generated and collected, also analyzed.

* Variety: It means different types of data we are having and using it.

* Veracity: It refers to the quality and security of the data.

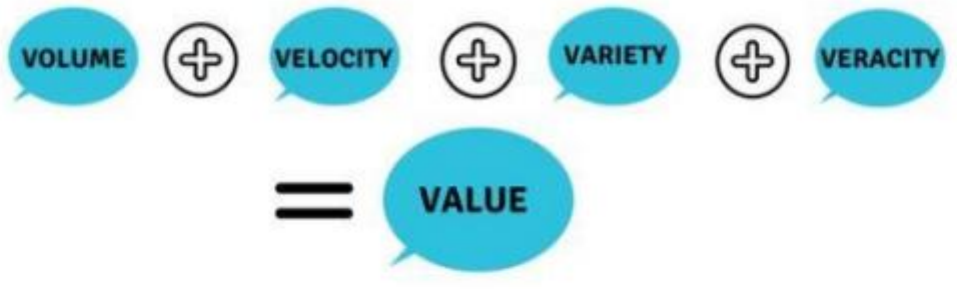


Figure 2: value in 5vs of BIG DATA

Value: In this we refer to the worth of the data that is going to be extracted.

We can perform any type of analysis using big data with the following way:

□ Collection of data

□ Classification of data

□ Identification of data

□ Finally the prediction

□ Visualization

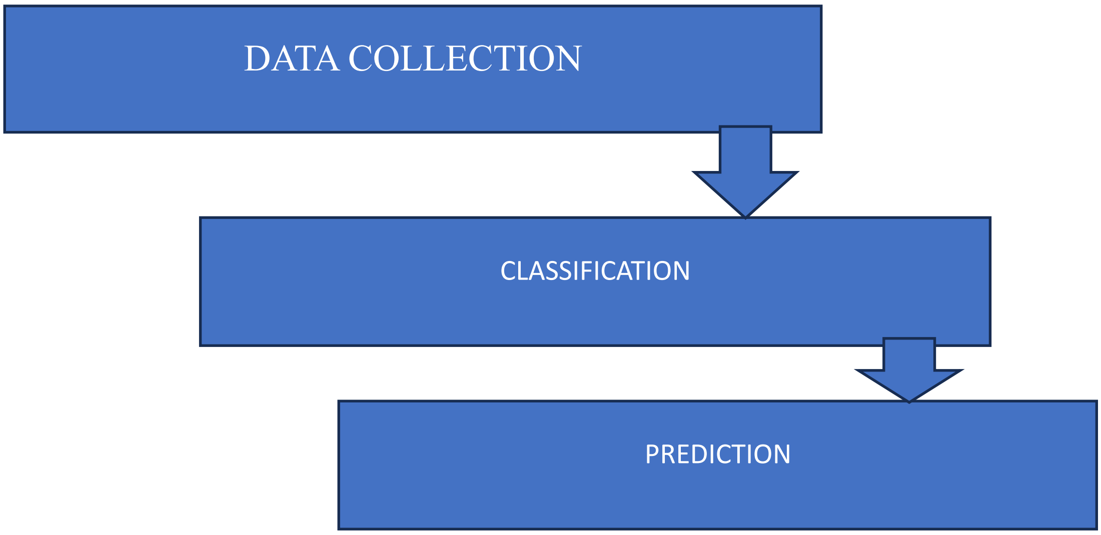


Figure 3: steps for Rating Analysis

1.1.1ANALYSIS AND GOOGLE PLAY STORE APPS

In today's scenario we can see that mobile app playing an important role in any individual's life. It has been seen that the development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry. With enormous challenge from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The Google play store is observed to be biggest application platform. It has been seen that in spite of the fact that it creates more then two-fold the downloads then the Apple App Store yet makes just a large portion of the cash contrasted with App Store. In this way, I scratched information from the play store to direct examination on it. With the fast development of advanced cells, portable applications (Mobile Apps) have turned out to be basic pieces of our lives. Be that as it may, it is troublesome for us to follow along the fact and to understand everything about the apps as new applications are entering market each day. It is accounted for that Android marker achieved a large portion of a million applications in September 2011. Starting at now, 0.675 million android applications are accessible on Google play app store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination extend. We trust versatile application clients consider online application

1.1.2 GOOGLE PLAYSTORE DATA

The dataset taken is of Google play store application and is taken from Kaggle, which is the world's largest community for data scientists to explore, analyse and share data. Google play is an online store that offers users apps, TV shows, books, movies, and more. It collects data as people use anything the store provides. Google play's main information page says the information gathered helps developers learn how people find their apps and how much they'll pay for them. Then, once a person has a Google play product on their device, the collected data includes how much they engaged with it and lets developers read user feedback. If a play store app has paid features, developers receive user data showing when and how often people make purchases and whether they become repeat buyers.

1.1.3 MACHINE LEARNING

Apache Machine learning is defined as a data warehouse system built on top of Apache Hadoop for analysis and applying query on a largest dataset. It changes over SQL- like queries into Map Reduce for simple execution and preparing of amazing substantial volumes of information.

The three significant functionalities for which machine learning is conveyed are: data summarization, data analysis and data query. The 1 query language that is solely upheld by machine learning is the machine learning QL. This language interprets the SQL- like 1queries into map reduce jobs for conveying it on Hadoop. Machine learning QL likewise underpins map reduce contents that can be connected to the machine learning queries. Machine learning builds the patterns structure adaptability and further more data serialization and deserialization.

A typical machine learning tasks are to provide a recommendation. Recommender systems are a common application of machine learning, and they use historical data to provide personalized recommendations to users. In the case of Netflix, the system uses a combination of collaborative filtering and content-based filtering to recommend movies and TV shows to users based on their view in history, ratings, and other factors such as genre preferences.

WHY MACHINE LEARNING?

The Apache Machine Learning is mostly utilized for information querying, analysis and summarization. It improves the designer profitability which comes at the expense of expanding inactivity, and diminishing proficiency. Machine learning stands tall when contrasted with SQL and an awesome one for sure. Machine learning has the future of your organization. In particular, we see tremendous impact occurring within the customer care industry, whereby machine learning is allowing people to get things done more quickly and efficiently. Through Virtual Assistant solutions, machine learning automates tasks that would otherwise need to be performed by a several practical applications that drive the kind of real business results - such as time and money savings - that have the potential to dramatically impact live agent - such changing a password or checking an account balance. This frees up valuable agent time that can be used to focus on the kind of customer care that humans perform best: high touch, complicated decision-making that is not as easily handled by a machine. At Interactions, we further improve the process by eliminating the decision of whether a request should be sent to a human or a machine: unique Adaptive Understanding technology, the machine learns to be aware of its limitations, and bailout to humans when it has low confidence in providing the correct solution.

A subset of artificial intelligence (AI), machine learning (ML) is the area of computational science that focuses on analysis and interpreting patterns and structures in data to enable learning, reasoning, and decision making outside of human interaction. Simply put, machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analysis and make data-driven recommendations and decisions based on only the input data. If any corrections are identified, the algorithm can incorporate that information to improve its future decision making.

1.1.4 HADOOP AND DISTRIBUTED FILE SYSTEM

The Hadoop Distributed File System (HDFS) is a distributed file system. It is a core part of Hadoop which is used for data storage. It is designed to run on commodity hardware.

Unlike other distributed file system, HDFS is highly fault-tolerant and can be deployed on low-cost hardware. It can easily handle the application that contains large data sets.

Let's see some of the important features and goals of HDFS.

Highly Scalable - HDFS is highly scalable as it can scale hundreds of nodes in a single cluster.

Replication - Due to some unfavourable conditions, the node containing the data may be loss. So, to overcome such problems, HDFS always maintains the copy of data on a different machine.

Fault tolerance - In HDFS, the fault tolerance signifies the robustness of the system in the event of failure. The HDFS is highly fault-tolerant that if any machine fails, the other machine containing the copy of that data automatically become active.

Distributed data storage - This is one of the most important features of HDFS that makes Hadoop very powerful. Here, data is divided into multiple blocks and stored into nodes.

Portable - HDFS is designed in such a way that it can easily portable from platform to another.

Goals of HDFS

Handling the hardware failure - The HDFS contains multiple server machines. Anyhow, if any machine fails, the HDFS goal is to recover it quickly.

Streaming data access - The HDFS applications usually run on the general-purpose file system. This application requires streaming access to their data sets.

Coherence Model - The application that runs on HDFS require to follow the write-once-ready-many approach. So, a file once created need not to be changed. However, it can be appended and truncate.

1.1.5 DEEP LEARNING

Deep learning is based on the branch of machine learning, which is a subset of artificial intelligence. Since neural networks imitate the human brain and so deep learning will do. In deep learning, nothing is programmed explicitly. Basically, it is a machine learning class that makes use of numerous nonlinear processing units so as to perform feature extraction as well as transformation. The output from each preceding layer is taken as input by each one of the successive layers.

Deep learning models are capable enough to focus on the accurate features themselves by requiring a little guidance from the programmer and are very helpful in solving out the problem of dimensionality. Deep learning algorithms are used, especially when we have a huge no of inputs and outputs.

Since deep learning has been evolved by the machine learning, which itself is a subset of artificial intelligence and as the idea behind the artificial intelligence is to mimic the human behaviour, so same is "the idea of deep learning to build such algorithm that can mimic the brain. the help of Neural Networks, and the idea behind the motivation of Neural Network is the biological neurons, which is nothing but a brain cell.

1.1.6 NEURAL NETWORK

Model in accordance with the human brain, *a* Neural Network was built to mimic the functionality of a human brain. The human brain is a neural network made up of multiple neurons, similarly, an Artificial Neural Network (ANN) is made up of multiple perceptron (explained later).

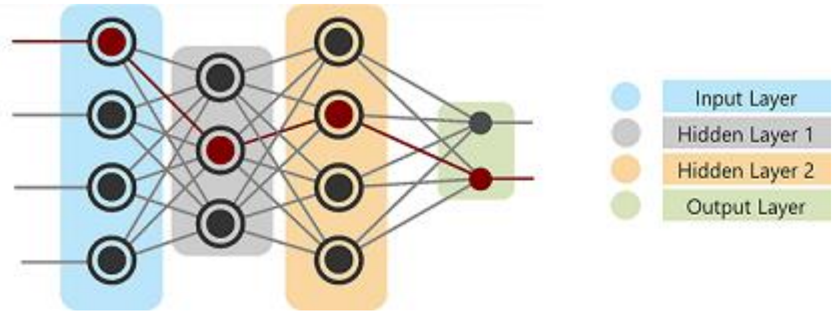


Figure 4: NEURAL NETWORK

A neural network consists of three important layers:

DEEP NEURAL NETWORK

An artificial neural network (ANN) or a simple traditional neural network aims to solve trivial tasks with a straightforward network outline. An artificial neural network is loosely inspired from biological neural networks. It is a collection of layers to perform a specific task. Each layer consists of a collection of nodes to operate together.

These networks usually consist of an input layer, one to two hidden layers, and an output layer. While it is possible to solve easy mathematical questions, and computer problems, including basic gate structures with their respective truth tables, it is tough for these networks to solve complicated image processing, computer vision, and natural language processing tasks.

For these problems, we utilize **deep neural networks**, which often have a complex hidden layer structure with a wide variety of different layers, such as a convolutional layer, max-pooling layer, dense layer, and other unique layers. These additional layers help the model to understand problems better and provide optimal solutions to complex projects. A deep neural network has more layers (more depth) than ANN and each layer adds complexity to the model while enabling the model to process the inputs concisely for outputting the ideal solution.

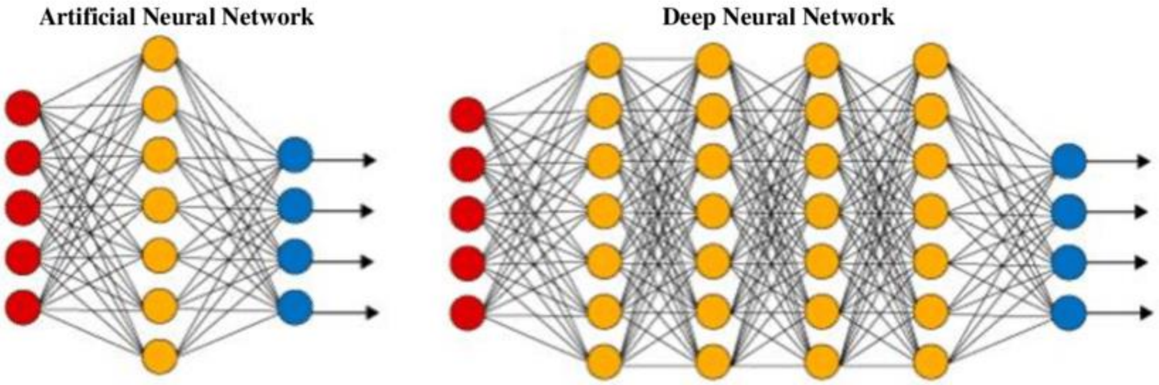


Figure 5: The building block of deep neural network

1.1.7 WHAT IS THE DIFFERENCE BETWEEN MACHINE LEARING AND DEEP LEARNING?

Machine Learning and Deep Learning are the two main concepts of Data Science and the subsets of Artificial Intelligence. Most of the people think the machine learning, deep learning, and as well as artificial intelligence as the same buzzwords. But in actuality, all these terms are different but related to each other.

In this topic, we will learn how machine learning is different from deep learning. But before learning the differences, lets first have a brief introduction of machine learning and deep learning.

1.1.8 WHY DEEP LEARNING?

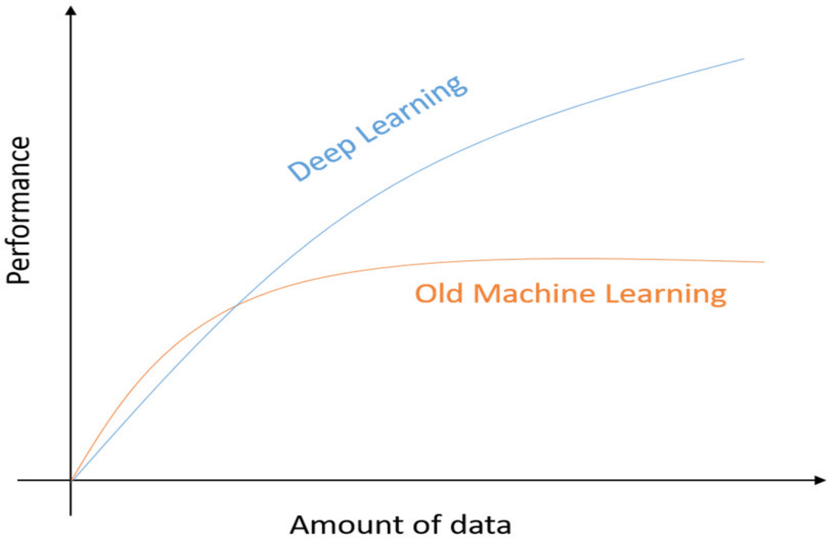


Figure 6: performance of deep learning

Deep learning is all the rage today, as companies across industries seek to use advanced computational techniques to find useful information hidden across huge swaths of data. While the field of artificial intelligence is decades old, breakthroughs in the field of artificial neural networks are driving the explosion of deep learning.

Attempts at creating artificial intelligence go back decades. In the wake of World War II, the English mathematician and codebreaker Alan Turning penned his definition for true artificial intelligence. Dubbed the Turing Test, a conversational machine would have to convince a human that he was talking to another human.

It took 60 years, but a computer finally passed the Turing Test back in 2014, when a chat bot developed by the University of Reading dubbed “Eugene” convinced 33% of the judges convened by the Royal Society in London that he was real. It was the first time that the 30% threshold had been exceeded.

Since then, the field of deep learning and AI has exploded as computers get closer to delivering human-level capabilities. Consumers have been inundated with an array of chat bots like Apple‘s Siri, Amazon‘s Alexa, and Microsoft‘s Cortana that use natural language processing and machine learning to answer questions.

Now, companies across all industries are looking to leverage their big data sets as a training ground to develop sharper AI programs that can interact with the world in ever-more natural ways, and extract useful information from it in ways that have never been done before. Researchers have found that the combination of advanced neural networks, ready availability of huge masses of training data, and extremely powerful distributed GPU-based systems have given us the building blocks for creating intelligent, self-learning machines that can begin to rival humans in their perception.

1.2 PROBLEM STATEMENT

In this project we have focused on our two objectives. we have taken the dataset and observed it nicely and as per our need we have taken various attributes to analysis and further display the result. By doing this we have clearly and easily observe the data set. Moreover, Firstly, I will analysis the different attributes given in the dataset. Secondly, I will do prediction of those different attributes like predict whether the user review is positive or negative.

1.3 OBJECTIVE

Project objectives are goals, plain and simple. These are the business objectives that you want the project to accomplish. Within project management, it is of the utmost importance that a project Objectives are stated clearly, as these will impact every decision in the project lifecycle.

Project objectives must be measurable and contain key performance indicators that will be used to assess a project overall success. These indicators will often include criteria such as budget, quality, and time to completion.

1.4 METHODOLOGY

The implementation of the project is divided into 2 parts:

1.4.1ANALYSIS

In this, various attributes like application name, category, rating, reviews, size, installs, type, price, content rating genres, last updated, version, android version are analysis using Machine
Learning QL.

Steps for implementing the analysis part:

- First, I have selected the dataset.
- Then that information is sent to HDFS.
- Then in machine learning I have created a database.
- In the above created database tables are created.

To create the table named Play store:

Create table play store (app string, category string, rating float, reviews int, size string. Installs string, type string, price int, content string, genres string, last up string, current string, android string) row format delimited fields terminated by ',' stored as text file;

To load the data in the table created by using command, we will have to use the following
command:

Load data local input," googleplaystore.csv" into table play store;

Below mentioned are the various queries that a user can use to retrieve information:

// Names of all free apps

// Names of all paid apps

Select play store .app, playstore. type from google. playstore where type='Paid';

// Names of top free apps

Select playstore.app,playstore.type,playstore.rating

// Names of top free apps

Select playstore.app,playstore.type,playstore.rating

// Names of top free apps

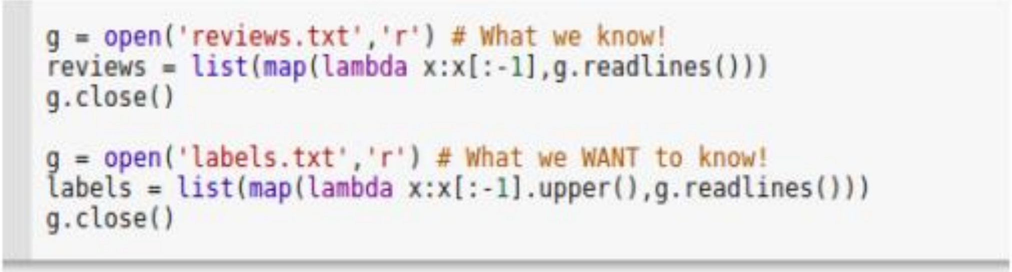
Select playstore.app, playstore.type,playstore.rating from google.playstore where
type='Free' and rating&gt;=4;

// Name of top free apps

Select playstore.app,playstore.type,playstore.rating from google.playstore where
type=’Paid’ and rating&gt;=4;
// Names of all distinct categories
Select distinct play store. category from google. play store;
// Names of top reviewed apps
Select playstore.app, plays tore. reviews from google. playstore were
Reviews&gt; =20000000;
// Editor’s choice
Select playstore.app, playstore. rating, playstore. reviews from google. playstore were
Rating&gt; =4 and reviews&gt; =20000000;

.5PREDICTION

□ Initially I have read the given dataset which is in the text from and labeled it accordingly, as shown in figure:



```
g = open('reviews.txt','r') # What we know!
reviews = list(map(lambda x:x[:-1],g.readlines()))
g.close()

g = open('labels.txt','r') # What we WANT to know!
labels = list(map(lambda x:x[:-1].upper(),g.readlines()))
g.close()
```

Figure 7: Reading the dataset

- I have broken the dataset into training and testing data. The training input data and the testing input data is one hot encoded by seeing the words in a review. For example, if there is a review like “this is a good application”. So, the positioning of each word will be setup in an array and the position of the word in an array is set to one and remaining will be zero. let us assume this has a position of 3 in an array, good has a position of I, has a position of zero, application has a position of 4 and is has a position of 2. Total words are 10. Let’s assume so the array will look like

Import numpy as np
arr= np. zeros (10)
arr [0] =1
arr [1] =1
arr [2] =1
arr [3] =1
arr [4] =1
so, this will be the input to our neural network and out will be either positive or negative.

Positive output will be set to 1

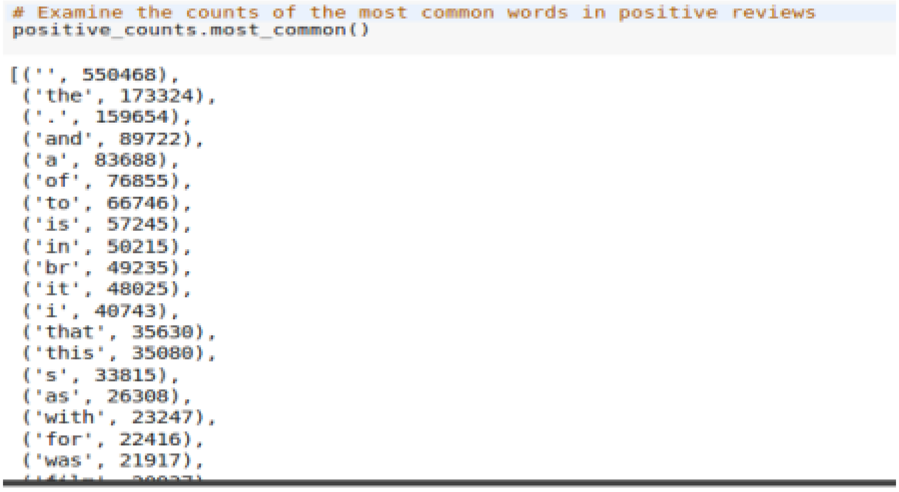
Negative output will be set to 0

So, in this case “this is a good movie” our input look like

[1,1,1,1,1,0,0,0,0,0,] output will be I because this is a positive review

Similarly other reviews will be stored in this fashion.

From the above array of data, I have tried to find out the most common words used in positive review as can be seen from the figure below:

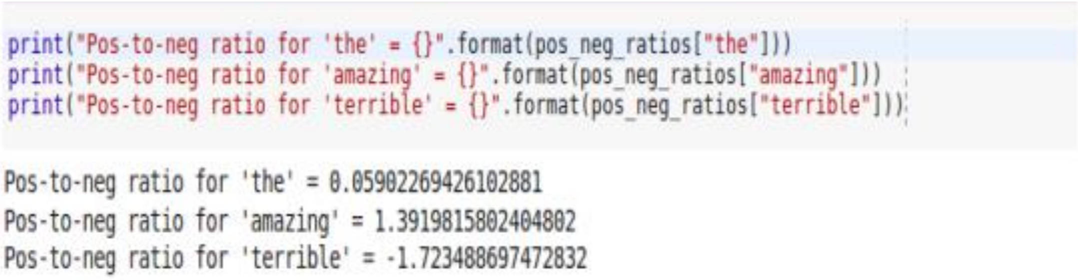


```
# Examine the counts of the most common words in positive reviews
positive_counts.most_common()

[('', 550468),
 ('the', 173324),
 ('.', 159654),
 ('and', 89722),
 ('a', 83688),
 ('of', 76855),
 ('to', 66746),
 ('is', 57245),
 ('in', 50215),
 ('br', 49235),
 ('it', 48025),
 ('i', 40743),
 ('that', 35630),
 ('this', 35080),
 ('s', 33815),
 ('as', 26308),
 ('with', 23247),
 ('for', 22416),
 ('was', 21917),
```

Figure 8: most common words in positive review

After finding the most common words in positive or negatives review I have calculate the positive to negative ratio as shown in figure :



```
print("Pos-to-neg ratio for 'the' = {}".format(pos_neg_ratios["the"]))
print("Pos-to-neg ratio for 'amazing' = {}".format(pos_neg_ratios["amazing"]))
print("Pos-to-neg ratio for 'terrible' = {}".format(pos_neg_ratios["terrible"]))

Pos-to-neg ratio for 'the' = 0.05902269426102881
Pos-to-neg ratio for 'amazing' = 1.3919815802404802
Pos-to-neg ratio for 'terrible' = -1.723488697472832
```

Figure 9: post-to-neg ratio

- ➢ Since the English alphabetical data is converted to mathematical value. And now will be passed through the neural network for testing purposes.
- ➢ The data is passed through the neural network with some randomized weights on each link using feed forward process.
- ➢ Based on the weight model predict the value and based on these predictions the value is compared with the actual output.

- Error functions are calculated based on these predictions and actual values. Backpropagation technique is used to adjust these weights by minimizing the error function. These error is minimized using stochastic gradient descent and after that new weight is adjusted. This step is performed several times keeping in mind that model doesn't over fit the given training data.
- After training process testing is done and based on this dataset our accuracy factor is calculated. In this model we set the probabilistic values i.e. if we get the actual output
- 0.5 then review would be positive and if the output
- 0.5 then the review would be negative otherwise neutral

LITERATURE SURVEY

LITERATURE SURVEY

In the recent years, enormous work is carried out in the domain of Weather forecasting. Weather forecasting is one of the applications to predict state of climate in future at a given location.

2.1 "BIG DATA TECHNIQUES FOR EFFICIENT STORAGE AND PROCESSING OF WEATHER" [1]

This Research paper proposes an efficient Big Data technique for storage and processing of weather data. In general Apache Hadoop framework is most popularly use for storing and processing of enormous dataset. During this study, Apache Spark and Cassandra integration is experimented to judge the time taken to efficiently store any datasets and process it and therefore the result is evaluated with Hadoop Map Reduce

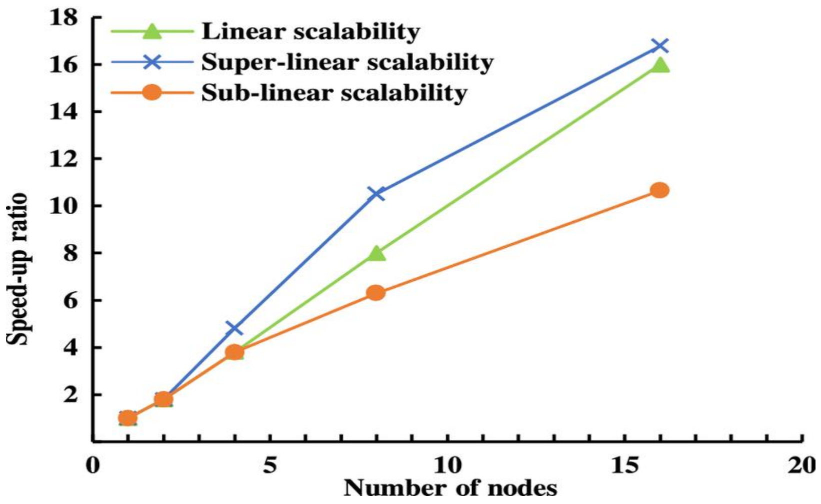
Weather datasets is collected from National Climatic Data Centre (NCDC). In weather forecasting the raw information is received through satellite delivered over to the various weather stations and this data stored in cluster. Traditional Database like SQL are not best to handle unstructured data or weather data. Input datasets contain field like location, date, temperature, humidity, pressure, rain, wind etc.

METHODOLOGY USED IN THIS RESEARCH PAPER: -

Hadoop Map Reduce implementation: -Hadoop Framework works on parallel processing distributed system which are conventionally based on map reduce jobs. The Input data is splitted into split. These splits are passed to mapper and the result output is given as input to reducer. Hence in this research paper spark used to process weather data compared to Hadoop Map Reduce.

➤ **Spark implementation**: - Due to its in-memory computation it can perform 10x better than Map Reduce. Core concept in Apache Spark is RDD which act as a table in database and can hold various type of data and store on different partition.

> **NoSQL Database Cassandra**: -NoSQL Database provides a provision for storage and retrieval of unstructured data unlike the traditional database which use tabular relations. NoSQL Database are being efficiently used for real time web application and high speed online transactional data.



Graph 1: Hadoop Map Reduce v/s Spark Cassandra Benchmarking

2.2 "COMMERCIAL PRODUCT ANALYSIS USING HADOOP MAP REDUCE"

It examines how an association can find certified open entryways in solidifying disengaged and data to give cleverness on how consolidating separated and online data can be helpful. Associations proposition estimations which have the above favorable circumstances. Proposition computations are best seen for their use on online business Web destinations. Here they use customer's interests as a commitment to make a record of endorsed things.

The first one is called content based sifting. Content based separating can moreover be called as intellectual sifting, which endorses things dependent on an examination between the substance of the things and a customer profile.

Additionally, the next one is community oriented sifting. It relies on not just the attributes of the things yet rather how person's for example various customers respond to comparable articles. Affiliations need to get all of the data characteristics, detached and on the web, into a lone database, which would be moreover refined by front line examination procedures, and use the solidified data for exactness concentrating on.

2.3 "REVIEW PAPER ON HADOOP AND MAP REDUCE"

Gigantic Data is a data whose scale, superior to average collection, and multifaceted nature require new planning, procedures, figuring, and examination to guide it and center respect and masked picking up from it.

Hadoop is the center stage for dealing with Big Data, and manages the issue of making it huge for examination purposes. Hadoop is an open source programming experience that draws in the appropriated treatment of huge enlightening accumulations crosswise over packs of thing servers. It is proposed to scale up from a singular server to a large number machines, with an anomalous condition of acclimation to inside disillusionment.

This paper totally inspects why hadoop is better in all edges. Seeking after focuses base on the upsides of hadoop.

Flexibility surpasses desires at dealing with data of complex nature and its open-source nature makes it much surely understood. In this paper the structure of hadoop and guide diminish is cleared up. Guide Reduce has been appeared as a free stage as preference layer legitimate for various need by cloud suppliers. It in addition empowers clients to value the data dealing with and investigating.

2.4 "SENTIMENT ANALYSIS OF TWITTER DATA WITHIN BIG DATA DISTRIBUTED ENVIRONMENT FOR STOCK PREDICTION"

The paper examined a securities exchange expectation probability dependent on gathering of information originating from the required tweets from Twitter small scale blogging stage.

Tweets just in dialect English were utilized in this undertaking work. Retweeted posts were viewed as excess for arrangement, so were evacuated. After information pre- handling, each tweet was spared as a model of pack of-words, a standard procedure for disentangling the spoke to data utilized in data recovery.

The structure of the framework comprised of four noteworthy parts:

➤Retrieving, pre-processing and sparing the twitter information to our database

Stock showcase information recovery

Extremity examination is that part of sentiment investigation, in which the information is gathered either as positive or negative. Customized estimation identification of tweets was pulled off by utilizing Senti Word Net. Future stock costs expectation is performed in this paper by joining the consequences of grouped assumption tweets and stock prices from some past interim.

Viewing huge amounts of information as sorted and the reality they are a composed content, the Naïve Bayes calculation was chosen for its quick procedure of preparing even with huge amount of preparing information and the way that it very well may be expanded. Considering huge amounts of information additionally result choice to execute the guide diminish rendition of Naïve Bayes calculation

2.5 “PYTHON THE FASTEST GROWING PROGRAMMING LANGUAGE”

This paper started with introduction of python as a high state programming. PROGRAMMING Why python is growing so rapidly is discussed next because of its attributes like mobility, simple to learn, open source, etc. Further its drawbacks are discussed like its moderate nature and besides it is hard to keep up. Various activities have been recorded in python and it has been used in Irobot, Google, and Intel, etc.

2.6 "MACHINE LEARNING ALGORITHMS: A REVIEW"

In this investigation paper, distinctive AI estimations have been discussed. This paper gives quick and dirty illumination of the classes of the computations. In the coordinated learning arrangement, three counts have been inspected for example gullible bayes, bolster vector machine and choice tree. In the unsupervised learning, two algorithms have been discussed for example K-implies grouping and vital part examination.

2.7 "AN OVERVIEW OF DEEP LEARNING"

Deep learning approaches are essential for us to take care of numerous issues. In this paper, we present deep learning models and structures in detail. Deep learning various types of models and structures, and it has had numerous applications in numerous perspectives. From these, we can see that deep learning has an incredible advancement potential. In future, it is predictable that deep learning could set up ideal speculations to clarify its exhibitions. In the meantime, its capacities of unsupervised learning will be improved since there are the great many information on the planet however it isn't relevant to add names to every one of them. It is additionally anticipated that neural system structures will turn out to be increasingly unpredictable with the goal that they can separate all the more semantically important highlights. In addition, profound learning will consolidate with support adapting better and we can utilize these points of interest to achieve more assignments

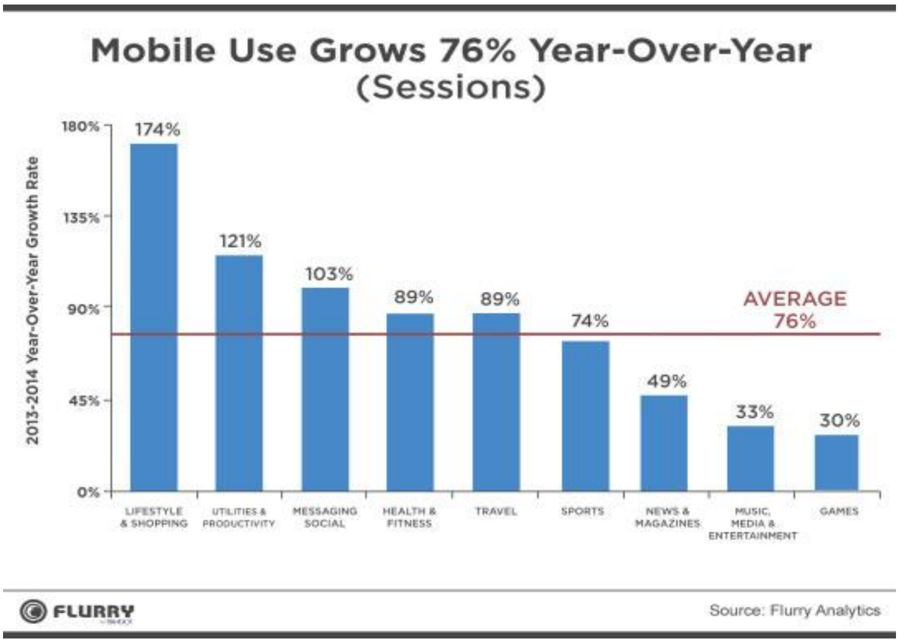
2.8 "THE POWER OF MOBILE APPLICATIONS”

The importance of cell phones and applications has been developing massively for past years as far as use measurements, download checks and number of applications on the business sectors. As indicated by Flurry Analytics (2015), portable application utilization developed by %76 in 2014

Mobile applications have become an integral part of our lives. They have revolutionized the way we interact with technology and have made our lives more convenient. With the advent of low-code platforms like **Microsoft Power Apps**, creating mobile applications has become easier than ever before .

Mobile applications offer several benefits to businesses, including the ability to work on the go and in the field, simplifying processes, and building better customer relationships [2]. A successful mobile application should provide an excellent user experience, be easy to use, and bring data together to help improve customer experiences .

In addition to the benefits of mobile applications, there are also technical challenges that need to be addressed. For example, running deep convolutional neural networks (CNNs) for complex tasks such as ImageNet classification on mobile devices is challenging .

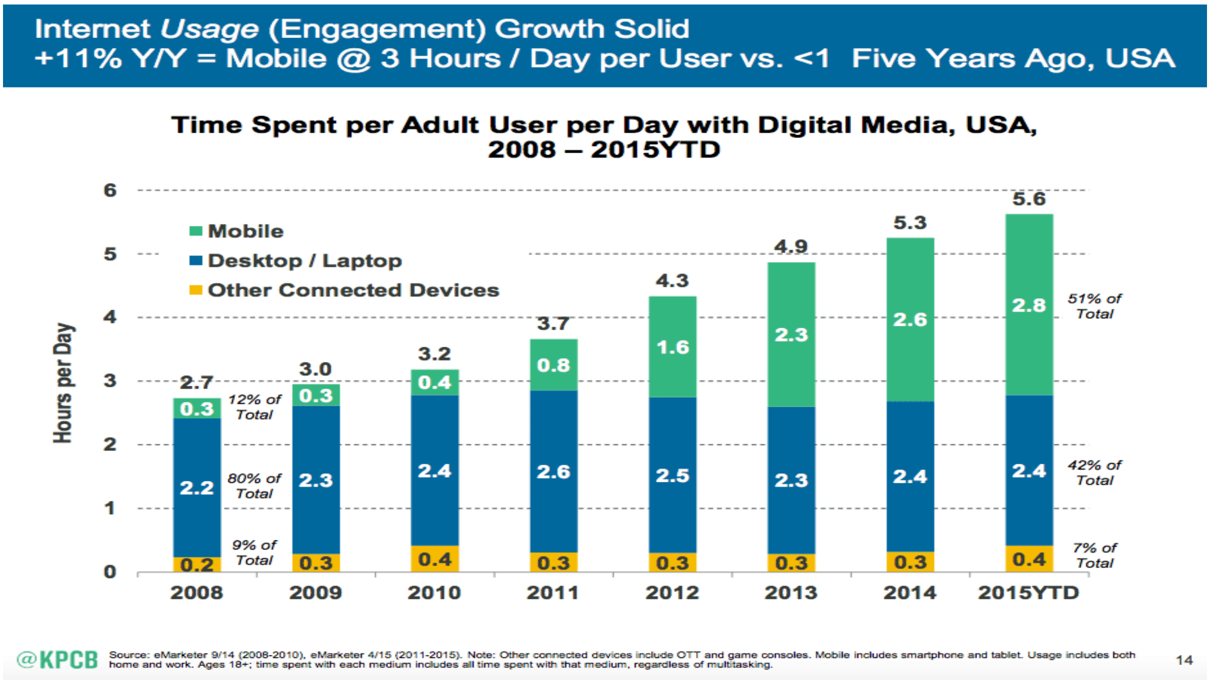


Mobile Use Grows 76% Year-Over-Year (Sessions)

Source: Flurry Analytics

Graph 2: Mobile Use Grows Year-Over-Year [8]

In addition, late insights uncovers that versatile media use overwhelms most of absolute media utilization. Kleiner Perkins Caufield and Byers' report (2015) demonstrates that time spent by clients on versatile advanced media is the most astounding contrasted with others 3.0 80% of 2.3 89% 2500 HEALTH & FITNESS 0.3 2009 89%



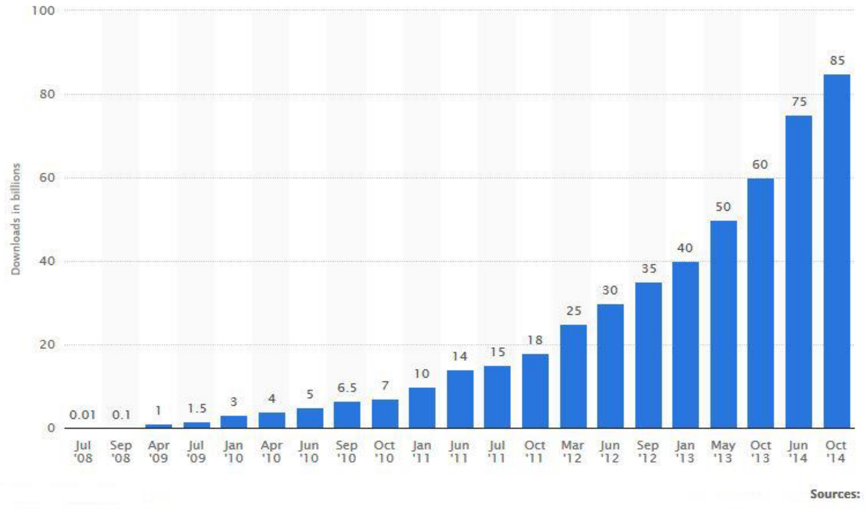
Internet *Usage* (Engagement) Growth Solid
+11% Y/Y = Mobile @ 3 Hours / Day per User vs. <1 Five Years Ago, USA

Time Spent per Adult User per Day with Digital Media, USA, 2008 – 2015YTD

Source: eMarketer 9/14 (2008-2010), eMarketer 4/15 (2011-2015). Note: Other connected devices include OTT and game consoles. Mobile includes smartphone and tablet. Usage includes both home and work. Ages 18+; time spent with each medium includes all time spent with that medium, regardless of multitasking.

Graph 3: Time spent per day with digital media [8]

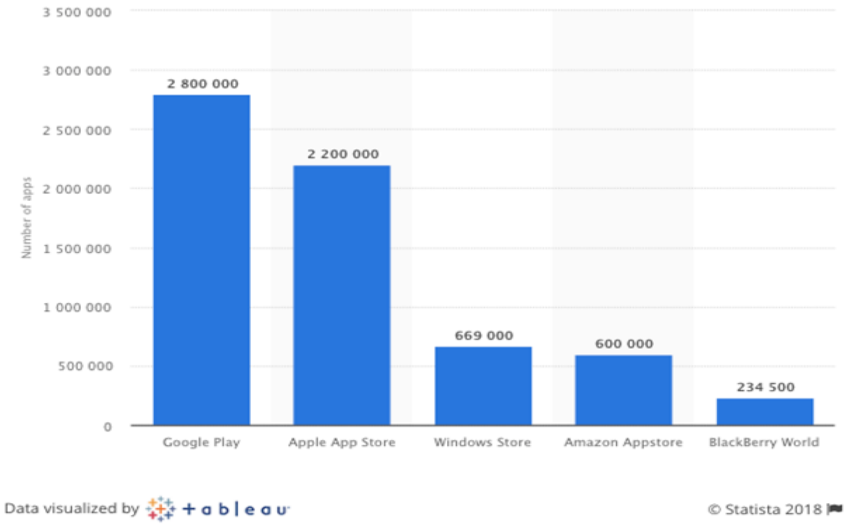
App as of late reported that the App Store passed 100 billion application Apple downloads since it is opened in 2008 (CNET, 2015). Apple's CEO Tim Cook made the declaration at Apple's Worldwide Developers Conference (WWDC) while he noticed that it has been paid $30 billion dollars to application designers by Apple, the realistic by Statista (2013) demonstrates the development in number of downloads in App Store from 2008 to 2015. Criminative number of apps downloaded from the Apple App Store from July 2008 to June 2015 (in billions) 37 M19 26 Graph 4: Number of Apps Downloaded [S] 02 Statista Increment in the fame of versatile condition makes the field focused for individual engineers, organizations and merchants (for example Apple, Google, Windows, Amazon and so forth.). Since engineers need to get required to portable segment, the quantity of applications in stores increments and make the challenge increasingly extreme. The quantity of accessible applications for July, 2015 is appeared in the Figure 4 which uncovers that there are in excess of 3 million applications in Google Play and Apple App Store (Statista, 2015).



Sources:

© Statista 2015

Graph 4: Number of apps downloaded



Graph 5: Number of total apps in mobile markets

This circumstance causes issues for the two designers and buyers. For customer's case, the principle issue is essentially to pick right application for the correct reason in a great many applications. Then again, engineers ought to keep up solid input procedure to improve their applications, make new features and defeat the shortcomings (Chen et al., 2014) so asto draw in more clients. Regardless of other programming circulation channels, versatile stores offer clients to capacity to rate applications and make remarks.

2.9 Mobile App Analytics

Bitterer (2011) portrayed Mobile Business Intelligence (Mobile BI) as one of the new innovations which have the capability of disturbing the Business Intelligence (BI) advertises. Thinking about the intensity of versatile promotions (Snider, 2012), the nature of the portable environment, which offers capacity to gather customized and area explicit substance, has given significant open doors for Business Intelligence and Analytics (Chen et. al, 2012). One of the significant information which versatile application investigation decides to process is "Star Ratings" of utilizations. Star evaluations which exhibit the normal voted rating of the applications can impact incomes and point of confinement the rate of development (Vasa et. al., 2012). So as to exist in this exceedingly aggressive market, engineers ought to demonstrate the nature of their applications with high application evaluations (Khalid, 2013).

As it is expressed by Chevalier and Myzalin (2006), he said that review of the customer has a great impact on sales of particular application. He also said that review contain some important information including functionality, about failure of apps, weakness of user interface, bugs at the time of update etc. The main motive of the developers is how to respond those feedbacks so that all those vulnerabilities can be removed easily from the particular app. Hence, there is a requirement for mechanized arrangements dissecting surveys and changing to instructive information.

Google Analytics for Mobile Apps is a free tool that provides easy-to-use SDKs and reports designed with app developers in mind. It enables developers to understand the number of users in their app, their characteristics, and where they come from. It also helps to measure what actions users are taking, in-app payments, revenue, and visualize user navigation paths .

SYSTEM DEVELOPMENT

SYSTEM DEVELOPMENT

3.1 DESIGNING

DATA MODELS

In machine learning, a data model is an algorithm that can identify patterns or make predictions on unseen datasets. These models are the mathematical engines of artificial intelligence .

There are many types of machines learning models, including regression and classification algorithms [3]. Regression algorithms are used to predict a continuous outcome using independent variables, while classification models are used to predict the probability of an event

TABLES

Tables are an essential component of a data model. They are used to organize and store data in a structured manner [1]. In a data model, tables are used to represent entities, and the relationships between these entities are represented by the tables' attributes [2].

In the context of machine learning, tables are used to represent datasets. Each row in the table represents an instance of the dataset, while each column represents a feature of the dataset [3].

If you are looking for information on how to use tables in Excel's data model, you can use Cube functions to access data in the data model

DEEP NEURAL NETWORK MODELS

Machine learning techniques have been widely applied in various areas such as pattern recognition, natural language processing, and computational learning. During the past decades, machine learning has brought enormous influence on our daily life with examples including efficient web search, self-driving systems, computer vision, and optical character recognition (OCR). Especially, deep neural network models have become a powerful tool for machine learning and artificial intelligence. A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. Note that the terms ANN vs. DNN are often incorrectly confused or used interchangeably. The success of deep neural networks has led to breakthroughs such as reducing word error rates in speech recognition by 30% over traditional approaches (the biggest gain in 20 years) or drastically cutting the error rate in an image recognition competition since 2011 .

Neuron receives multiple signals through the synapses contacting its dendrites and sends a single stream of action potentials out through its axon. The complexity of multiple inputs is reduced by categorizing its input patterns. Inspired by this intuition, artificial neural network models are composed of units that combine multiple inputs and produce a single output.

3.2 ARCHITECTURE

3.2.1 MACHINE LEARNING ARCHITECTURE

Machine Learning architecture is defined as the subject that has evolved from the concept of fantasy to the proof of reality. As earlier machine learning approach for pattern recognitions has lead foundation for the upcoming major artificial intelligence program. Based upon the different algorithm that is used on the training data machine learning architecture is categorized into three types i.e. Supervised Learning, Unsupervised Learning, and Reinforcement Learning and the process involved in this architecture are Data Aquisition, Data Processing, Model Engineering, Excursion, and Deployment

Architecting the Machine Learning Process

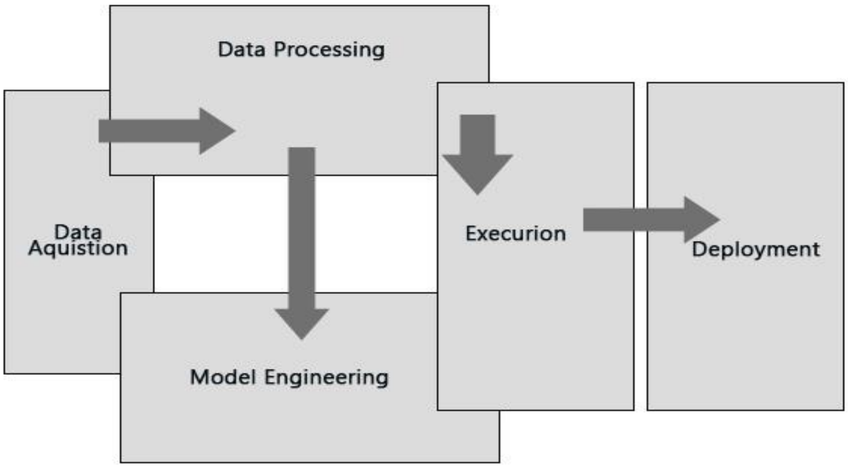


Figure 10: Apache Machine learning Architecture

1. Data Acquisition

As machine learning is based on available data for the system to make a decision hence the first step defined in the architecture is data acquisition. This involves data collection, preparing and segregating the case scenarios based on certain features involved with the decision making cycle and forwarding the data to the processing unit for carrying out further categorization.

2. Data Processing

The subjected to advanced integration and processing and involves normalization of the data, data cleaning, transformation, and encoding. The data processing is also dependent on the type of learning being used. For e.g., if supervised learning is being used the data shall be needed to be segregated into multiple steps of sample data required for training of the system and the data thus created is called training sample data or simply training data.

3. Data Modelling

This layer of the architecture involves the selection of different algorithms that might adapt the system to address the problem for which the learning is being devised, These algorithms are being evolved or being inherited from a set of libraries. The algorithms are used to model the data accordingly, this

4. Execution.

This stage in machine learning is where the experimentation is done, testing is involved and tunings are performed. The general goal behind being to optimize the algorithm in order to extract the required machine outcome and maximize the system performance, The output of the step is a refined solution capable of providing the required data for the machine to make decisions.

5. Deployment

Like any other software output, ML outputs need to be operationalized or be forwarded for further exploratory processing. The output Cavan be considered as a non-deterministic query which needs to be further deployed into the decision-making system.

3.2.2 HDFS ARCHITECTURE

Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware devices. It is used for storing large data in a Hadoop cluster. HDFS has a master/slave architecture, consisting of a single Name Node and a number of Data Nodes. The Name Node is the master server that manages the file system namespace and regulates access to files by clients. Data Nodes, usually one per node in the cluster, manage storage attached to the nodes that they run on.

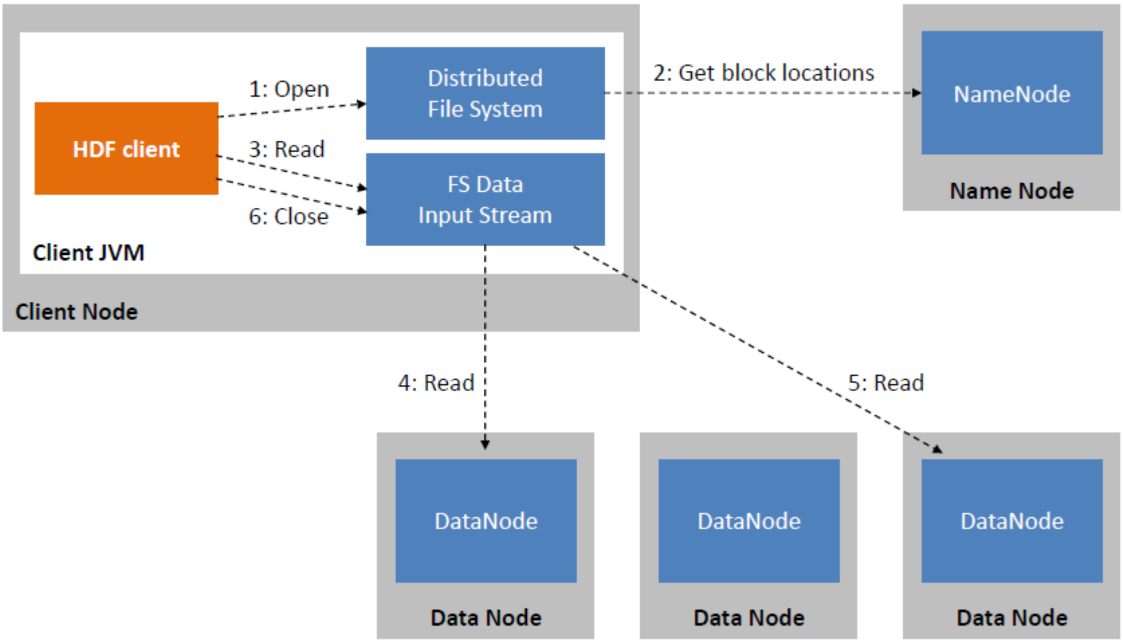


Figure 11: HDFS Architecture

3.3 DATA FLOW

3.3.1 DATA FLOW IN MACHINE LEARNING

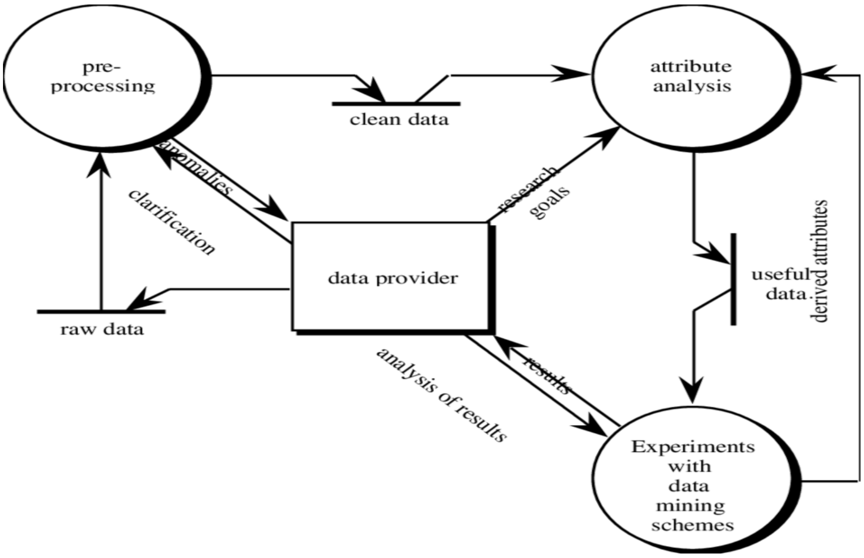


Figure 12: Data flow diagram

3.3.2 DATA FLOW IN DEEP NEURAL NETWORK

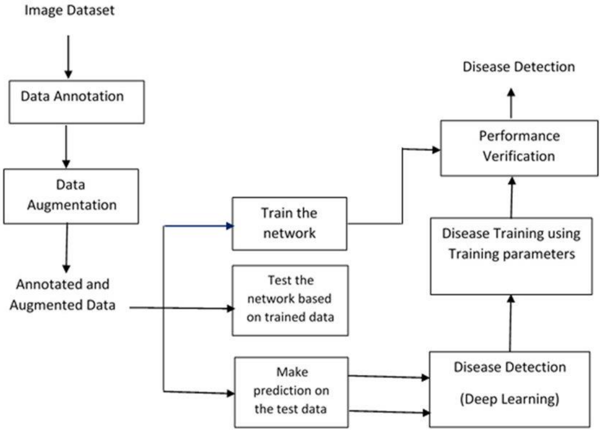


Image Dataset
Data Annotation
Data Augmentation
Annotated and Augmented Data
Train the network
Test the network based on trained data
Make prediction on the test data
Disease Detection (Deep Learning)
Disease Training using Training parameters
Performance Verification
Disease Detection

Figure 13: data flow in deep neural network

3.4 REQUIREMENTS

3.4.1 HARDWARE REQUIREMENTS

Types of Hardware used

Processor -multiprocessor-based with a 2.00 GHz 64-bit

Hardware - Quard core Intel i3 has 4 GB RAM

Memory - 4 GB (2.4 GB on virtual machine)

Disk space -8 GB free disk space.

Requirements increase as data is gathered and stored in

HDFS.

3.4.2 SOFTWARE REQUIREMENTS:

Type of Software Versions

Operating System Linux (Ubuntu)/Windows

VM ware workstation 14.1.2

Horton works sandbox HDP 1.2.4

Hadoop 2.7.3

NetBeans IDE 4.8

Web browser Mozilla Firebox 28

Python Jupiter 4.3.0.

3.4.3 ADVABTAGES OF VMWARE WORKSTATIONS:

1.ost-effective virtualization option in terms of pricing

2.Saves money on computing

3.Reduces downtime and increases the uptime of your system

4.DatCa backup and recovery

5.Boosts profit

6.Easy to customize the machine operating system with RAM, Hard Disk, CPU, etc.

7.Allows anyone to create a virtual machine in order to prevent viruses

8.Can be used to test a wide range of purposes

9.Supports multiple operating systems

10.Easy to add and delete virtual machines

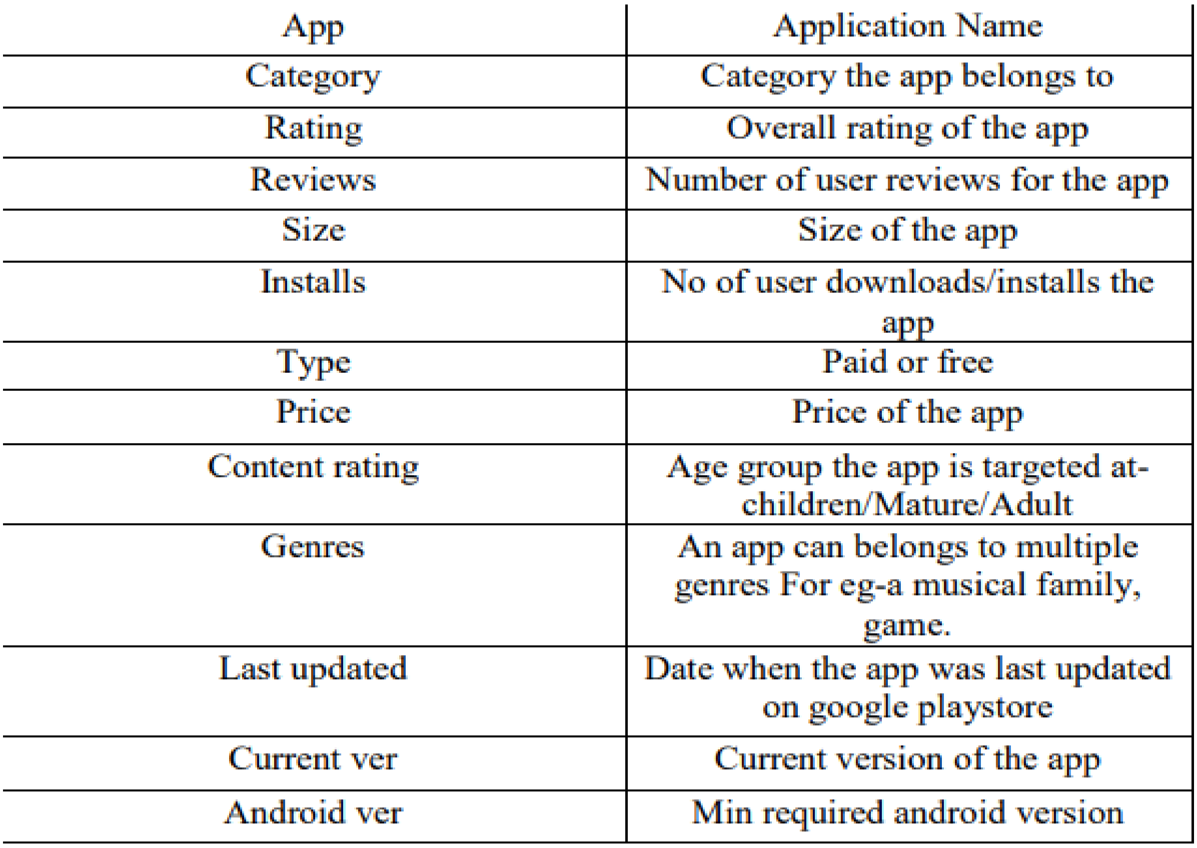
3.5 TEST PLAN

3.5.1 DATA SET

A data set is a structured collection of data organized in a tabular format, typically with rows representing individual instances or observations and columns representing attributes or features of those instances.it is a fundamental component in data science, machine learning, and statistical analysis. A data set is organized into some type of data structure. In a database, for example, a data set might contain a collection of business data (names, salaries, contact information, sales figures, and so forth). The database itself can be considered a data set, as can bodies of data within it related to a particular type of information, such as sales data for a particular corporate department.The term data set originated with IBM, where its meaning was similar to that of file. In an IBM mainframe operating system, a data set s a named collection of data that contains individual data units organized (formatted) in a specific, IBM-prescribed way and accessed by a specific access method based on the data set organization. Types of data set organization include sequential, relative sequential, indexed sequential, and partitioned. Access methods include the Virtual Sequential Access Method (VSAM) and the Indexed Sequential Access Method (ISAM).

google play store. CSV

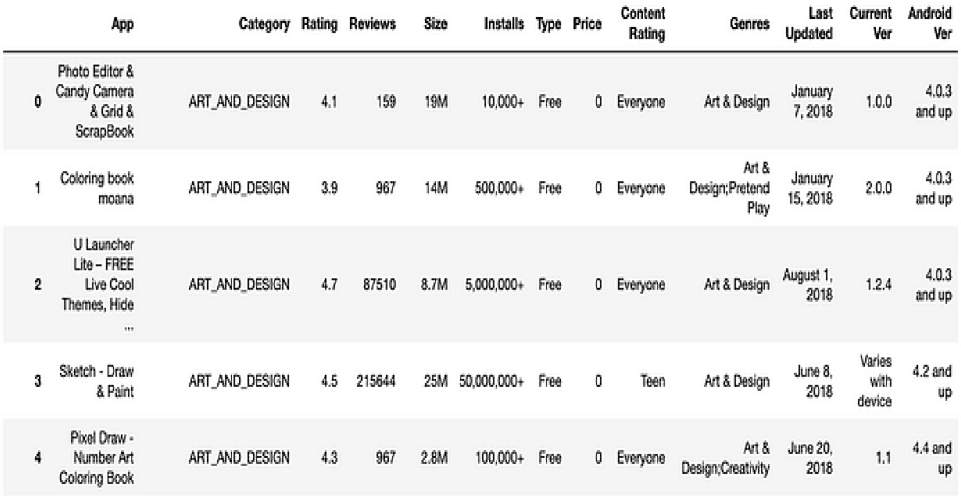
This data set is for web scratched information of 10k play store application for examining the android advertises .This is the offline dataset which can be utilized by a client to have the android market of different utilization of various classes music ,camera etc. with the assistance of this, client can foresee whether any given application will get lower or higher rating level, which app is free or which is paid, about ht e reviews, about the latest version of the particular app and many more things to analyse .This dataset can be additionally utilized for further references for the suggestion of any application .In addition ,this offline dataset is chosen in order to decide the forecast precisely as online information gets refreshed all around much of the time .likewise, to monitor its old clients to comprehend their inclinations better.



App	Application Name
Category	Category the app belongs to
Rating	Overall rating of the app
Reviews	Number of user reviews for the app
Size	Size of the app
Installs	No of user downloads/installs the app
Type	Paid or free
Price	Price of the app
Content rating	Age group the app is targeted at-children/Mature/Adult
Genres	An app can belongs to multiple genres For eg-a musical family, game.
Last updated	Date when the app was last updated on google playstore
Current ver	Current version of the app
Android ver	Min required android version

Table 1: Dataset columns and its specifications

Sample record from the offline-dataset (google play store.csv) is shown in the following table:



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Google play store_user_reviews.csv

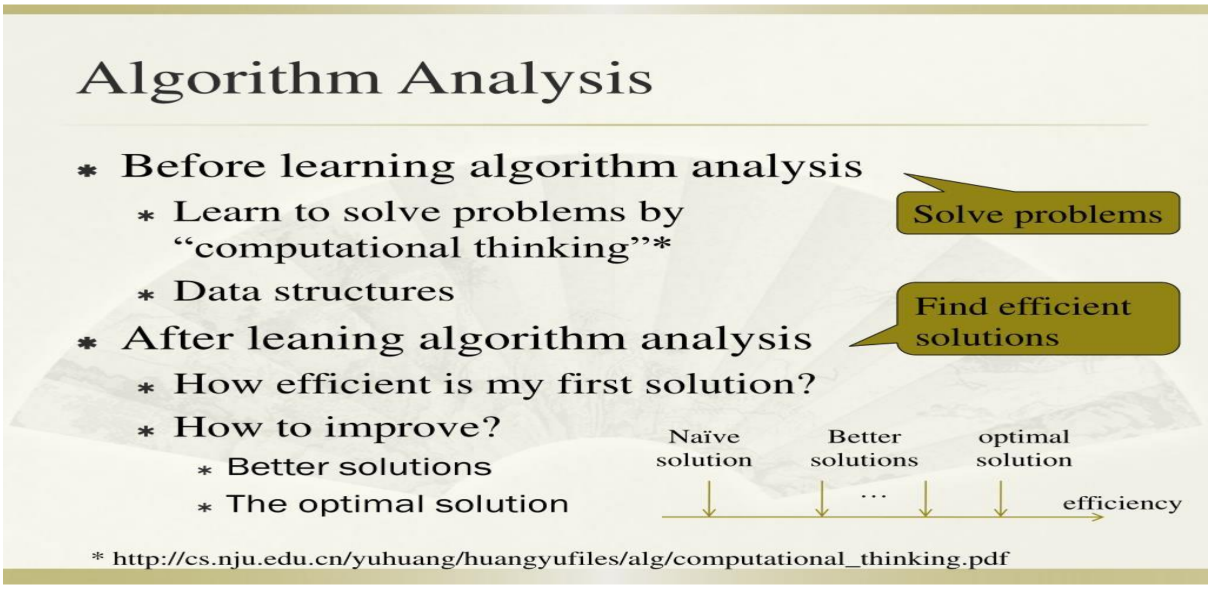
"Google play store_user_reviews.csv" likely refers to a CSV (Comma-Separated Values) file that contains data related to user reviews on the Google Play Store. In this context, the file could include structured information about the reviews that users have submitted for various apps available on the Google Play Store.

App	Name of the app
Translated review	User review
Sentiment	Positive/negative/neutral
Sentiment_ polarity	Sentiment polarity score
Sentiment_subjectivity	Sentiment subjectivity score

Table 2:Dataset columns and its specifications

3.6 ALGORITHMS

3.6.1 ALGORITHM USED FOR ANALYSIS PART



Algorithm Analysis

- Before learning algorithm analysis
  - Learn to solve problems by "computational thinking"*
  - Data structures
- After leaning algorithm analysis
  - How efficient is my first solution?
  - How to improve?
    - Better solutions
    - The optimal solution

Solve problems

Find efficient solutions

Naïve solution

Better solutions

optimal solution

efficiency

* http://cs.nju.edu.cn/yuhuang/huangyufiles/alg/computational_thinking.pdf

3.6.1.1 MAP REDUCE ALGORITHM

A MapReduce is a data processing tool which is used to process the data parallelly in a distributed form. It was developed in 2004, on the basis of paper titled as "MapReduce: Simplified Data Processing on Large Clusters," published by Google.

The MapReduce is a paradigm which has two phases, the mapper phase, and the reducer phase. In the Mapper, the input is given in the form of a key-value pair. The output of Mapper is fed to the reducer as input. The reducer runs only after the Mapper is

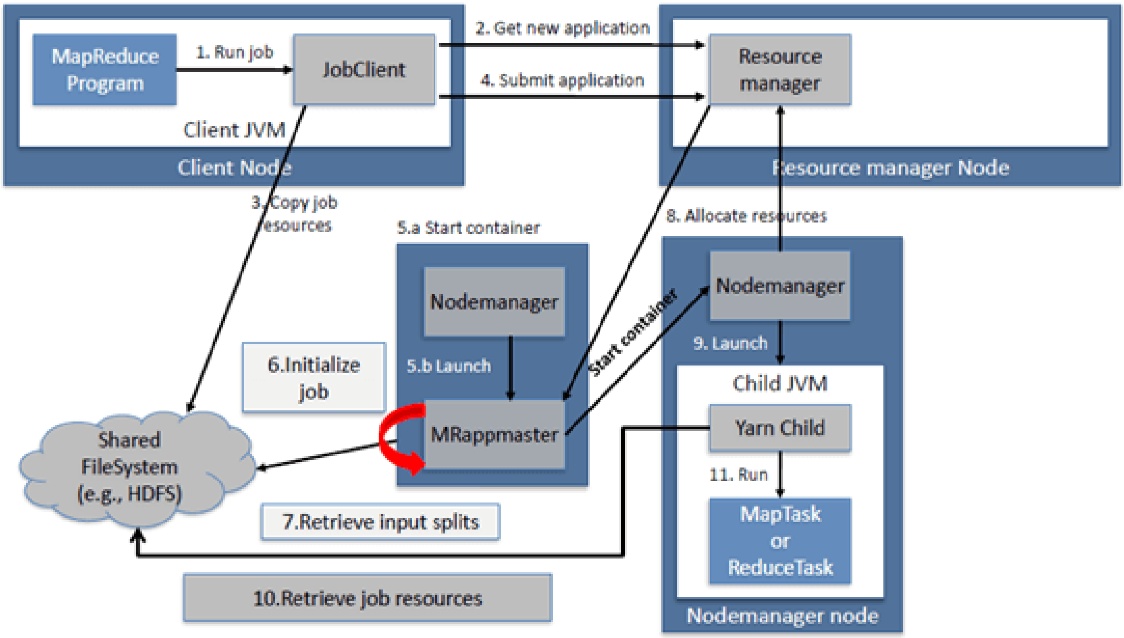


Figure 14: Map reduce algorithm

3.6.2 ALGORITHM USED FOR PREDICTION PART

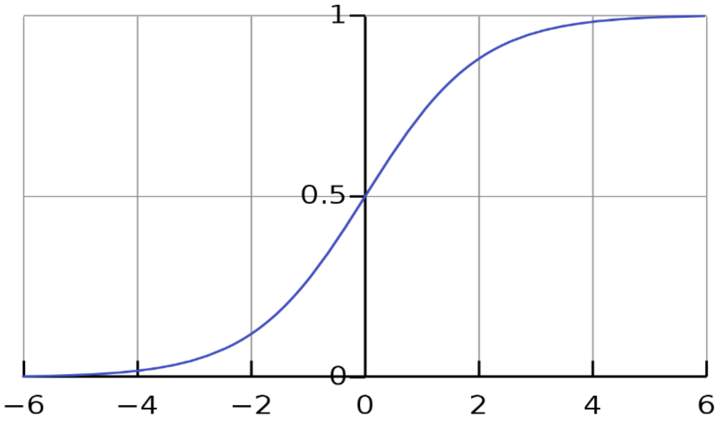
3.6.2.1 BACKPROPAGATION:

Backpropagation is one of the important concepts of a neural network. Our task is to classify our data best. For this, we have to update the weights of parameter and bias, but how can we do that in a deep neural network? In the linear regression model, we use gradient descent to optimize the parameter. Similarly here we also use gradient descent algorithm using Backpropagation.

For a single training example, Backpropagation algorithm calculates the gradient of the error function. Backpropagation can be written as a function of the neural network. Backpropagation algorithms are a set of methods used to efficiently train artificial neural networks following a gradient descent approach which exploits the chain rule.

The main features of Backpropagation are the iterative, recursive and efficient method through which it calculates the updated weight to improve the network until it is not able to perform the task for which it is being trained. Derivatives of the activation function to be known at network design time is required to Backpropagation.

Now, how error function is used in Backpropagation and how Backpropagation works? Let start with an example and do it mathematically to understand how exactly updates the weight using Backpropagation.



Graph 6: sigmoid function

The Back propagation algorithm in neural network computes the gradient of the loss function for a single weight by the chain rule. It efficiently computes one layer at a time, unlike a native direct computation. It computes the gradient, but it does not define how the gradient is used. It generalizes the computation in the delta rule.

FEED FORWARD

A feedforward neural network is an artificial neural network where the nodes never form a cycle. This kind of neural network has an input layer, hidden layers, and an output layer. It is the first and simplest type of artificial neural network.

In the feed-forward neural network, there are not any feedback loops or connections in the network. Here is simply an input layer, a hidden layer, and an output layer.

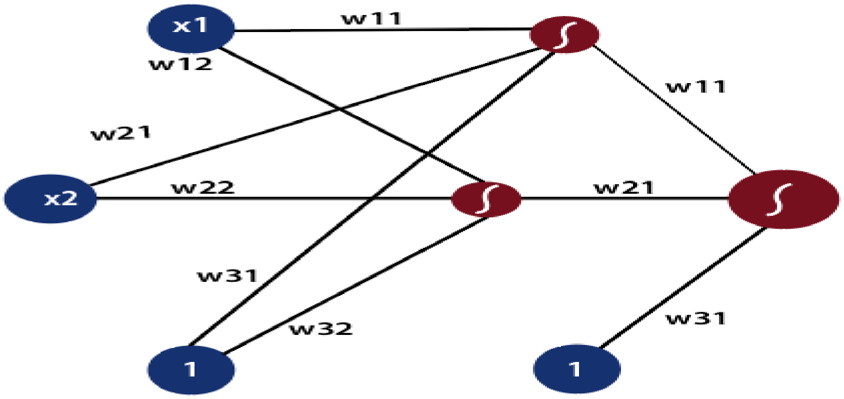


Figure 15: feed forward

There can be multiple hidden layers which depend on what kind of data you are dealing with. The number of hidden layers is known as the depth of the neural network. The deep neural network can learn from more functions. Input layer first provides the neural network with data and the output layer then make predictions on that data which is based on a series of functions. ReLU Function is the most commonly used activation function in the deep neural network.

IMPLEMENTATION

There are six modules in the project are:

1. Descriptive Analysis
2. Exploratory Data Analysis
3. Data Preprocessing and Visualization
4. Feature Engineering
5. Model building and Evaluation
6. Model Deployment

1. **Descriptive Analysis**:

They are biologically inspired algorithms that have several neurons like units arranged in layers. The units in neural networks are connected and are called nodes. Data enters the network at the point of input, seeps through every layer before reaching the output.

2. **Exploratory Data Analysis**:

Exploratory Data Analysis (EDA) refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables. EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling

3. **Data Preprocessing and Visualization**:

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.

4. **Feature Engineering**:

Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictivemodel using machine learning or statistical Modeling.

5. **Model building and Evaluation**:

After shortlisting the models, we first split the data into train, validation, and test set. Now we select a simple base model from these shortlisted models. Now we train this base model on the train set and fine-tune it by adjusting its hyperparameters using the train and the validation set. Performance is mastered using performance metrics, such as accuracy, precision, recall, etc. After the final tuning, we log the performance of the model.

6. **Model Deployment**:

Model deployment is considered to be a challenging stage for data scientists. This is because it is often not considered their core responsibility, and due to the technological and mindset differences between model development and training and the organizational tech stack, like versioning, testing and scaling which make deployment difficult. These organizational and technological silos can be overcome with the right model deployment frameworks, tools and processes.

LIST OF ALGORITHMS:

List of algorithms used in implementation of our experiment are:

1. Support Vector Machine
2. K-nearest algorithm
3. Artificial Neural Network
4. Random Forest algorithm
5. Logistic regression algorithm

1.SUPPORT VECTOR MACHINE:

Pseudocode:

```
 Importing the necessary packages Examples: import pandas as pd
 def SVM
Step 1: START
Step 2: Reading the dataset.pd.read.csv (file name) #reads the dataset file
Step 3: Data cleaning and preprocessing of data
 Resampling the data as normal and fraud class i.e normal=0 and fraud=1 under
 Under sampling of data is done.
 Data is scaled (if any value then eliminated) and normalized.
 Dataset is spitted into two set as train data and test data using split() on training data is used to split the data.
Step 4: Training the data using the SVM algorithm
 SVM classifier is called as classifier. Predict () #which predicts whether transaction fraud or non-fraud.
Step 5: Calculating the fraud transactions and valid transactions, then calculating the recall, precision and accuracy and stored in the respective locations.
Step 6: STOP
```

2. K-NEAREST NEIGHBOUR

Pseudocode:

```
Step 1: START
Step 2: Loading of dataset pd read.csv (csv file) # reads the csv file and loads
Step 3: cleaning and normalization of data
Normal=0
Fraud=1#resampling. Data is scaled and normalized
Train_test_split () #splitting of dataset into train and test data
Step 4: Train the model then fit the trained model
Trained the data using Knn classifier
K Neighbors Classifier () # Knn classifier which does classification of transactions
Step 5: Calculating the number of frauds, valid transactions and recall, precision and accuracy calculated.
Step 6: STOP
```

3. ARTIFICIAL NEURAL NETWORK

Pseudocode:

The ANN algorithm has two parts: Training part and testing part.

```
Training part: Def ANN
Step 1: START
Step 2: loading and observing the dataset
Pd read.csv(.csv) #reads the dataset
resampling of data
StandardScaler () #scaling and normalization of data
Step 3: Data preprocessing
Train_test_split ()#splitting of data
Step 4: Training the model
Dense () #Adding data to activation function
Step 5: Analyzing the model prediction of fraud is made and this trained data is stored. it can used to astest (training the model takes longer time so it is stored)
Step 6: STOP
Testing part:
```

Def ANN it is carried out similar way only difference is that the stored trained model is used to test the data and classify it

4. RANDOM FOREST TREE ALGORITHM:

Random Forest is a widely used classification and regression algorithm. As classification and regression are the most significant aspects of machine learning, we can say that the Random Forest Algorithm is one of the most important algorithms in machine learning. The capacity to correctly classify observations is helpful for various business applications, such as predicting whether; a specific user would buy a product or a loan will default or not. Classification algorithms in data science include logistic regression, support vector machines, naive Bayes classifiers, and decision trees. On the other hand, the random forest classifier is near the top of the classifier hierarchy. This article will deep dive into how a Random Forest classifier works with real-life examples and why the Random Forest is the most effective classification algorithm. Let&#39; s start with a basic definition of the Random Forest Algorithm. The Random Forest Algorithm is composed of different decision trees, each with the same nodes, but using different data that leads to different leaves. It merges the decisions of multiple decision trees in order to find an answer, which represents the average of all these decision trees. The random forest algorithm is a supervised learning model; it uses labeled data to “learn” how to classify unlabeled data. This is the opposite of the K-means Cluster algorithm, which we learned in a past article was an unsupervised learning model. The Random Forest Algorithm is used to solve both regression and classification problems, making it a diverse model that is widely used by engineers.

5. LOGISTIC REGRESSION ALGORITHM

Logistic Regression is one of the most used Machine Learning algorithms for binary classification. It is a simple Algorithm that you can use as a performance baseline, it is easy to implement and it will do well enough in many tasks. Therefore, every Machine Learning engineer should be familiar with its concepts. The building block concepts of Logistic Regression can also be helpful in deep learning while building neural networks. In this post, you will learn what Logistic Regression is, how it works, what are advantages and disadvantages and much more. These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the sigmoid function. The Sigmoid-Function is an S-shaped curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. This values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier.

UMLS

1.1 UML SOFTWARE DESIGN

A Unified Modeling Language (UML) diagram provides a visual representation of an aspect of a system. UML diagrams illustrate the quantifiable aspects of a system that can be described visually, such as relationships, behaviour, structure, and functionality. For example, a class diagram describes the structure of the system or the details of an implementation, while a sequence diagram shows the interaction between objects over time. In a UML diagram, the diagram elements visually represent the classifiers in a system or application. These classifiers are the diagrammatic representation of a source element. UML diagrams provide views of source elements; however, diagram elements do not have semantic value.

UML diagrams can help system architects and developers understand, collaborate on, and develop an application. High-level architects and managers can use UML diagrams to visualize an entire system or project and separate applications into smaller components for development. System developers can use UML diagrams to specify, visualize, and document applications, which can increase efficiency and improve their application design. UML diagrams can also help identify patterns of behaviour, which can provide opportunities for reuse and streamlined applications. The visual representation of a system that UML diagrams provide can offer both low-level and high-level insight into the concept and design of an application. You can use a wide variety of diagram types to model a system or application, based on the system, audience, and detail of the model you create. Depending on the diagram choice, you can select the detail and level of abstraction that the diagrams display. A typical UML model can consist of many different types of diagrams, with each diagram presenting a different view of the system that you are Modeling. Some examples of UML 2.1 and later diagrams include use case diagrams, state diagrams, sequence and communication diagrams, and topic and browse diagrams. Some UML 2.1 and later diagrams also allow the use of freeform and non-UML shapes. You can use freeform diagrams to represent general high-level views of a model or alternate model views that cannot be represented by standard UML notation. There are two types of freeform diagrams:

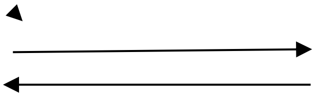
 the pure freeform diagram, which has no UML semantic context

 the UML diagrams that allow the use of freeform diagram elements, which are the use- case, class, component, and deployment diagrams.

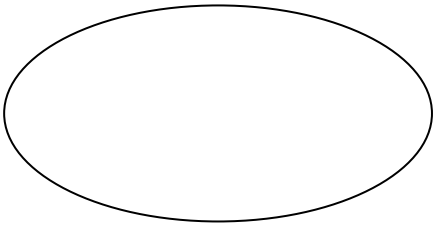
DATA FLOW DIAGRAMS

DFD is the abbreviation for Data Flow Diagram. The flow of data of a system or a process is represented by DFD. It also gives insight into the inputs and outputs of each entity and the process itself. DFD does not have control flow and no loops or decision rules are present. Specific operations depending on the type of data can be explained by a flowchart. It is a graphical tool, useful for communicating with users, managers and other personnel. it is useful for analyzing existing as well as proposed system. It should be pointed out that a DFD is not a flowchart. In drawing the DFD, the designer has to specify the major transforms in the path of the data flowing from the input to the output. DFDs can be hierarchically organized, which helps in progressively partitioning and analysing large systems.

1.Dataflow: Data move in a specific direction from an origin to a destination



2. **Process**: People, procedure, or devices that use or produce (Transform) data. The physical component is not identified.

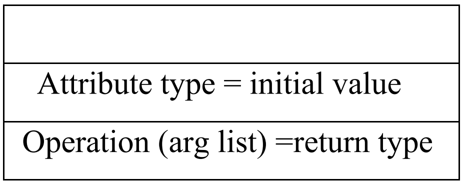


3.Source: External sources or destination of data, which may be people, program, organizations or other entity.

4. **Data Store**: Here data are stored or referenced by a process in the system.

CLASS DIAGRAM

Class diagrams are the backbone of almost every object-oriented method including UML. They describe the static structure of a system. Basic class Diagram Symbols and Notations Class represent an abstraction of entities with common characteristics. Associations represent the relationships between classes. Illustrate classes with rectangles divided into compartments. Place the name of the class in the first partition (cantered, bolded, and capitalized), list the attributes in the second partition, and write.


Attribute type = initial value
Operation (arg list) =return type

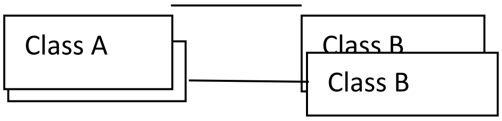
Active class

Active classes initiate and control the flow of activity, while passive classes store data and serve other classes. Illustrate active classes with a thicker border.

Active class

ASSOCIATIONS

In UML class diagrams, associations represent relationships between classes. These relationships define how instances of one class are related to instances of another class. Associations are crucial for understanding the structure of a system and the interactions between its components.



COMPOSITION AND AGGREGATION

Composition and aggregation are two types of association relationships in Unified Modeling Language (UML) class diagrams. Both represent a form of "whole-part" relationships between classes, indicating how one class is composed of or aggregated with other classes. However, they differ in terms of the strength and lifetime of the relationships

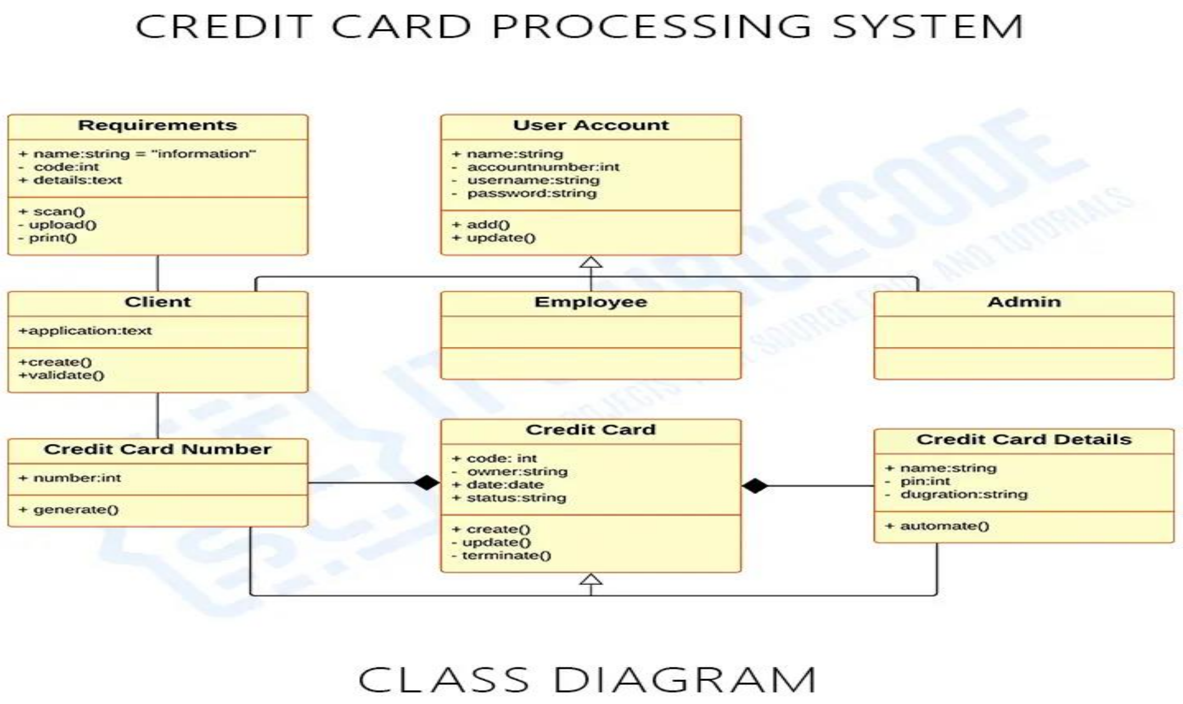


Figure 16: **Class Diagram on credit Card processing system**

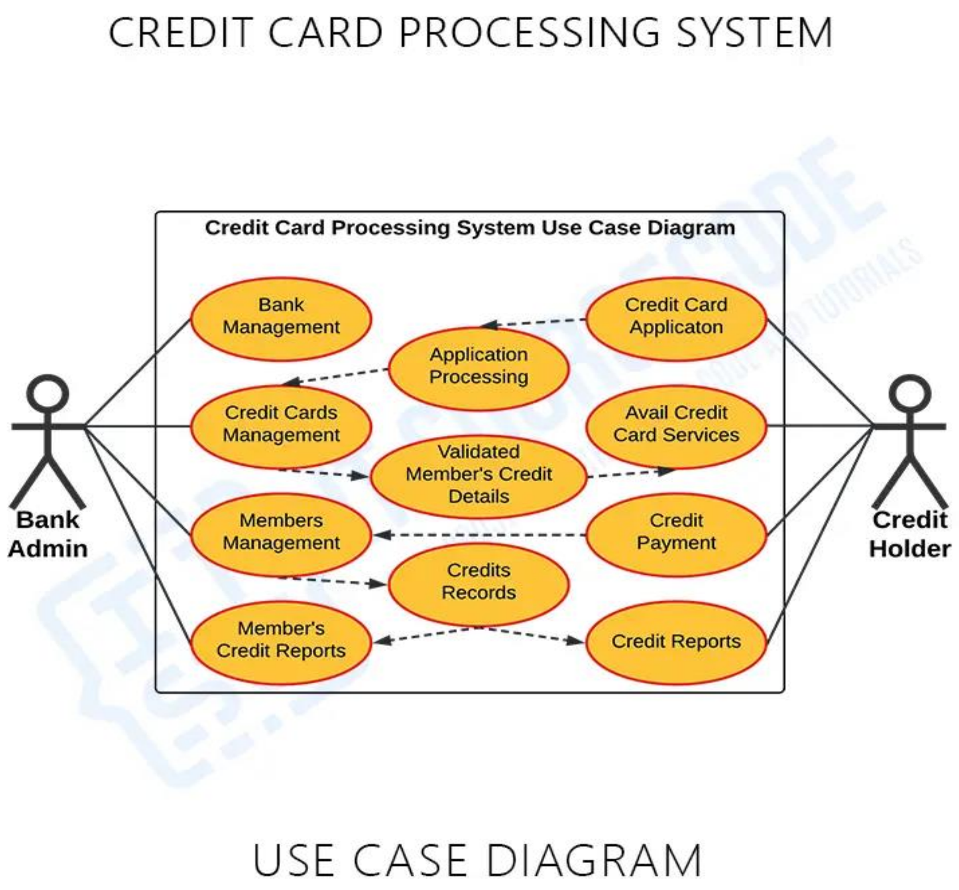
USE CASE DIAGRAM

Use case diagrams model the functionality of a system using actors and use cases.Use cases are services or functions provided by the System to its users.

BASIC USE CASE DIAGRAM SYMBOLS AND NOTATIONS:

System

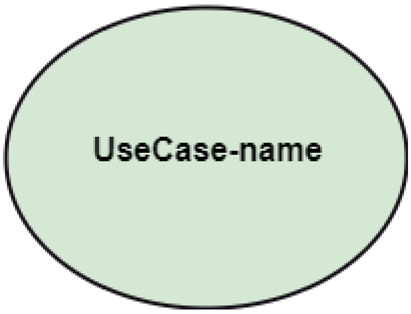
Draw your system's boundaries using a rectangle that contains use cases. Place actors outside the system's boundaries.



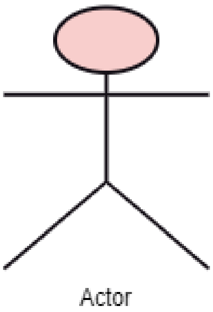
Use case diagram on credit card processing system

Use case:

Use case is the core concept of object-oriented Modeling. It portrays a set of actions executed by a system to achieve the goal.

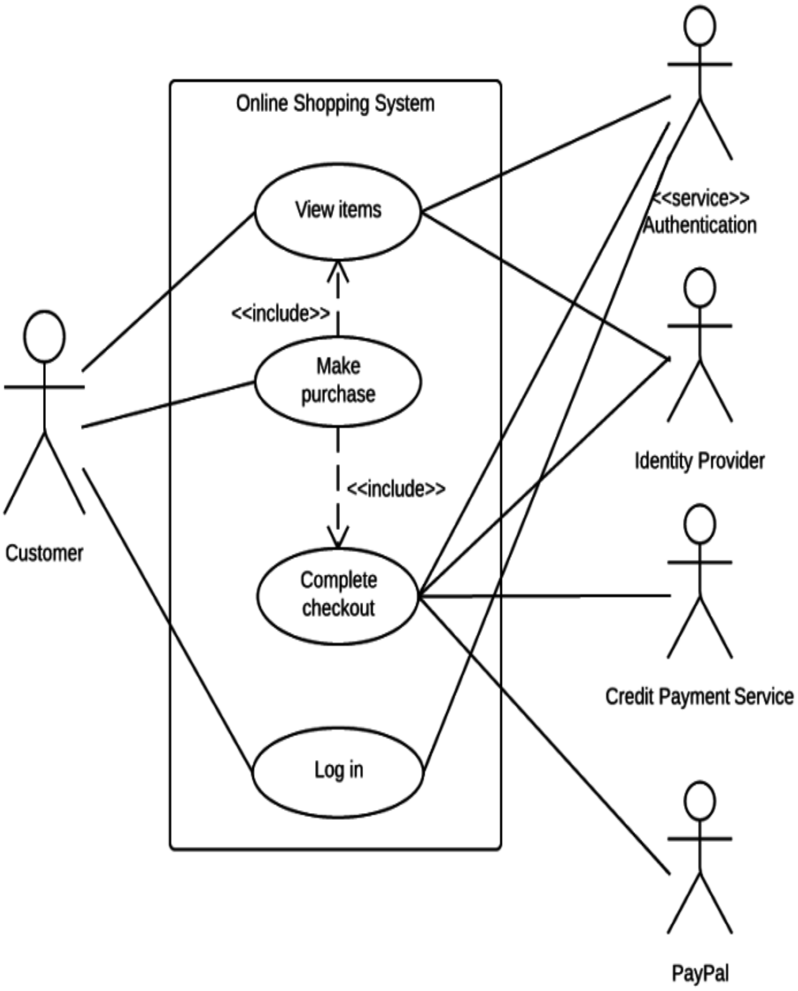


Actor: It comes under the use case diagrams. It is an object that interacts with the system, for example, a user.



Relationships

The relationship between two use cases basically models the dependencies between two use cases. By reusing existing use cases using different types of relationships, the overall effort required to develop the system is reduced. Use case diagrams show use cases, actors, and the relationships between them.



USE CASE DIAGRAM

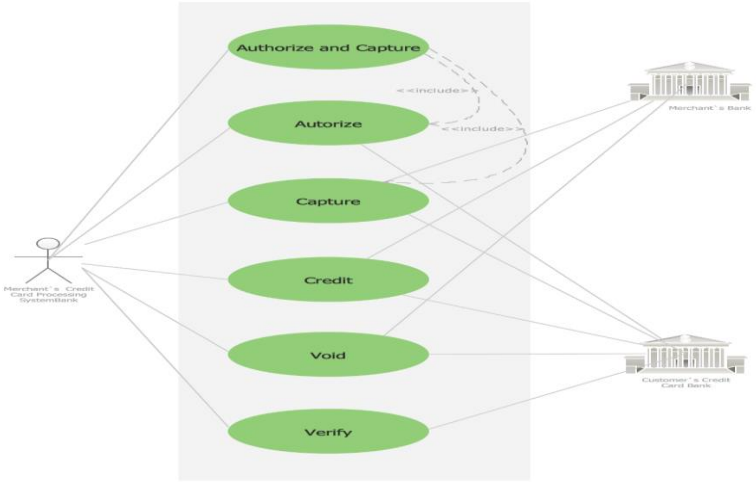


Figure 17: Use case diagram for credit card processing system

ACTIVITY DIAGRAM

An activity diagram is a type of Unified Modeling Language (UML) flowchart that shows the flow from one activity to another in a system or process. It's used to describe the different dynamic aspects of a system and is referred to as a 'behaviour diagram' because it describes what should happen in the modelled system.

Even very complex systems can be visualized by activity diagrams. As a result, activity diagrams are often used in business process Modeling or to describe the steps of a use case diagram within organizations. They show the individual steps in an activity and the order in which they are presented. They can also show the flow of data between activities.

Activity diagrams show the process from the start (the initial state) to the end (the final state). Each activity diagram includes an action, decision node, control flows, start node, and end node.

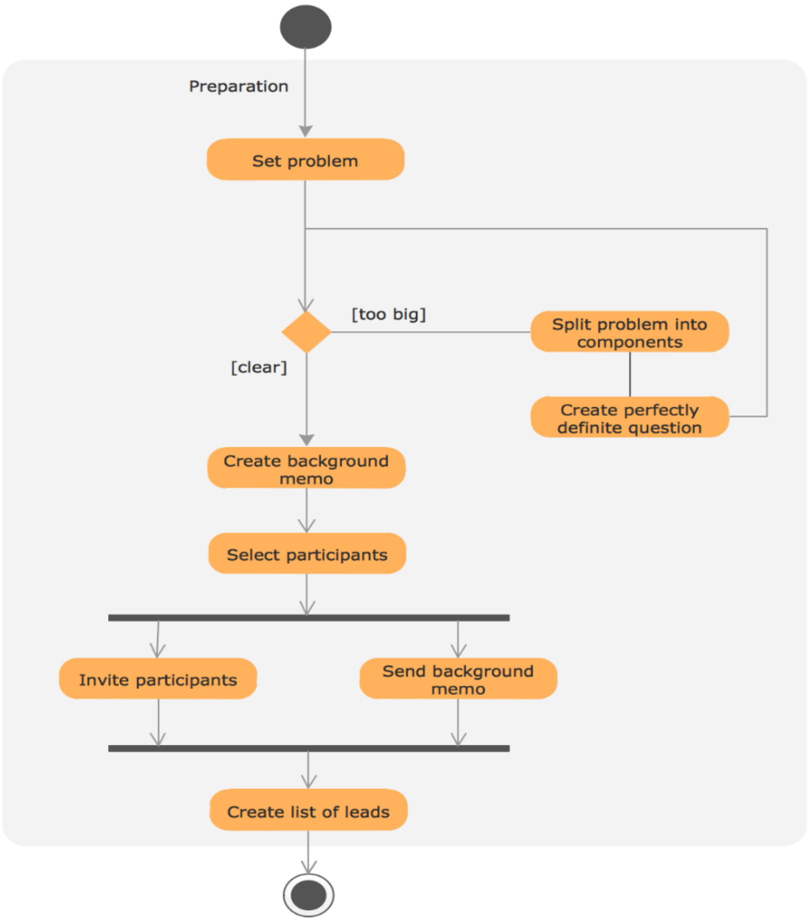


Figure 18: Activity diagram on credit card processing system

SEQUENCE DIAGRAM

_ Sequence Diagrams are interaction diagrams that detail how operations are carried out. They capture the interaction between objects in the context of a collaboration. Sequence Diagrams are time focus and they show the order of the interaction visually by using the vertical axis of the diagram to represent time what messages are sent and when. Model high-level interaction between active objects in a system

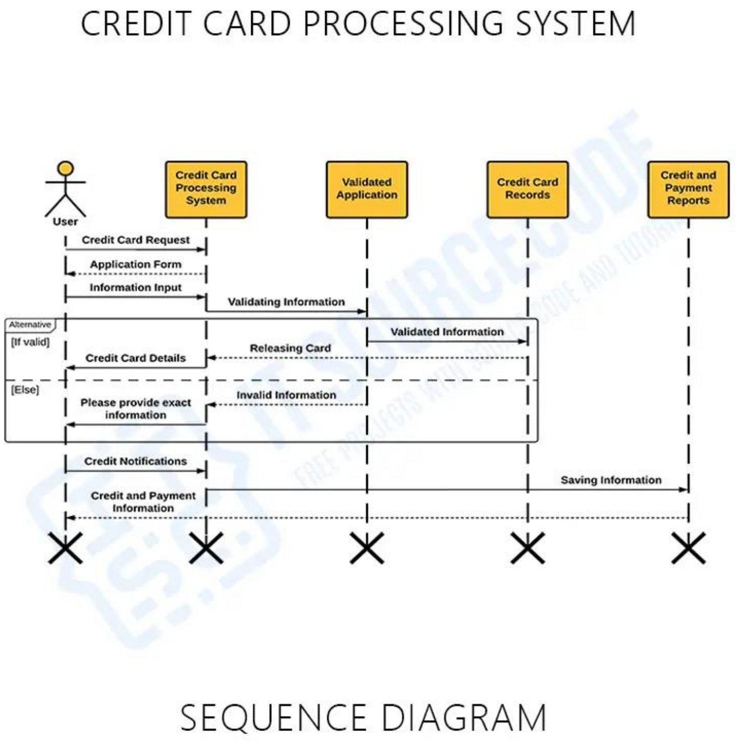


Figure 19: sequence diagram on credit card processing system

COMPONENT DIAGRAM

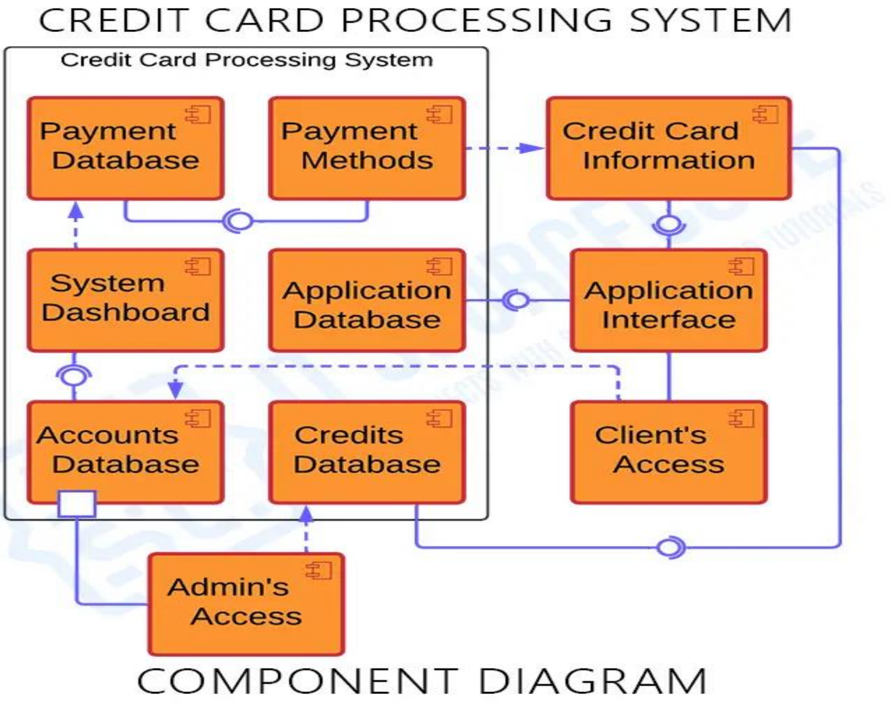


Figure 20: component diagram on credit card processing system

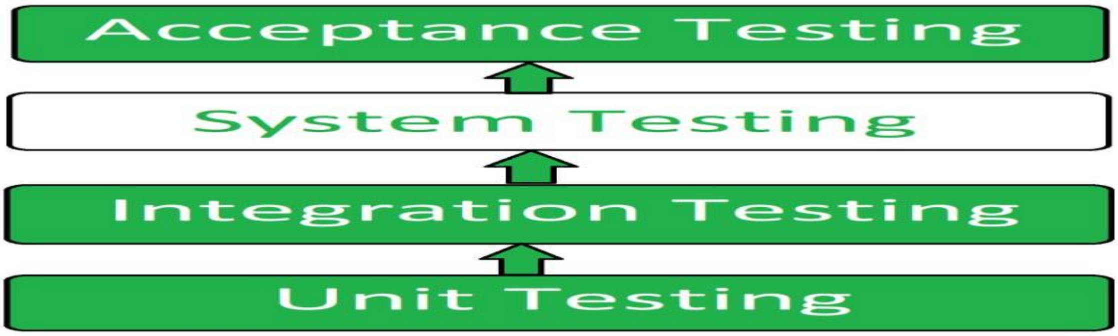
SYSTEM TESTING AND IMPLEMENTATION

4.SYSTEM TESTING AND IMPLEMENTATION

INTRODUCTION:

System testing is a type of software testing that evaluates the overall functionality and performance of a complete and fully integrated software solution. It tests if the system meets the specified requirements and if it is suitable for delivery to the end-users. This type of testing is performed after the integration testing and before the acceptance testing.

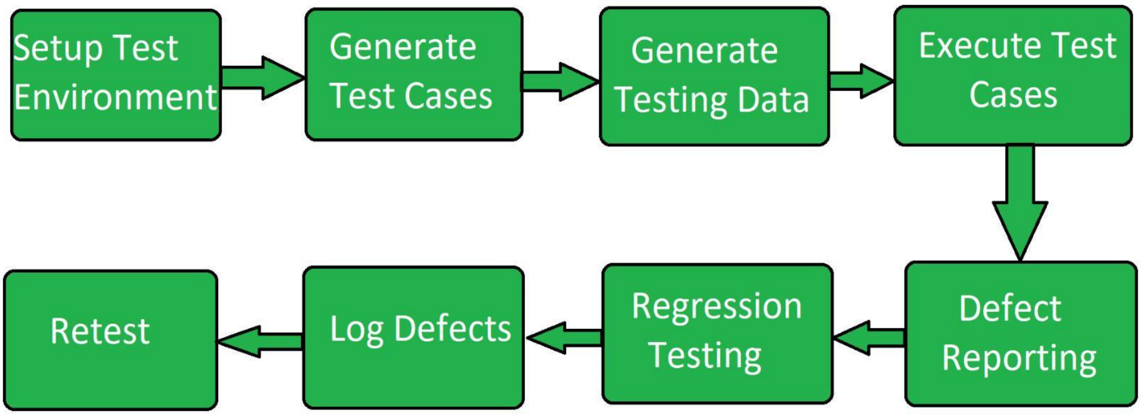
System Testing is a type of software testing that is performed on a complete integrated system to evaluate the compliance of the system with the corresponding requirements. In system testing, integration testing passed components are taken as input. The goal of integration testing is to detect any irregularity between the units that are integrated together. System testing detects defects within both the integrated units and the whole system. The result of system testing is the observed behavior of a component or a system when it is tested. **System Testing** is carried out on the whole system in the context of either system requirement specifications or functional requirement specifications or in the context of both. System testing tests the design and behavior of the system and also the expectations of the customer. It is performed to test the system beyond the bounds mentioned in the software requirements specification (SRS). System Testing is basically performed by a testing team that is independent of the development team that helps to test the quality of the system impartial. It has both functional and non-functional testing. **System Testing is a black-box testing**. System Testing is performed after the integration testing and before the acceptance testing.



System Testing Process: System Testing is performed in the following steps:

- **Test Environment Setup:** Create testing environment for the better-quality testing.
- **Create Test Case:** Generate test case for the testing process.
- **Create Test Data:** Generate the data that is to be tested.

- **Execute Test Case:** After the generation of the test case and the test data, test cases are executed.
- **Defect Reporting:** Defects in the system are detected.
- **Regression Testing:** It is carried out to test the side effects of the testing process.
- **Log Defects:** Defects are fixed in this step.
- **Retest:** If the test is not successful then again test is performed.



Types of System Testing:

- **Performance Testing:** Performance Testing is a type of software testing that is carried out to test the speed, scalability, stability and reliability of the software product or application.
- **Load Testing:** Load Testing is a type of software Testing which is carried out to determine the behavior of a system or software product under extreme load.
- **Stress Testing:** Stress Testing is a type of software testing performed to check the robustness of the system under the varying loads.
- **Scalability Testing:** Scalability Testing is a type of software testing which is carried out to check the performance of a software application or system in terms of its capability to scale up or scale down the number of user request load.

Performance testing

Performance testing is a type of software testing that evaluates the speed, responsiveness, and stability of a system under a specific workload. The primary goal of performance testing is to identify and address performance-related issues before the software is deployed in a production environment. Performance testing helps ensure that a system can handle the expected user load and provides a satisfactory user experience. Here are some common types of performance testing:

1.Load Testing:

□ Objective: To determine how the system performs under anticipated user loads.

□ Process: Gradually increase the number of simultaneous users or transactions until the system reaches its
capacity limits.

2. Stress Testing:

□ Objective: To evaluate the system&#39;s behavior under extreme conditions or beyond its expected capacity.

□ Process: Increase the load beyond the system&#39;s capacity to identify its breaking point.

3. Volume Testing:

□ Objective: To assess how the system handles a large amount of data.

□ Process: Test the software&#39;s performance with a significant amount of data, such as a large database or
high file volume.

4. Scalability Testing:

□ Objective: To evaluate the system&#39;s ability to scale up or down in terms of hardware, network, or other
resources.

□ Process: Assess the system&#39;s performance as resources are added or removed.

5. Endurance Testing:

□ Objective: To check if the system can handle a sustained load over an extended period.

□ Process: Apply a constant load for an extended duration to identify potential performance degradation or
issues over time.

6. Concurrency Testing:

 Objective: To evaluate how well the system supports multiple users or transactions simultaneously.

 Process: Simulate concurrent user interactions to identify and address issues related to concurrency.

7. Reliability Testing:

 Objective: To ensure the system&#39;s stability and reliability under different conditions.

 Process: Test the system&#39;s performance over an extended period, including restarts and recovery
scenarios.

8. Isolation Testing:

 Objective: To identify and isolate performance bottlenecks in the system.

 Process: Analyze and test individual components or subsystems to identify areas for improvement.

Performance testing tools, such as Apache JMeter, LoadRunner, and Gatling, are commonly used to automate and streamline the testing process. It&#39;s important to conduct performance testing throughout the software development life cycle to catch and address performance issues early on, ensuring a smooth user experience when the software is deployed.

SECURITY TESTING

Security testing is a crucial part of the software development process that focuses on identifying and addressing vulnerabilities to protect against potential security threats. It involves various techniques, including:

 Vulnerability Assessment: Identifying system weaknesses using automated tools and manual analysis.

 Penetration Testing: Simulating real-world attacks to uncover and exploit security vulnerabilities.

 Security Scanning: Automated tools scan applications or networks for known security issues.

 Code Review (Static Analysis): Reviewing source code to find and fix security vulnerabilities.

 Authentication and Authorization Testing: Evaluating the effectiveness of access control mechanisms.

 Data Security Testing: Ensuring the confidentiality, integrity, and availability of sensitive data.

 Security Architecture Review: Assessing the overall security design and implementation.

 Security Auditing: Verifying compliance with security policies and standards.

By integrating security testing throughout the development lifecycle, teams can proactively identify and mitigate potential security risks, protecting both the software and its users from security threats.

SYSTEM SECURITY

System testing is the phase of software testing in which the entire integrated system is tested to evaluate its
compliance with specified requirements and to ensure that it functions as intended in a real-world environment.

ALPHA BETA TESTING

Alpha and beta testing are two distinct phases of software testing:

Alpha Testing: In-house testing conducted by the development team to identify and address issues within
the software before releasing it to a limited external user group.

Beta Testing: External users or a select group of customers test the software in a real-world environment
to provide feedback on usability, functionality, and to uncover potential issues before the official release to
the wider audience.

BLACK BOX TESTING

Black box testing is a software testing method where the tester assesses the functionality of a system without knowledge of its internal code or implementation details. The focus is on evaluating inputs and outputs, ensuring that the software behaves according to its specifications and requirements.

WHITE BOX TESTING

White box testing, also known as clear box testing or structural testing, is a software testing method that examines the internal structure and logic of a system. Testers have knowledge of the internal code, architecture, and design, allowing them to create test cases based on this understanding. The goal is to verify the correctness of the code, ensure all branches and paths are tested, and assess the overall system&#39;s reliability.

UNIT TESTING

Unit testing is a software testing method where individual units or components of a software application are tested in isolation to ensure they function correctly. The focus is on verifying that each unit of code performs as intended and meets its design specifications. Unit tests are typically automated and are an essential part of the development process to catch and fix bugs early in the software development lifecycle.

INTEGRATION TESTING

Integration testing is a software testing method that focuses on verifying the interactions between different modules or components of a system when integrated together. The primary goal is to identify and address issues that may arise from the combination of these components. Integration testing ensures that the units, when combined, work as expecte and that data flows correctly between them. It helps detect interface defects, communication problems, and other issues that may arise from the interactions between integrated components.

ACCEPTANCE TESTING

Acceptance testing is a phase of software testing where a system is evaluated for its compliance with business requirements and whether it meets the acceptance criteria set by the stakeholders. It is the final phase before the software is deployed to a production environment. Acceptance testing can be divided into
two main types:

User Acceptance Testing (UAT): Conducted by end-users or clients to ensure that the software meets their business needs and requirements.

Business Acceptance Testing (BAT): Performed by business stakeholders to verify that the software aligns with the overall business objectives and goals.

The purpose of acceptance testing is to confirm that the software is ready for production use and that it
satisfies the specified criteria agreed upon by both developers and stakeholders.

FUNCTIONAL TESTING

Functional testing is a type of software testing that evaluates the functions or features of a software application to ensure they work according to specified requirements. The primary focus is on verifying that the software performs its intended operations and produces the correct results. Functional testing can be performed at various levels of the software development life cycle and includes different types of testing such as unit testing, integration testing, and system testing.

5.RESULTS AND PERFORMANCE ANALYSIS

5.RESULTS AND PERFORMANCE ANALYSIS

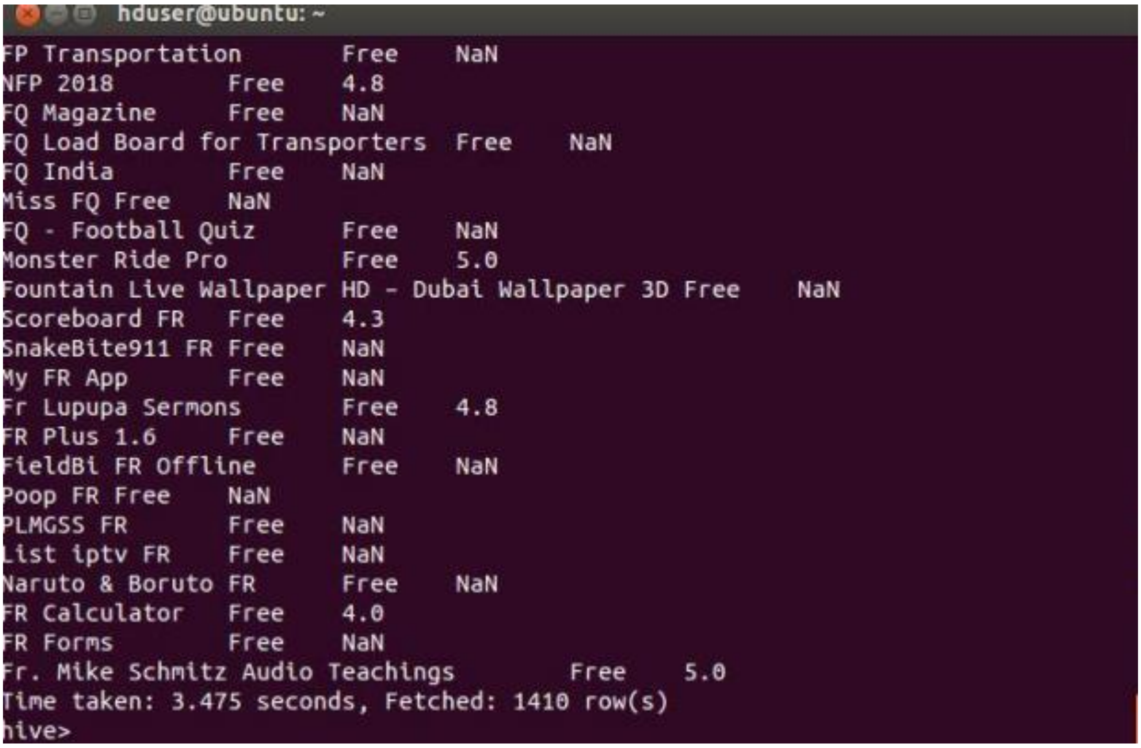
After implementing the project on machine learning and deep neural network I have observed the following results.

5.1 ANALYSIS

5.1.1 USING MACHINE LEARNING

Preprocessing will take away null values and duplicates from a reliable dataset. The dataset is separated into 2 sections: a training dataset and a testing dataset using the train test split function. A training dataset is needed for the development of machine learning models. This approach makes use of machine learning models that enable vector regression and random forest regression. The test data set will be used to train models, which can then be used to create predictions. Based on the results of the 2 machine learning approaches, the final projection is calculated.

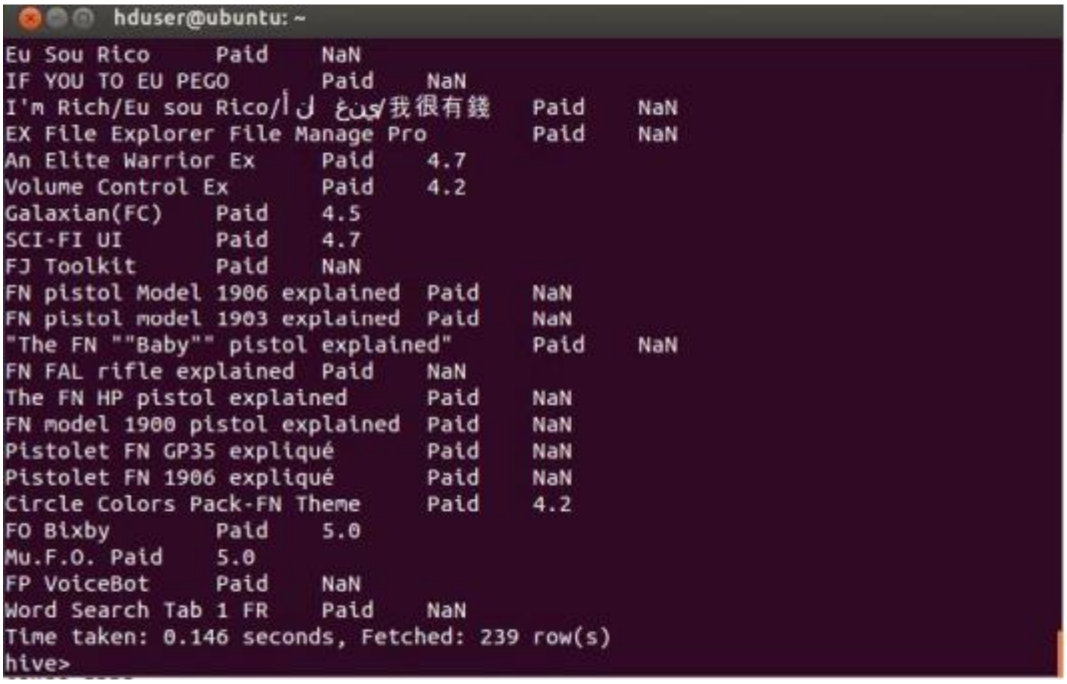
1 Firstly, I have to tried find out the names of those applications which come on the top of the entire free app with their respective rating. This picture shows all the names of top free apps.



```
FP Transportation       Free    NaN
NFP 2018        Free    4.8
FQ Magazine     Free    NaN
FQ Load Board for Transporters  Free    NaN
FQ India        Free    NaN
Miss FQ Free    NaN
FQ - Football Quiz      Free    NaN
Monster Ride Pro        Free    5.0
Fountain Live Wallpaper HD - Dubai Wallpaper 3D Free    NaN
Scoreboard FR   Free    4.3
SnakeBite911 FR Free    NaN
My FR App       Free    NaN
Fr Lupupa Sermons       Free    4.8
FR Plus 1.6     Free    NaN
FieldBi FR Offline      Free    NaN
Poop FR Free    NaN
PLMGSS FR       Free    NaN
List iptv FR    Free    NaN
Naruto & Boruto FR      Free    NaN
FR Calculator   Free    4.0
FR Forms        Free    NaN
Fr. Mike Schmitz Audio Teachings        Free    5.0
Time taken: 3.475 seconds, Fetched: 1410 row(s)
hive>
```

Figure 21: Top free applications

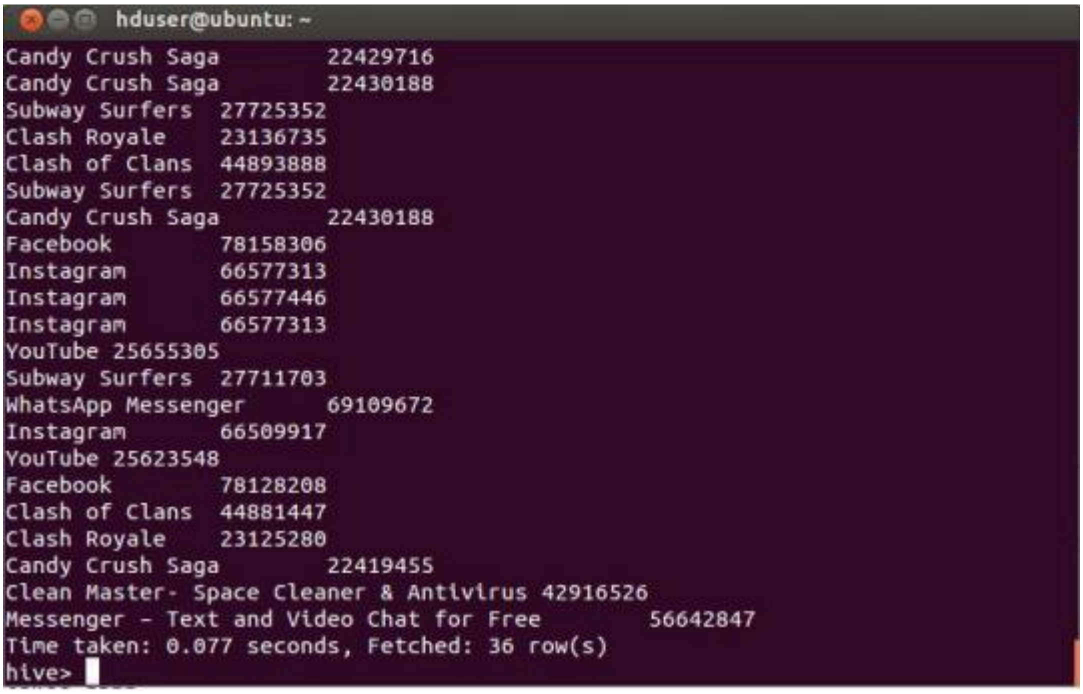
In this I have tried to find out names of those applications which comes on the top of all the paid apps with their respective rating. This picture shows all the names of top paid apps among all the presented apps in given dataset



```
hduser@ubuntu: ~
Eu Sou Rico     Paid    NaN
IF YOU TO EU PEGO       Paid    NaN
I'm Rich/Eu sou Rico/أ ل  غني/我很有錢   Paid    NaN
EX File Explorer File Manage Pro        Paid    NaN
An Elite Warrior Ex     Paid    4.7
Volume Control Ex       Paid    4.2
Galaxian(FC)    Paid    4.5
SCI-FI UI       Paid    4.7
FJ Toolkit      Paid    NaN
FN pistol Model 1906 explained  Paid    NaN
FN pistol model 1903 explained  Paid    NaN
"The FN ""Baby"" pistol explained"      Paid    NaN
FN FAL rifle explained  Paid    NaN
The FN HP pistol explained      Paid    NaN
FN model 1900 pistol explained  Paid    NaN
Pistolet FN GP35 expliqué       Paid    NaN
Pistolet FN 1906 expliqué       Paid    NaN
Circle Colors Pack-FN Theme     Paid    4.2
FO Bixby        Paid    5.0
Mu.F.O. Paid    5.0
FP VoiceBot     Paid    NaN
Word Search Tab 1 FR    Paid    NaN
Time taken: 0.146 seconds, Fetched: 239 row(s)
hive>
```

Figure 22: Top paid applications

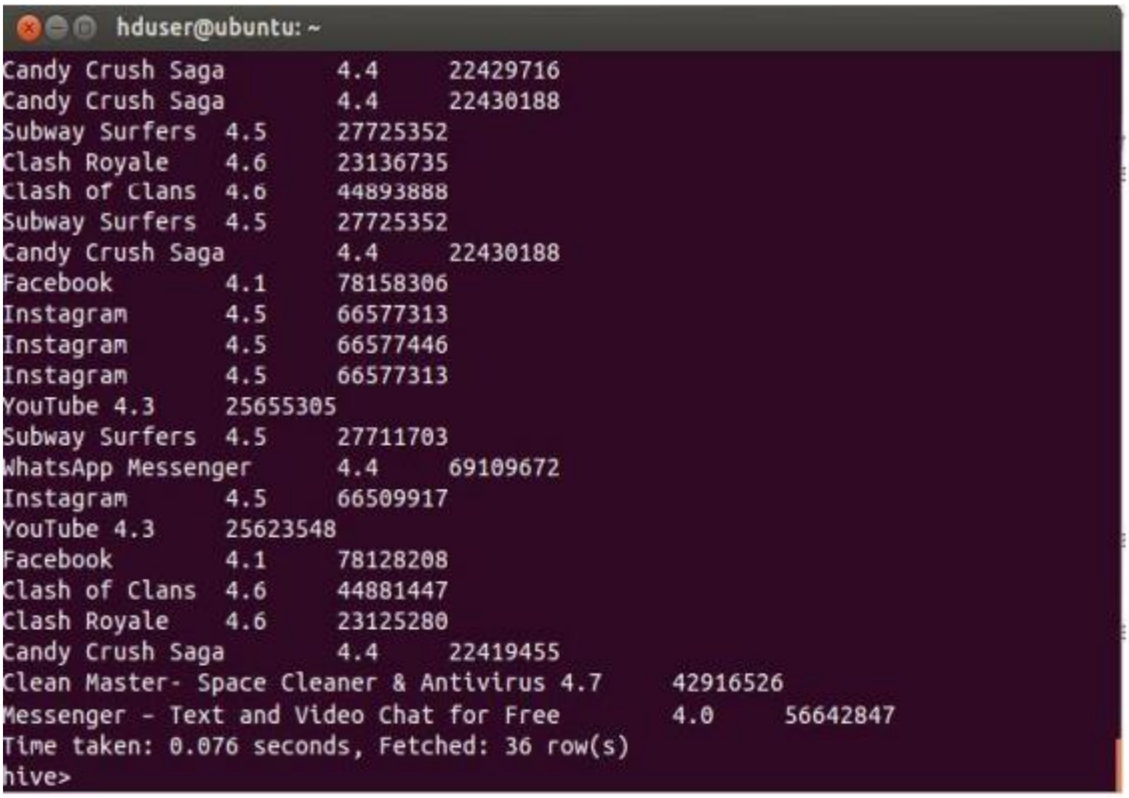
(iii) In this I have tried to find out names of those applications which come on the top of all the reviewed apps. This picture shows all the names of apps which are most reviewed by the users among all the presented apps in given dataset



```
hduser@ubuntu: ~
Candy Crush Saga        22429716
Candy Crush Saga        22430188
Subway Surfers  27725352
Clash Royale    23136735
Clash of Clans  44893888
Subway Surfers  27725352
Candy Crush Saga        22430188
Facebook        78158306
Instagram       66577313
Instagram       66577446
Instagram       66577313
YouTube 25655305
Subway Surfers  27711703
WhatsApp Messenger      69109672
Instagram       66509917
YouTube 25623548
Facebook        78128208
Clash of Clans  44881447
Clash Royale    23125280
Candy Crush Saga        22419455
Clean Master- Space Cleaner & Antivirus 42916526
Messenger - Text and Video Chat for Free        56642847
Time taken: 0.077 seconds, Fetched: 36 row(s)
hive>
```

Figure 23: Top reviewed applications

(iv) In this I have tried to find out names of those applications which come on the top of Editor's choice. This picture shows all the names of the apps which come under the editor's choice whose rating is above 4 and has more reviews among all the presented apps in given dataset



hduser@ubuntu: ~

```
Candy Crush Saga        4.4     22429716
Candy Crush Saga        4.4     22430188
Subway Surfers  4.5     27725352
Clash Royale    4.6     23136735
Clash of Clans  4.6     44893888
Subway Surfers  4.5     27725352
Candy Crush Saga        4.4     22430188
Facebook        4.1     78158306
Instagram       4.5     66577313
Instagram       4.5     66577446
Instagram       4.5     66577313
YouTube 4.3     25655305
Subway Surfers  4.5     27711703
WhatsApp Messenger      4.4     69109672
Instagram       4.5     66509917
YouTube 4.3     25623548
Facebook        4.1     78128208
Clash of Clans  4.6     44881447
Clash Royale    4.6     23125280
Candy Crush Saga        4.4     22419455
Clean Master- Space Cleaner & Antivirus 4.7     42916526
Messenger - Text and Video Chat for Free        4.0     56642847
Time taken: 0.076 seconds, Fetched: 36 row(s)
hive>
```

Figure 24: Editor’s choice apps

5.2 PREDICTION

In the branch of artificial intelligence known as "machine learning," algorithms and models are created that can learn from data and generate predictions. From banking and healthcare to transportation and marketing, machine learning prediction is now a crucial tool in many fields of study and industry. This article will examine machine learning prediction, including what it is, how it functions, and some of its uses.

The practise of using data to create predictions or foresee future events is known as machine learning prediction. Building models that can recognise patterns in data and utilise those patterns to create precise predictions about novel, unforeseen data is the aim of machine learning prediction. These forecasts can be used to guide decisions, such as identifying the customers most likely to purchase a product, the individuals most likely to contract an illness, or the investments most likely to provide large returns.

Machine learning prediction involves several steps:

Data gathering: Gathering pertinent data for the current issue is the first step in machine learning prediction. Many other sources, including sensors, surveys, and databases, can provide this data.

Data preparation: After data has been gathered, it needs to be ready for analysis. This entails preparing the data for analysis by cleaning it, eliminating outliers or errors, and converting it to an appropriate format.

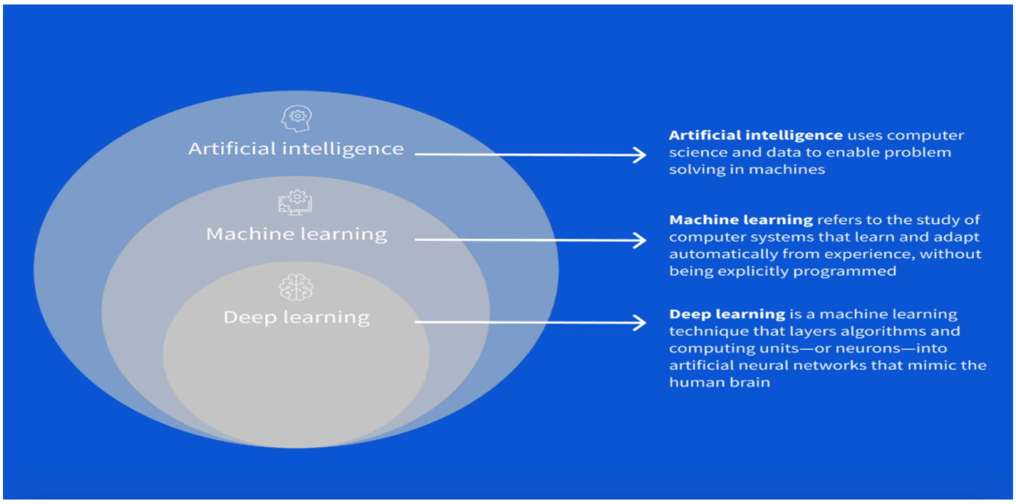
Model development is the following action in the machine learning process. In order to do this, a suitable algorithm must be chosen, the input variables (also known as features) must be defined, and the model must be trained using the prepared data.

Evaluation of the model: After the model has been trained, it is assessed using a different dataset known as the test set. This is done to make sure the model is not overfitting to the training data and is generalising well to new data.

5.3 DEEP LEARNING OVER MACHINE LEARNING

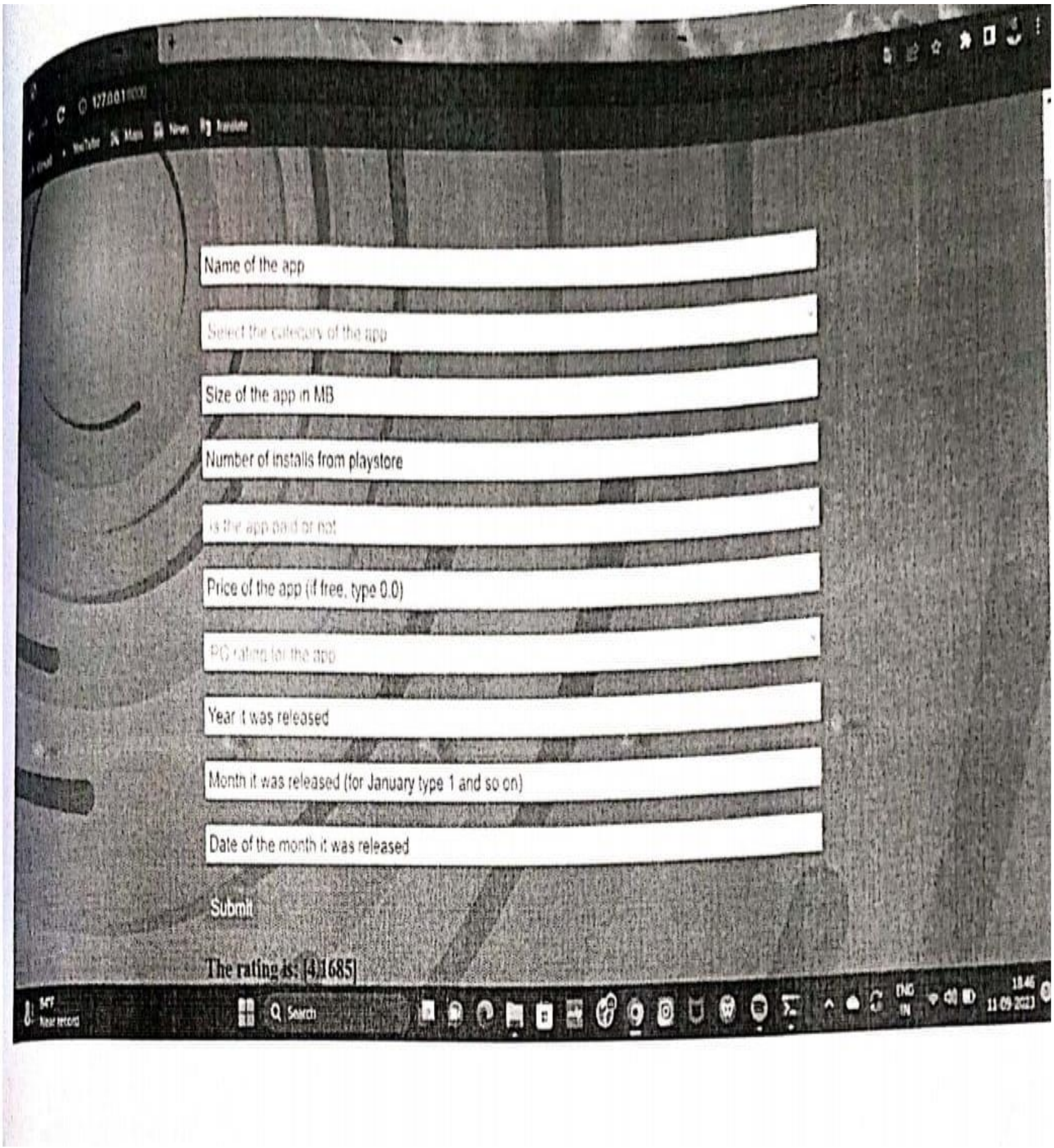
Machine learning and deep learning are both types of AI. In short, **machine learning** is AI that automatically adapt with minimal human interference. **Deep learning** is a subset of machine learning that uses artificial neural networks to mimic the learning process of the human brain.

Take a look at these key differences before we dive in further.


Artificial intelligence
Machine learning
Deep learning
Artificial intelligence uses computer science and data to enable problem solving in machines
Machine learning refers to the study of computer systems that learn and adapt automatically from experience, without being explicitly programmed
Deep learning is a machine learning technique that layers algorithms and computing units—or neurons—into artificial neural networks that mimic the human brain

SCREEN SHORTS

SCREEN SHORTS



Name of the app

Select the category of the app

Size of the app in MB

Number of installs from playstore

Is the app paid or not

Price of the app (if free, type 0.0)

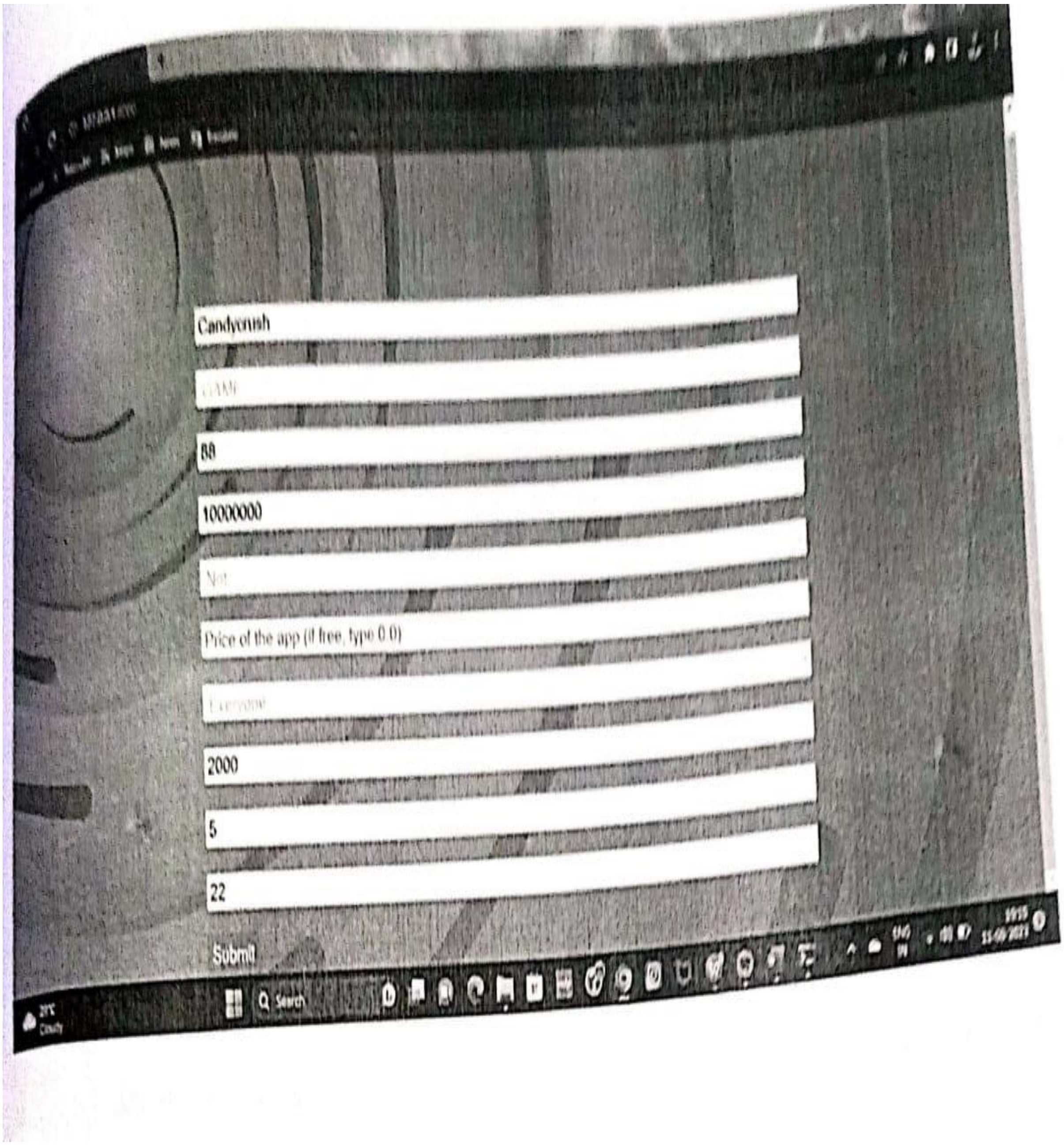
PG rating for the app

Year it was released

Month it was released (for January type 1 and so on)

Date of the month it was released

The rating is: [4.1685]



Candycrush

GAME

88

10000000

Nil

Price of the app (if free, type 0.0)

Everyone

2000

5

22

7.CHAPTER CONCLUSION

7.CHAPTER CONCLUSION

CONCLUSION

The Google Play Store is the biggest application advertises on this earth. It produces more than the download of the Apple App Store, yet not profits as the App Store. We scratched information from the Play Store to lead examine on it. In this project I have used Big data technique such as Machine learning to analysis the different attributes of the given dataset of Google play store application such as top free apps, top paid apps, most reviewed apps, apps undereditor's choice with the help of Machine learning QL and displayed the results as shown above.

We came to the conclusion that our hypothesis is correct after running through all of these algorithms and processes. As a result, it is possible to predict app ratings, but a large amount of preprocessing is required before the classification and regression processes can be started. The data collected from Google Play Store apps has huge potential to help app development companies succeed. Developers can use the information to their advantage to work on and conquer the Android market! In order to accurately estimate whether an app will have more than 100,000 downloads and be a success on the Google Play Store, we need to know the app's Size, Type, Price, Content Rating, and Genre.

FUTURE SCOPE

Only polarity and subjectiveness may be obtained from user reviews. Predictions are also important because of the enormous expansion in review-based data. This is a challenging but rewarding process, as user reviews are qualitative and ratings are mainly quantitative. Additionally, Google's numerical rating system may be distorted and amplified by the fact that higher ratings supplied by consumers may bring in disproportionately more new users. So this study investigated if ensemble classifiers might be used to predict numerical ratings for Google Play store apps based on user reviews. The Google App store assessments were used to test many ensemble classifiers. To predict numerical ratings in the future, deep learning technology will be applied

REFERENCES

[1] Statista, Numerous accessible applications within the Google Play store from December 2009 to March 2019, https://www.statista.com/statistics/266210/numberof-available-applications-in-the-google-play-store/, Online: accessed twentytwo May 2019.

[2] Statista, various mobile app downloads worldwide in 2017, 2018, and 2020 (in billions), https://www.statista.com/statistics/ 271644/worldwide-free-and-paid-mobile-app-store downloads/, Online: accessed twenty-two May 2019.

[3] Online shopping, pew internet, and American life project, Washington, DC, 2018, http://www.pewinternet.org/Repor ts/2008/online shopping/01-Summary-of-Findings.aspx Online: accessed eight August. 2014.

[4] D. Pagano and W. Maalej, User feedback within the AppStore: an empirical study, in Proc. IEEE Int. Requirements Eng. Conf. (Rio de Janeiro, Brazil), July 2013, pp. 125–134.

[5] T. Chumwatana, using sentiment analysis technique for analyzing Thai client satisfaction from social media, 2015.

[6] T. Trivia et al., Mobile apps feature extraction supported user reviews using machine learning, 2019.

[7] H. Hanyang et al., learning the consistency of star ratings and reviews of standard free hybrid android and iOS apps, Empirical Softw. Eng. 24 (2019), no. 7, 7–32.

[8] N. Kumari and S. Narayan Singh, Sentiment analysis on e-commerce application by using opinion mining, in Proc. Int. Conf.- Cloud Syst. Big Data Eng. (Noida, India), Jan. 2016, pp. 320–325.

[9] R. M. Duwairi and that I. Qarqaz, Arabic sentiment analysis using supervised classification, in Proc. Int. Conf. Future Internet Things Cloud (Barcelona, Spain), Aug. 2014, pp. 579–583.

[10] H. S. Le, T. V. Le, and T. V. Pham, Aspect analysis for opinion mining of Vietnamese text, in Proc. Int. Conf. Adv. Comput. Applicant. (Ho Chi Minh, Vietnam), Nov. 2015, pp. 118–123.

[11] H. Wang, L. Yue, and C. Zhai, Latent aspect rating analysis on review text data: a rating regression approach, in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining (Washington, D.C., USA), July 2010, pp. 783–792.

[12] K. Dave, S. Lawrence, and D. M. Pennock Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in Proc. Int. Conf. World Wide Web (New York, USA), 2003, pp. 519–528.