

Machine Learning

Machine Learning Assignment

**Module Co-ordinator: Besher Alhalabi**

Name: Muhammad Nawab Anjum

ID: 21171479

Email:Muhammad.anjum2@mail.bcu.ac.uk

Table of Contents

[Machine Learning Subfields 1](#_Toc90027400)

[Domain Description 2](#_Toc90027401)

[Problems Definition 3](#_Toc90027402)

[Scenario 1 3](#_Toc90027403)

[Dataset Description 3](#_Toc90027404)

[Dataset Exploration (pre-processing and wrangling) 3](#_Toc90027405)

[Main Python libraries that I have used for this project. 4](#_Toc90027406)

[Data Cleaning 5](#_Toc90027407)

[Different Approaches to overcome the Outlier 6](#_Toc90027408)

[Data Visualisation 6](#_Toc90027409)

[Histograms 6](#_Toc90027410)

[BarPlot 7](#_Toc90027411)

[Box Plot 7](#_Toc90027412)

[Pie Chart 8](#_Toc90027413)

[Feature Selection 9](#_Toc90027414)

[Coloration b/w features 9](#_Toc90027415)

[Correlation Analysis. 9](#_Toc90027416)

[Experiment and Evaluation 10](#_Toc90027417)

[Supervised learning 10](#_Toc90027418)

[Analysis and Results 10](#_Toc90027419)

[Conclusions 10](#_Toc90027420)

[References 11](#_Toc90027421)

[Appendixes 11](#_Toc90027422)

# Machine Learning Subfields

Machine learning and artificial intelligence is most advanced technologies. Machine learning is a branch of AI artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that human lean, gradually improving its accuracy. (IBM).Machine learning algorithm is vastly used to train data and get exact result according to our requirements. Most popular Google search engine used machine learning algorithm. Social networking sites also used machine learning algorithm, Facebook, Twitter, YouTube, Instagram, and email spam filter is also used to save time. (Das & Dey, 2015)

There are many machine learning techniques that are as follows:

1. Machine learning (Supervised & Unsupervised Learning)
2. Inductive Deductive Learning & reinforcement learning
3. Deductive Transductive Learning algorithm
4. Multi task and active learning
5. Transfer and Online learning

Mostly the supervised and unsupervised technique is used in our daily lives. Some other learning algorithm are being used for different purposes and have different perspective. Deep learning and its other part is also a machine learning artificial intelligence neural network part. Artificial algorithm is sub domain of other domain and its work together to perform task. The classification of image processing using deep learning algorithm either the person is male and female, face recognition and image labeller. Artificial neural network can train large set of images datasets and its easy method to identify images. (Machine Learning Mastery, 2021)

# Domain Description

Mobile apps are one of the fastest growing segments in the downloadable apps market. Of all the markets, we choose the Google Play Store because of its growing popularity and its recent rapid growth. One of the main reasons for this popularity is that about 81% of applications are free. With titan companies like Samsung, LG, Motorola and HTC all launching Android phones, it soon became the most popular mobile operating system, reaching more than one billion active users by 2014.

As Android grew, so did Google's control over the operating system. Initially, manufacturing partners were able to customize most of the platform to their liking; however, Google has added more mandatory services and conditions each year, ensuring benefits for its own app package.

Google Play has grown tremendously over the last decade, reaching $ 38.6 billion by 2020. By 2020, more than 2.9 million apps were available in stores, downloaded 108 billion times.

Developers and consumers play a critical role in determining how industry impacts affect future technologies. However, the lack of a clear understanding of the inner workings and functionality of the popular software industry affects developers and users alike. In this article, we will try to explain the strengths of the Google Play Store and how we can use various configuration files for prediction purposes.

In this article, I provide a longitudinal study of the metadata for Google Play applications, which provides unique information that is not available with a standard approach to take a single application. Use the feature longitudinal analysis of the application to determine whether or not an application will be successful. Our analysis is divided into four phases: data extraction, data purification, data visualization, and application of different models. I will first collect data from the Kaggle website. In the next step I will try to delete the data from the dataset to reduce the error rate. Once the dataset is complete, we try to analyse the dataset using different maps and replace the unnecessary things in the dataset. In the last step, I will use different supervise and unsupervised algorithms on the data set to find out which offers the highest percentage accuracy. Finally, I will present the results of the analysis to provide a clear picture of the relationship between interests. The relevance and future directions of research are discussed in detail in the last chapter, Conclusions and future papers.

# Problems Definition

## Scenario 1

App analysis will be used to ﬁnd whether an app will be successful or not?

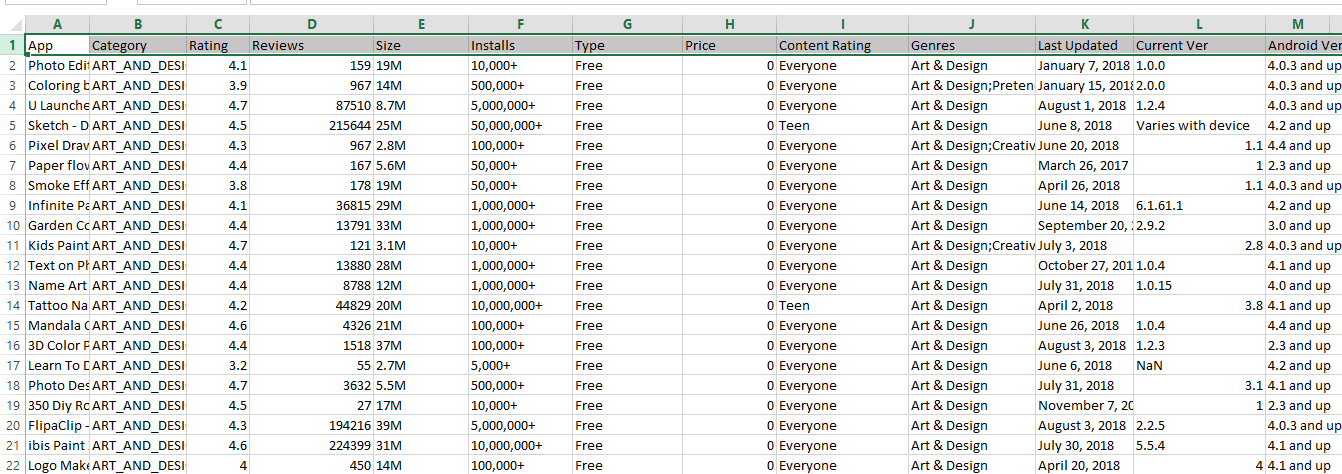
# Dataset Description

**Oxford Dictionary** defines a dataset as **“a collection of data that is treated as a single unit by a computer”.** This means that a dataset contains a lot of different data but can be used to train algorithms for the purpose of finding predictable patterns throughout the dataset.

In my case I have download the dataset from the [Kaggle](https://www.kaggle.com/lava18/google-play-store-apps/code) this is a web scrapping repository where [Kaggle](https://www.kaggle.com/lava18/google-play-store-apps/code) user contribute such kind of data. My Play Store apps dataset has total of 10841 rows and 13 columns.

The columns of the dataset are as follows:

1. App (Name) The name of the App on the play store
2. Category (App)
3. Rating (App)
4. Reviews (User)
5. Size (App)
6. Installs (App)
7. Type (Free/Paid)
8. Price (App)
9. Content Rating (Everyone/Teenager/Adult)
10. Genres (Detailed Category)
11. Last Updated (App)
12. Current Version (App)
13. Android Version (Support)



Sample Dataset

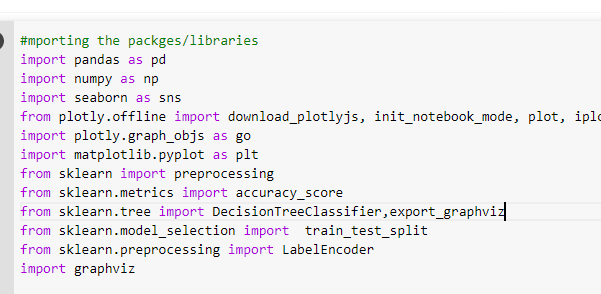
# Dataset Exploration (pre-processing and wrangling)

Most of the time, the data come with garbage/waste values, which need to be addressed before affecting performance. Trained models that predict outcome. Several steps are used to pre-process this data. Reprocessing is important into transitioning raw data into more desirable format. Undergoing the pre-processing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need pre-processing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e. there can be some missing values.

In this Section, I will discuss more about Google Play Store Apps Based on the Given Information, using the notebook I have created to understand my dataset.

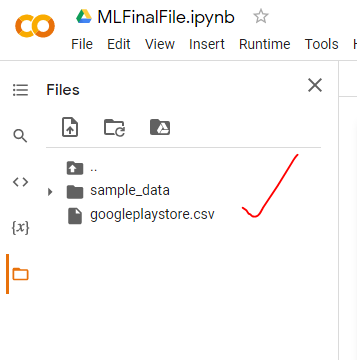
## Main Python libraries that I have used for this project.

* **Pandas**: The library is highly optimized for performance, with critical code paths written in python or C. It deal with DataFrame,time series data, reading and writing data between in-memory data structures and different file formats.
* **Scikit Learn**: This machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines. Also ,it can work with other libraries and packages like Numpy.
* **Graphviz:** Graphviz is an open-source python module that is used to create graph objects which can be completed using different nodes and edges. It is based on the DOT language of the Graphviz software and in python it allows us to download the source code of the graph in DOT language.
* **NumPy:** Is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
* **Matplotlib:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.
* **Seaborn:** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.



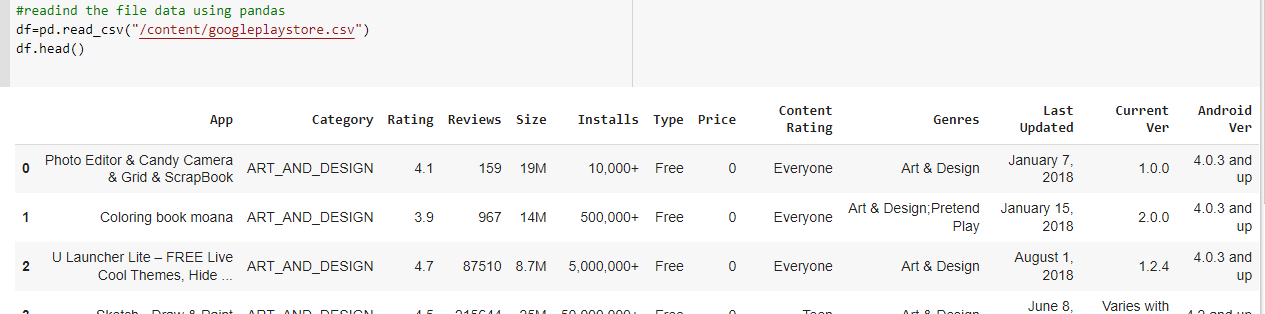
Libraries importing

The first step for any data science project is to enter your data. Often you will be dealing with data in Comma Separated Value (CSV) files. Now next main task is to read the dataset using **Pandas** Read**\_**CSV It is an easy way to read the CSV file data as a dataset. Before you can use Panda to import your data, you need to know where your data is in your file system and what your current task list is. I’m my case I have upload the CSV file into local directory of the **Colab**



df=pd.read\_csv("/content/googleplaystore.csv")

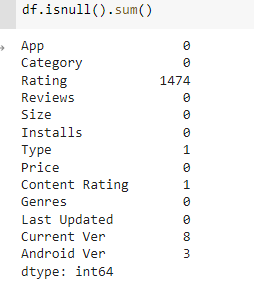
After reading the file, just checking that file is read or not by printing the head of the dataset.



Now I’m successful to reading the file, next step is star to clean the imported data.

## Data Cleaning

Checking the total null values, I’m using the isnull.sum() method this method will return the sum of missing values



In the above picture we can see that Rating field has 1474 Null values type has 1 and content rating 1 as well as android ver 3. Category has 1.9 that is considered as a mislabelled ,so we have to remove this using the below code

mislabel = df.loc[df["Category"] == "1.9"]

df = df.drop(int(mislabel.index.values),axis=0)

print(df["Category"].unique())

Also Rating has some null values so removing the null values from the dataset.

#dropping null values

df = df.drop(df[df['Rating'].isnull()].index, axis=0)

## Different Approaches to overcome the Outlier

* **Mean:** If we have a sample of numeric values, then its mean or the average is the total sum of the values (or observations) divided by the number of values
* **Median:** The median of a sample of numeric data is the value that lies in the middle when we sort the data. The data may be sorted in ascending or descending order, the median remains the same.
* **Mode:** The mode is the most frequent observation (or observations) in a sample. If we have the sample [4, 1, 2, 2, 3, 5], then its mode is 2 because 2 appears two times in the sample whereas the other elements only appear once.

## Data Visualisation

Data visualization in python is perhaps one of the most utilized features for data science with python in today’s day and age.

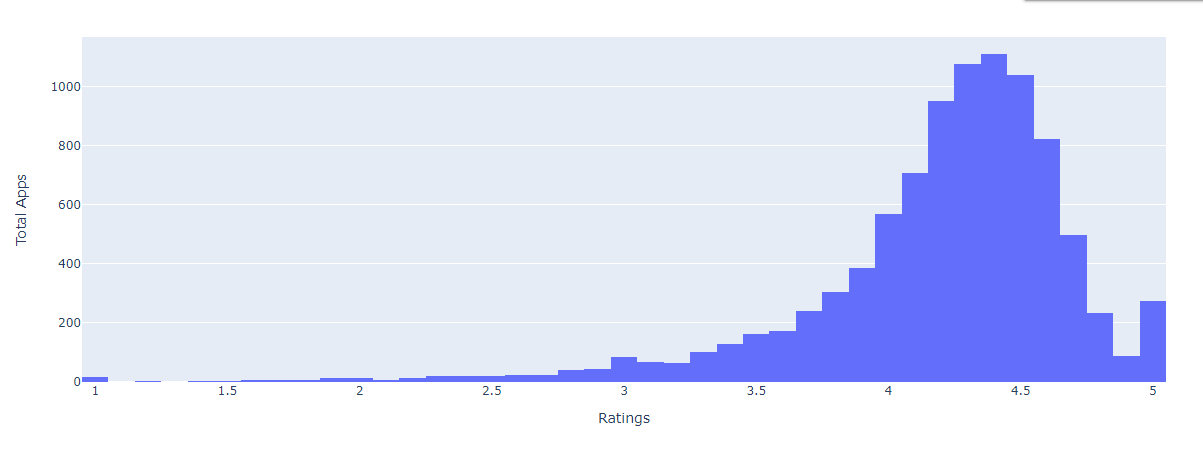
A good visualization must follow the **Gestalt Laws.**

* **Law of Proximity**: We view objects close to each other as belonging to a group.
* **Law of Similarity**: We look for likeness and differences in objects and link similar objects as belonging to a group.
* **Law of Closure**: Our minds tend to see complete figures or forms even if a picture is incomplete.
* **Law of Enclosure**: We view objects as belonging to a group when they are enclosed in a way that creates a boundary or border around them.
* **Law of Connectedness**: We view objects connected to each other as a single group as opposed to objects that are not linked in the same way.
* **Law of Continuity**: Our propensity is to see shapes as continuous to the greatest degree possible. The human eye follows lines, curves or a sequence of shapes to create pathways.

The libraries in python come with lots of different features that enable users to make highly customized, elegant, and interactive plots

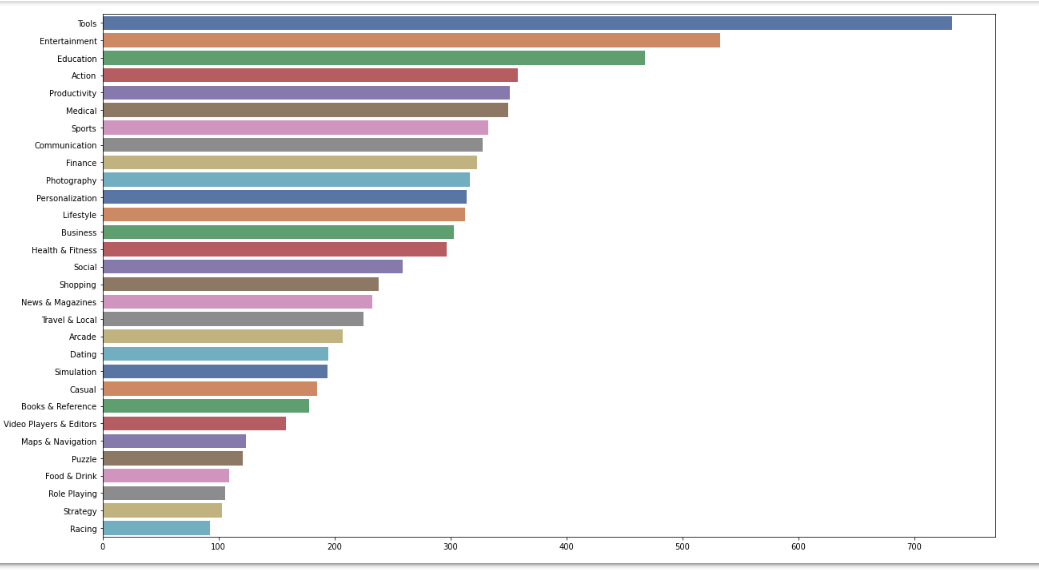
### Histograms

A histogram is a graph showing frequency distributions. It is a graph showing the number of observations within each given interval.



In this chart I’m showing the Total Apps on yaxis and Rating on xaxis. There are total almost 1099 application that are coming under the 4.5 rating and 275 is rated as 5 , whereas 27 applications is in 1 rating that’s are fail/flops or bad in the play store.

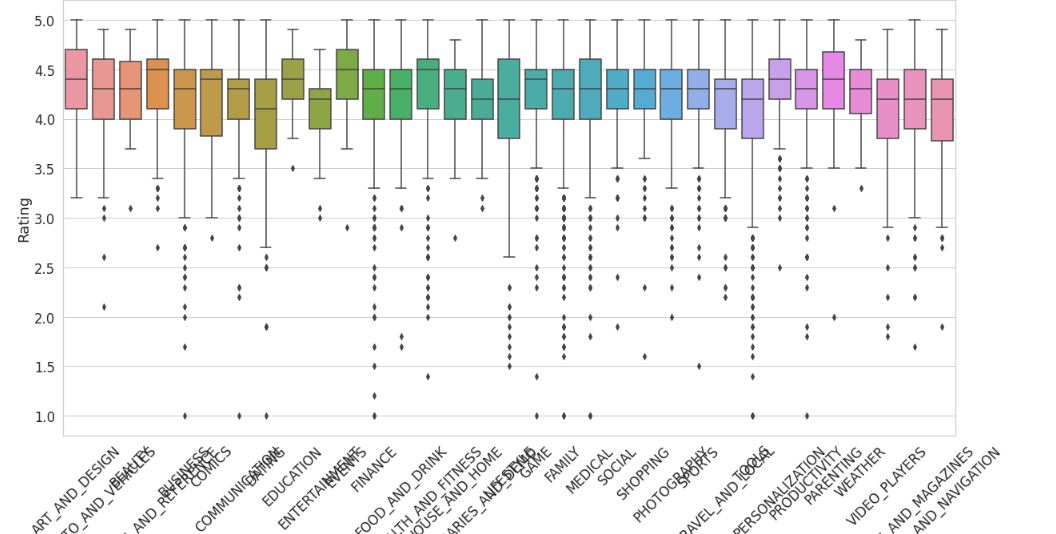
### BarPlot

A barplot shows the relationship between a numeric and a categorical variable. Each entity of the categorical variable is represented as a bar. The size of the bar represents its numeric value

In my case I did Show top 30 app genres with different colour Tools, Entertainment and education category is the main category that are the most top category in the playstore.

### Box Plot

A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum. In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median. Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution.

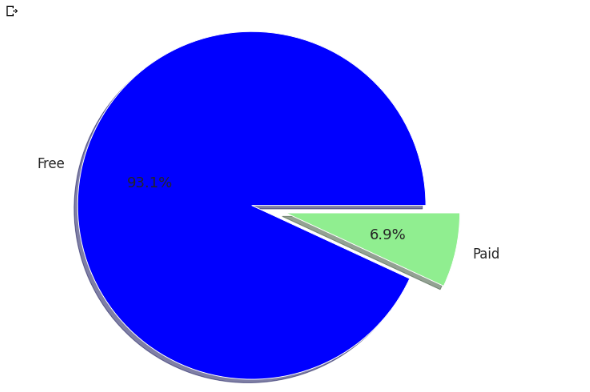


I did show the Rating and category in BoxPlot, xaxis showing the Category whereas yaxis showing the rating. As we can see that the graph showing **4.5** rating the maximum rating of an application that is considered as a good or successful application in the playstore.

### Pie Chart

A pie chart show a part-to-whole relationship in our data. Each slice represents one component and all slices added together equal the whole. Useful to show up data that is classified into nominal or ordinal categories.

As we can see in our Pie chart, we can see that 93 %( Approx.) Of apps in the google play store are free and 7 %(Approx.) Are paid



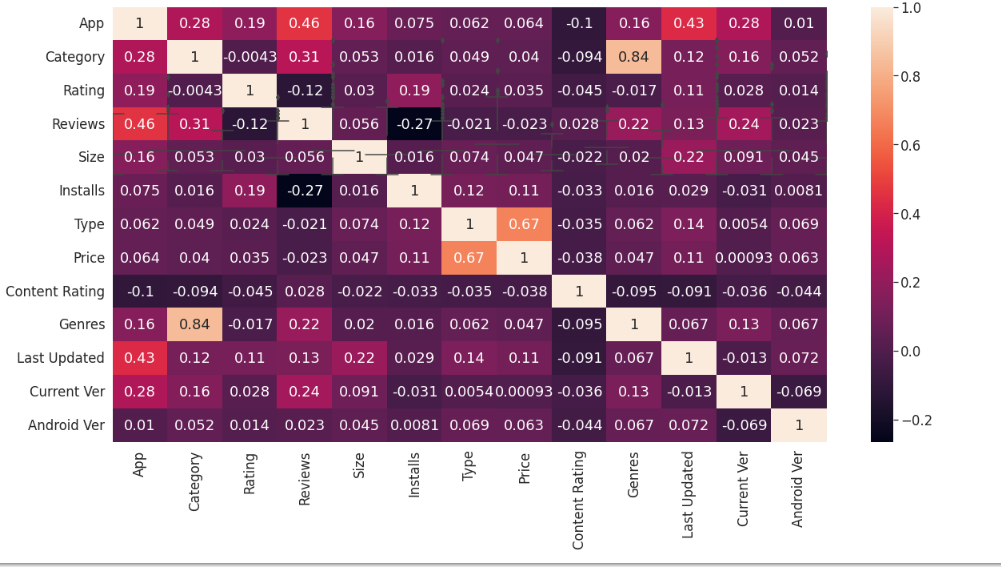
## Feature Selection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model

## Coloration b/w features

Correlation is a statistical term which in common usage refers to how close two variables are to having a linear relationship with each other.

Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features.



Correlation Matrix

Correlation is mostly used in Machine leering to figure out the least correlating variable that tends to be the best for classification.

### Correlation Analysis.

We used correlation for two purposed Feature engineering, regression

* **Feature engineering:** To optimize feature , like we have too much feature in our dataset but we are focusing on related to our business
* **Regression:** To know the relation b/w variable

At the end of correlation we will found the 1(positive) or -1(negative) values or 0 these are some term that are mostly using **Positive correlation,** **Negative Correlation, Zero Correlation**

* **Positive correlation :** When value of A increase the value of B must be increased
* **Negative Correlation:** When value of A increase the value of B decreased
* **Zero Correlation:** not relationship

# Experiment and Evaluation

Now this is our main concern and topic, for that I have created a complete description about **ML**,**Dataset**,**Domain**, **Visualisation** etc. Our main concern here to implement at least 3 algorithm which must be from the both **supervised** and **unsupervised** .

## Supervised learning

Let’s talk about some basics about supervised machine learning. **Supervised learning** machine learning is a type of machine learning, The main goal in supervised learning is to learn a model from labled training data that allows us to make predictions about unseen or future data, also the desired output labels are already known. Examples of Supervised Learning are ,Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

There are two main terms are being used Independent, dependent Variables,

* **Independent** **Variables:** Independent variables that do not have any dependence to another variables.Another name for independent variables is Predictor variables. In data, independent variables will also be known as regressors, controlled variables, manipulated variables, explanatory variable, exposure variable, and/or input variable.
* **Dependent Variables:** These variables are opposite to independent variable the outcome of these variables is totally depends on the Independent variables. Another name for dependent variables is Predicted variables. In data, dependent variables will also be known as response variable, regressand, measured variable, observed variable, responding variable, explained variable, outcome variable, experimental variable, and/or output variables

# 

# Analysis and Results

# Conclusions

# References

<https://www.python-graph-gallery.com/barplot/>

https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf

# Appendixes