**`11Internship Report**

**on**

**PREDICTING APP RATINGS**

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR, ANANTHAPURAMU**

*In Partial Fulfillment of the Requirements for the Award of the degree of*

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE & ENGINEERING – DATA SCIENCE**

**Submitted By**

**K.LAKSHITHA - 20691A3217**

****

**MADANAPALLE INSTITUTE OF TECHNOLOGY & SCIENCE**

**(UGC – AUTONOMOUS)**

**(Affiliated to JNTUA, Ananthapuramu)**

**(Accredited by NBA, Approved by AICTE New Delhi)**

**AN ISO 9001:2015 Certified Institution**

**P. B. No: 14, Angallu, Madanapalle – 517325**

**2023 - 24**

**2009-2013**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING – DATA SCIENCE**

**BONAFIDE CERTIFICATE**

This is to certify that the internship work entitled **“Predicting App Ratings”** is a bonafide work carried out by

**K.LAKSHITHA - 20691A3217**

Submitted in partial fulfillment of the requirements for the award of degree **Bachelor of Technology** in the stream of **Computer Science & Engineering** in **Madanapalle Institute of Technology & Science, Madanapalle,** affiliated to **Jawaharlal Nehru Technological University Anantapur, Ananthapuramu** during the academic year 2023-2024.

**Internship Coordinator Head of the Department**

Mrs. Manjula Prabakaran, Dr. S. Kusuma,

Assistant Professor, Assistant Professor & Head,

Department of CSE - DS Department of CSE - DS

**Submitted for the University Examination held on: -----------------**

**Examiner - I Examiner - II**

**ACKNOWLEDGEMENT**

I sincerely thank the **Management** of **Madanapalle Institute of Technology and**

**Science** for providing excellent infrastructure and lab facilities that helped me to complete this project.

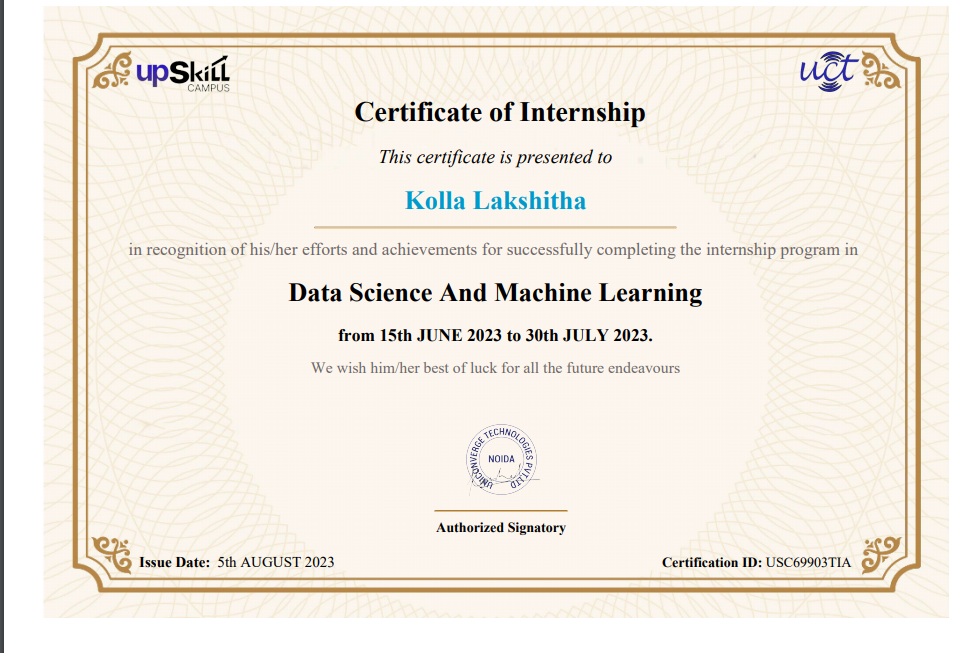
I sincerely thank **Dr. C. Yuvaraj , M.E., Ph.D., Principal** for guiding and providing facilities for the successful completion of our project at **Madanapalle Institute of Technology and Science,** Madanapalle.

I express our deep sense of gratitude to to **Dr. S. Kusuma, Ph.D., Assistant Professor and Head of the Department of CSE-DS** for her continuous support in making necessary arrangements for the successful completion of the Project.

I express our sincere thanks to the **Internship Coordinator, Mrs. Manjula Prabakaran. Assistant Professor, Department of CSE-DS** for her tremendous support for the successful completion of Project.

I also wish to place on record my gratefulness to other **Faculty members of CSE-DS Department** and our parents and friends for their help and cooperation during our project work.

**CERTIFICATE**

**DECLARATION**

I, the undersigned hereby declare that the results embodied in this Internship **“Predicting App Ratings”** is a bonafide record of the work done by me in partial fulfillment of the award of **Bachelor of Technology** in **Computer Science & Engineering** \_**Data Science** from **Jawaharlal Nehru Technological University Anantapur, Ananthapuramu.** The content of this reportis not submitted to any other University/Institute for award of any other degree.

|  |  |
| --- | --- |
| Place **: Madanapalle**  Date : | Name : K.Lakshitha  Roll No : 20691A3217  Department of Computer Science & Engineering\_Data Science  Madanapalle Institute of Technology & Science  Madanapalle |

**ABSTRACT**

Millions of users utilize Google Play Store, the company's official android market. It can be utilized to access magazine, gaming, music, movie, and television material, among other things. Users can rate and review apps in the android market after installing them to express their own experiences with them, and this goes both ways in that other user can be inspired by other user’s reviews. Users' experiences usually explain the app's usability, performance, and, in some cases, issues that they have encountered while using it. The goal is to categorise Google’s app evaluations using machine learning approach. The Machine learning approaches are employed to gather element identify things such as emotions, element polarities, and to build a sample utilizing these characteristics, that could then go into a series of procedures including data validation or cleaning, visualization, and finally classification into positive, neutral, and negative sentiments. Furthermore, the predictive model can help users choose apps that align with their preferences and expectations. By providing estimated ratings, users can make more informed decisions when downloading and using mobile applications.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE** | **NAME** | **PAGE NO.** |
| Fig 1.1 | Data Science | 1 |
| Fig 1.2 | Life Cycle of Data Science | 3 |
| Fig 1.3 | Types of Machine Learning | 4 |
| Fig 1.4 | Applications of Machine learning | 10 |
| Fig 3.4 | System Architecture | 14 |

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **CHAPTER** | **PAGE NO.** |
|  | **ACKNOWLEDGEMENT** | (iii) |
|  | **CERTIFICATE** | (iv) |
|  | **ABSTRACT** | (vi) |
|  | **LIST OF FIGURES** | (vii) |
| **1** | **INTRODUCTION** | 1 |
|  | 1.1 Data Science | 1 |
|  | 1.2 Applications of Data Science | 3 |
|  | 1.3 About Machine Learning | 4 |
|  | 1.4 Importance and Application of ML | 9 |
| **2** | **TOOLS AND TECHNOLOGY** | 11 |
|  | 2.1 Software Requirements | 11 |
|  | 2.2 Hardware Requirements | 11 |
| **3** | **PROJECT ANALYSIS** | 12 |
|  | 3.1 Problem Statement | 12 |
|  | 3.2 Project Overview | 12 |
|  | 3.3 Algorithm | 13 |
|  | 3.4 System Architecture | 14 |
|  | 3.5 Modules | 14 |
| **4** | **CONCLUSION** | 17 |
| **5** | **BIBLIOGRAPHY** | 18 |
|  | **APPENDIX** | 19 |

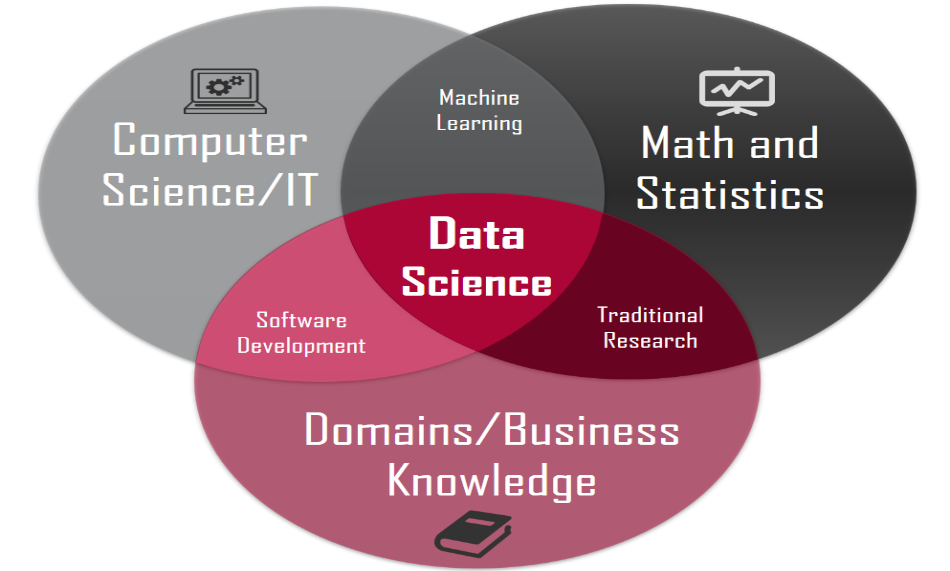
**CHAPTER 1**

**INTRODUCTION**

* 1. **DATA SCIENCE**

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization’s data. These insights can be used to guide decision making and strategic planning. The accelerating volume of data sources, and subsequently data, has made data science is one of the fastest growing field across every industry. Organizations are increasingly reliant on them to interpret data and provide actionable recommendations to improve business outcomes. Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

Data science uses complex [machine learning algorithms](https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article) to build predictive models .The data used for analysis can come from many different sources and presented in various formats. Now that you know what data science is, let’s see the data science lifestyle. Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand that they must advance past the traditional skills of analyzing large amounts of data, data mining, and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.



**Fig 1.1 : Data Science**

The data science lifecycle involves various roles, tools, and processes, which enables analysts to glean actionable insights. Typically, a data science project undergoes the following stages:

* **Data ingestion**: The lifecycle begins with the data collection--both raw structured and unstructured data from all relevant sources using a variety of methods. These methods can include manual entry, web scraping, and real-time streaming data from systems and devices. Data sources can include structured data, such as customer data, along with unstructured data like log files, video, audio, pictures, the Internet of Things (IoT), social media, and more.
* **Data storage and data processing:** Since data can have different formats and structures, companies need to consider different storage systems based on the type of data that needs to be captured. Data management teams help to set standards around data storage and structure, which facilitate workflows around analytics, machine learning and deep learning models. This stage includes cleaning data, deduplicating, transforming and combining the data using ETL (extract, transform, load) jobs or other data integration technologies. This data preparation is essential for promoting data quality before loading into a [data warehouse](https://www.ibm.com/topics/data-warehouse), [data lake](https://www.ibm.com/topics/data-lake), or other repository.
* **Data analysis:** Here, data scientists conduct an exploratory data analysis to examine biases, patterns, ranges, and distributions of values within the data. This data analytics exploration drives hypothesis generation for a/b testing. It also allows analysts to determine the data’s relevance for use within modeling efforts for predictive analytics, machine learning, and/or deep learning. Depending on a model’s accuracy, organizations can become reliant on these insights for business decision making, allowing them to drive more scalability.
* **Communicate:** Finally, insights are presented as reports and other data visualizations that make the insights—and their impact on business—easier for business analysts and other decision-makers to understand. A data science programming language such as R or Python includes components for generating visualizations; alternately, data scientists can use dedicated visualization tools.

Data science community - Connect with experts and peers to elevate technical expertise, solve problems and share insights.

* **Data Cleaning and Preprocessing:** Data cleaning (also known as data cleansing) is part of the pre-processing activity, where we wish to modify the data set in some manner to correct erroneous data, remove redundancies, or deal with incomplete or missing data.

A circular chart with colorful circles

Description automatically generated

**Fig 1.2 : Life cycle of Data Science**

The data science life cycle is flexible and may vary depending on the specific project and organization. It's important to note that data science is a dynamic field, and new tools and techniques are continually emerging, so staying up-to-date with the latest developments is crucial for success in this field.

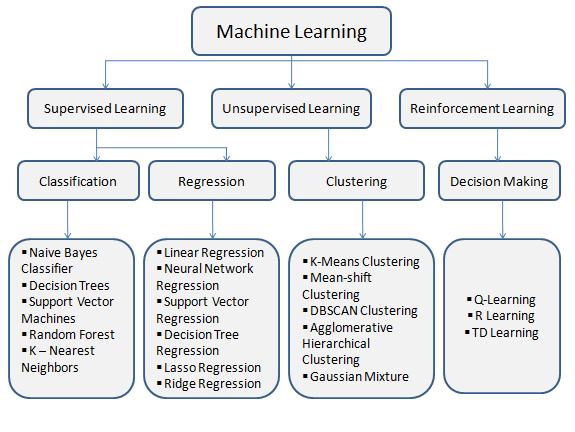
**1.2 APPLICATIONS OF DATASCIENCE**

* [Education](https://intellipaat.com/blog/applications-of-data-science-real-world-applications/#education)
* [Airline Route Planning](https://intellipaat.com/blog/applications-of-data-science-real-world-applications/#airline_route_planning)
* [Healthcare Industry](https://intellipaat.com/blog/applications-of-data-science-real-world-applications/#healthcare_industry)
* [Delivery Logistics](https://intellipaat.com/blog/applications-of-data-science-real-world-applications/#delivery_logistics)
* [Banking and Finance](https://intellipaat.com/blog/applications-of-data-science-real-world-applications/#banking_and_finance)
* [Filtered Internet Search](https://intellipaat.com/blog/applications-of-data-science-real-world-applications/#filtered_internet_search)
* [Product Recommendation Systems](https://intellipaat.com/blog/applications-of-data-science-real-world-applications/#product_recommendation_systems)
* [Digital Marketing and Advertising](https://intellipaat.com/blog/applications-of-data-science-real-world-applications/#digital_marketing_and_advertising)

**1.3 ABOUT MACHINE LEARNING**

Machine learning is a sub-domain of computer science. It uses data and artificial intelligence in its area of applications. It is considered as the top-notch pass to the most interesting and growing careers in the current world. It is used to make predictions and gain insights. This can be achieved by providing the data to train the model. At a broad level, machine learning can be classified into three types:

* Supervised Machine Learning
* Unsupervised Machine Learning
* Reinforcement learning

****

**Fig 1.3 Types of Machine Learning**

* **Supervised Machine Learning**

In supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the shape & size of the tail of cat and dog, Shape of eyes, color, height (dogs are taller, cats are smaller), etc. After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, color, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category.

This is the process of how the machine identifies the objects in Supervised Learning The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y). Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, etc.

**Categories of Supervised Machine Learning**

Supervised machine learning can be classified into two types of problems, which are given below:

* Classification
* Regression

**Classification**

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as **"Yes" or No, Male or Female, Red or Blue, etc.** The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**

Some popular classification algorithms are given below :

* Random Forest Algorithm
* Decision Tree Algorithm
* Logistic Regression Algorithm
* Support Vector Machine Learning Algorithm

**Regression**

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below

* Linear Regression Algorithm
* Non-Linear Regression Algorithm
* Polynomial Regression Algorithm

**Advantages and Disadvantages of Supervised Learning**

**Advantages**

* Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
* These algorithms are helpful in predicting the output on the basis of prior experience.

**Disadvantages**

* These algorithms are not able to solve complex tasks.
* It may predict the wrong output if the test data is different from the training data.
* It requires lots of computational time to train the algorithm.

### **Applications of Supervised Learning**

Some common applications of Supervised Learning are given below:

* **Image Segmentation :** Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.
* **Medical Diagnosis :** Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions. With such a process, the machine can identify a disease for the new patients.
* **Fraud Detection :** Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
* **Spam detection :** In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
* **Speech Recognition :** Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

**Un Supervised Machine Learning Algorithm**

In unsupervised machine learning, the machine is trained using the unlabelled dataset, and the machine predicts the output without any supervision . In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

**Categories of Unsupervised Machine Learning**

Unsupervised Learning can be further classified into two types, which are given below:

* Clustering
* Association

**Clustering**

The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

Some of the popular clustering algorithms are given below:

* K-Means Clustering algorithm
* **Mean-shift algorithm**
* **DBSCAN Algorithm**
* **Principal Component Analysis**
* **Independent Component Analysis**

### **Association**

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in**Market Basket analysis, Web usage mining, continuous production.**

Some popular algorithms of Association rule learning are **Apriori Algorithm, Eclat, FP-growth algorithm.**

### **Advantages and Disadvantages of Unsupervised Learning Algorithm**

### **Advantages**

* These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
* Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

**Disadvantages**

The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.

* Working with Unsupervised learning is more difficult as it works with the unlabeled dataset that does not map with the output.

### **Applications of Unsupervised Learning**

* **Network Analysis:** Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data for scholarly articles.
* **Recommendation Systems:** Recommendation systems widely use unsupervised learning techniques for building recommendation applications for different web applications and e-commerce websites.
* **Anomaly Detection:** Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraudulent transactions.
* **Singular Value Decomposition:** Singular Value Decomposition or SVD is used to extract information from the database. For example, extracting information of each user located at a particular location.

**Reinforcement Learning**

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance. Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards .In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only. The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards. Due to its way of working, reinforcement learning is employed in different fields such as Game theory, Operation Research, Information theory, multi-agent systems.

**Advantages and Disadvantages of Reinforcement Learning**

**Advantages**

* It helps in solving complex real-world problems which are difficult to be solved by general techniques.
* The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
* Helps in achieving long term results.

**Disadvantages**

* RL algorithms are not preferred for simple problems.
* RL algorithms require huge data and computations.
* Too much reinforcement learning can lead to an overload of states which can weaken the results.

**1.4 IMPORTANCE AND APPLICATIONS OF MACHINE LEARNING**

With the growing economy, the world is changing, and the internet has become the data generation machine. Machine learning helps the data analysts to organize and handle this data. It helps in analyzing the data and provides valuable insights. Machine learning allows the software’s to become more accurate.

It is known to everyone that large companies are describing the Machine Learning as “The future”. It has many applications in various domains. Few of them are listed below :

* + Image Recognition
  + Automatic language Translation
  + Medical Diagnosis
  + Stock market Trading
  + Online Fraud Detection
  + Virtual Personal Assistant
* Self-driving cars
* Recommendation Systems (Movie recommendation, Music Recommendation)
* Image recognition.
* Email Spam and Malware Filtering

A circular diagram of applications of machine learning

Description automatically generated

**Fig 1.4 Applications of Machine Learning**

SVMs use input-output pairs during training to constructing

an N-dimensional hyperplane that optimally separates the data

into two cat**Random Forest Algorithm**ccCHAPTER-4

**CHAPTER 2**

**TOOLS AND TECHNOLOGLY**

**2.1 SOFTWARE REQUIREMENTS**

Here, I have usedColab platform for executing my code. Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members just the way you edit documents in Google Docs. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.

As a programmer, you can perform the following using Google Colab.

* Write and execute code in Python
* Document your code that supports mathematical equations
* Create/Upload/Share notebooks
* Import/Save notebooks from/to Google Drive
* Import/Publish notebooks from GitHub
* Import external datasets e.g., from Kaggle
* Integrate PyTorch, TensorFlow, Keras, OpenCV
* Free Cloud service with free GPU

**2.2 HARDWARE REQUIREMNTS**

Device name : HP Laptop 15p-da0xxx

Processor : intel® Core ™ i5 -82504 CPU @ 1.60GHz 1.80

Installed Ram : 8.00 GB

System type : 64-bit Operating System , X64- based processor.

**CHAPTER 3**

**PROJECT ANALYSIS**

**3.1 PROBLEM STATEMENT**

Google Play Store team is about to launch a new feature wherein, certain apps that are promising, are boosted in visibility. The boost will manifest in multiple ways including higher priority in recommendations sections (“Similar apps”, “You might also like”, “New and updated games”). These will also get a boost in search results visibility. This feature will help bring more attention to newer apps that have the potential. Mobile applications have become an integral part of modern life, offering a wide range of services and entertainment to users. App ratings and reviews play a critical role in influencing user decisions, but the process of predicting an app's rating remains a challenging task. The problem at hand is to develop an accurate and reliable predictive model for estimating app ratings based on various attributes and features.

**3.2 PROJECT OVERVIEW**

The problem is to identify the apps that are going to be good for Google to promote. App ratings, which are provided by the customers, is always a great indicator of the goodness of the app. The problem reduces to: predict which apps will have high ratings.

Content: Dataset: Google Play Store data (“googleplaystore.csv”) Fields in the data –

• App: Application name Object/string data type

• Category: Category to which the app belongs Object/string data type

• Rating: Overall user rating of the app Object/string data type replace by as type(float)

• Reviews: Number of user reviews for the app Object/string data type replace by as type (int/float)

• Size: Size of the app Object/string data type replace by as type(int/float)

• Installs: Number of user downloads/installs for the app Object/string data type replace by as type(int)

• Type: Paid or Free

• Price: Price of the app

• Content Rating: Age group the app is targeted at - Children / Mature 21+ / Adult

• Genres: An app can belong to multiple genres (apart from its main category). For example, a musical family game will belong to Music, Game, Family genres.

• Last Updated: Date when the app was last updated on Play Store

• Current Ver: Current version of the app available on Play Store

• Android Ver: Minimum required Android version

Enhancing User and Developer Experiences" seeks to leverage data science and machine learning to improve the mobile app ecosystem. By providing accurate app rating predictions and insights, this project aims to empower both users and developers to make informed decisions that ultimately enhance user satisfaction and app quality.

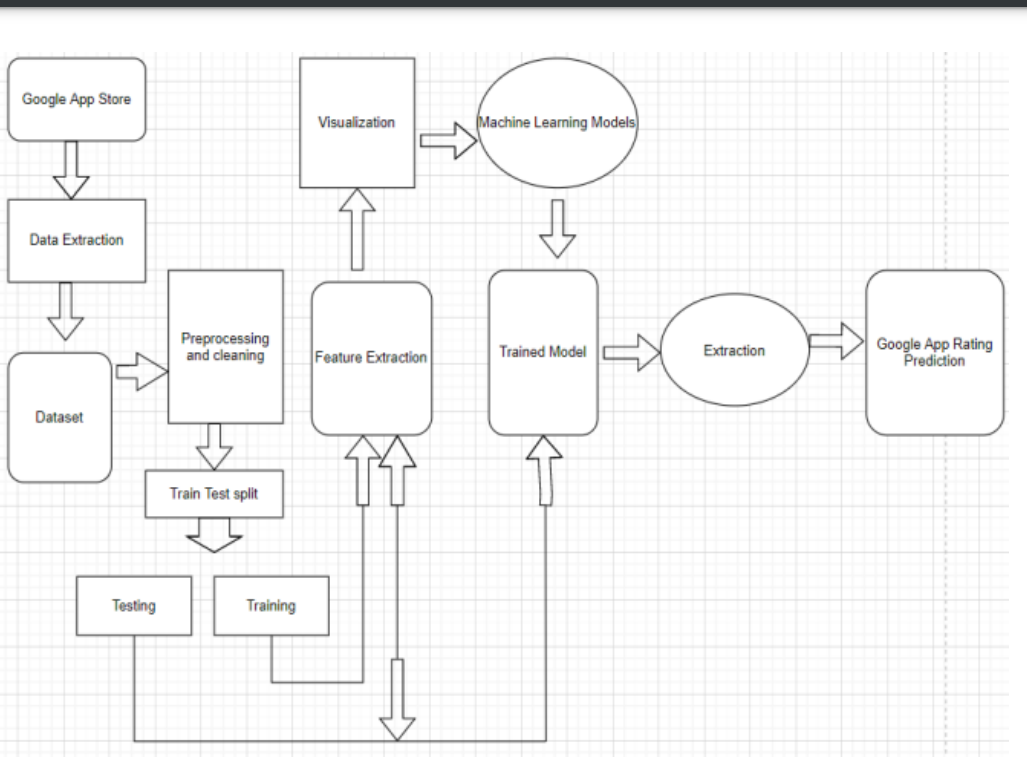
**3.3 ALGORITHM**

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not. It is a kind of statistical algorithm, which analyze the relationship between a set of independent variables and the dependent binary variables. It is a powerful tool for decision-making.

I have used Logistic Regression for Predicting App Ratings. It just the method of reviewing the dataset into the which a single or multiple distinct factors have impact on the outcome. The outcome is evaluated using a variable that is either true or false. The goal of the logistic regression is the to select the top model for explaining the association between such a number of autonomous factors and an interesting dualistic traits of interests.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

**3.4 SYSTEM ARCHITECTURE**



**Fig 3.4 System Architecture**

**3.5 MODULES**

The modules used for code are

* + Import libraries
  + Data Exploration
  + Data pre-processing
  + Model training
  + Model testing
  + Model Evaluation
  + Building a predictive system
* **Import libraries**

Python libraries make it very easy for us to handle the data and perform typical and complex tasks with a single line of code.

* [**Pandas**](https://www.geeksforgeeks.org/python-pandas-dataframe/)– This library helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.
* [**NumPy**](https://www.geeksforgeeks.org/python-numpy/)**–** NumPy arrays are very fast and can perform large computations in a very short time.
* [**Matplotlib**](https://www.geeksforgeeks.org/matplotlib-tutorial/)**/**[**Seaborn**](https://www.geeksforgeeks.org/introduction-to-seaborn-python/)– This library is used to draw visualizations.
* **Sk learn** – This module contains multiple libraries having pre-implemented functions to perform tasks from data pre-processing to model development and evaluation.
* **Data Exploration**

 Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

* **Data pre-processing**

In the second part, we will perform data preprocessing to treat outliers, impute missing values, remove duplicate entries, Gaussian transformation for a non-normal distribution, balance the imbalanced data, feature scaling (standardization and normalization), feature binning (converting continuous data to categorical values), feature encoding (converting categorical data to continuous values), and so on.

There are lots of tasks to consider for preprocessing, and this is the most important and time-consuming task of all the stages in machine learning. The data which is obtained from the primary sources is termed the raw data and required a lot of pre-processing before we can derive any conclusions from it or to some modelling on it. Those pre-processing steps are known as[data cleanin**g**](https://www.geeksforgeeks.org/data-cleansing-introduction/) and it includes, outliers removal, null value imputation, and removing discrepancies of any sort in the data inputs.

* **Model training**

Now we will separate the features and target variables and split them into training and the testing data by using which we will select the model which is performing best on the validation data.

* **Model testing**

Testing forms an integral part of any software development project. Testing helps in ensuring that the final product is by and large, free of defects and it meets the desired requirements. Proper testing in the development phase helps in identifying the critical errors in the design and implementation of various functionalities thereby ensuring product reliability.

* **Model evaluation**

From the above accuracies, we can say that Support Vector Machine classifier perform better on the validation data with less difference between the validation and training data. Let’s plot the [confusion matrix](https://www.geeksforgeeks.org/confusion-matrix-machine-learning/)as well for the validation data using the Logistic Regresssion model.

* **Building a predictive system**

Predictive analytics involves certain manipulations on data from existing data sets with the goal of identifying some new trends and patterns. These trends and patterns are then used to predict future outcomes and trends. By performing predictive analysis, we can predict future trends and performance. It is also defined as the prognostic analysis; the word prognostic means prediction. Predictive analytics uses the data, statistical algorithms and machine learning techniques to identify the probability of future outcomes based on historical data.

**CHAPTER 4**

**CONCLUSION**

We came to 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. the "Predicting App Ratings" project represents a significant stride in harnessing the power of data science and machine learning to enhance the mobile app ecosystem. By developing an accurate and reliable predictive model for estimating app ratings, this project has addressed the critical need for a solution that benefits both users and developers.

**CHAPTER 5**

**BIBLIOGRAPHY**

[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 twenty two 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. Thiviya et al., Mobile apps feature extraction supported user reviews using machine learning, 2019.

**APPENDIX**

**A.SOURCE CODE**

**IMPORTING LIBARIES**

Import pandas as pd

Import numpy as np

Import seaborn as sns

From sklearn import metrics

From sklearn .model\_selection import train\_test\_split

Import random

Import matplotlib.pyplot as plt

%matplotlib inline

Df =pd.read\_csv(‘/content/googleplaystore.csv’)

Df.head()

Df.info()

df.dropna(inplace = True)

df.describe()

df.isnull().sum()

**#data explorataion**

plt.rcParams['figure.figsize']=[20,5]

sns.heatmap(df.isnull())

plt.show()

x=df['Rating']

y=df['Installs']

plt.scatter(x,y)

plt.show()

x=df['Category']

y=df['Installs']

plt.bar(x,y)

plt.show()

**# preprocessing**

from sklearn.preprocessing import LabelEncoder

le=LabelEncoder()

df['Category']=le.fit\_transform(df['Type'])

df

**#training testing**

from sklearn.model\_selection import train\_test\_split

x= df [['Reviews']]

y=df['Reviews']

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,train\_size=0.25,random\_state=1)

**#Linear regression**

from sklearn.linear\_model import LinearRegression

log\_reg=LinearRegression()

log\_reg.fit(x\_train,y\_train)

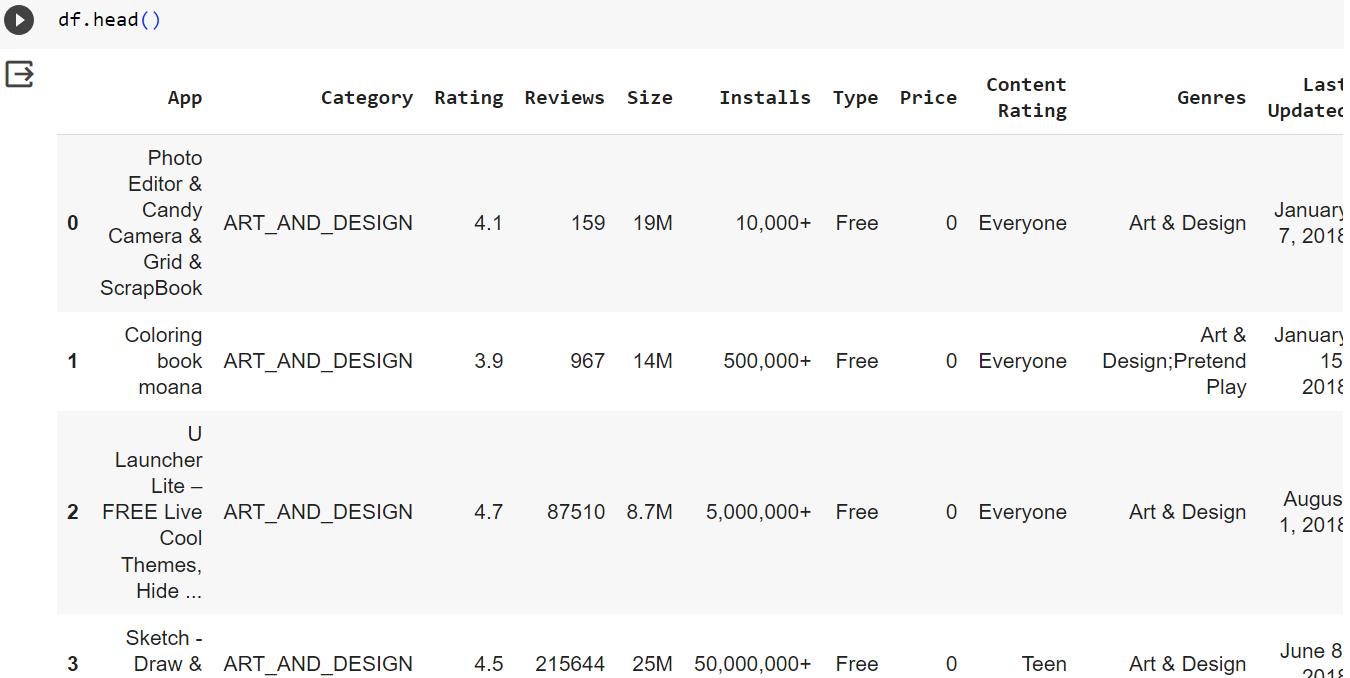
**#accuracy score**

log\_reg.score(x\_train,y\_train)

log\_reg.score(x\_test,y\_test)

sns.distplot((y\_test-y\_pred),bins=50);

**B.OUTPUT SCREENSHOTS**

****

**Fig.B.1 First five rows of dataset**

**A screenshot of a computer

Description automatically generated**

**Fig.B.2 Data Exploring by heat map**

**A screen shot of a graph

Description automatically generated**

**Fig.B.3 Data Exploring by Scatter Plot**

**A graph with blue bars

Description automatically generated with medium confidence**

**Fig.B.4 Data Exploring by Bar chart**

**A screenshot of a computer

Description automatically generated**

**Fig.B.5 Data PreProcessing**

**A screenshot of a computer program

Description automatically generated**

**Fig.B.6 Accuracy Using Linear Regression**

**A screenshot of a computer

Description automatically generated**

**Fig.B.7 DistPlot using Linear Regression**