

# Forensic Analysis Of Social Media Applications Using Machine Learning And Data Science

**Naveen Lakhchaura**

Maharaja Agrasen Institute of Technology  
E-mail address: [naveenlakhchaura1998@gmail.com](mailto:naveenlakhchaura1998@gmail.com)

## HIGHLIGHTS

- Presented a theoretical approach of modelling data on social media application for business.
- Identification of Relation among the use of Social media applications.
- Analysis of Data on Social Media Applications.
- Statistical approach to identify the solution to optimize it.

**Abstract:** *In this paper I present a method of analyzing social media applications and then using that for enhancing it for business purpose ,enhance user Experience on that social media .The data that has been used on the ML model is self collected by real user through google forms .Data Analytics has also been performed on it using the tableau Software. I have used a lot of ML algorithms on my ML model to find the algorithms with the best results .However the model that is been made here can be used in a lot of different datasets by performing certain changes in data preprocessing.*

## 1.Introduction

Social media Application are playing an important role in any person regular life they are key component in most of the peoples life .Many of these social media company are earning a lot of money through it using ads and various other approaches.Some of the most common social media applications include Instagram,facebook,twitter.There are a number of ways through which money can be made by these applications and start-up market are curiously looking in invest in such applications. In this paper I purpose a ML model that can help us to enhance market of such application and knowing approaches to improve user experience to these social media applications.

## 2. Data Preprocessing

Data Processing was an important part of my project because wee needed some actual user driven data can can help me in training our model using a ML classifier. This data was collected through a google form that was distributed among student so that they can fill it.

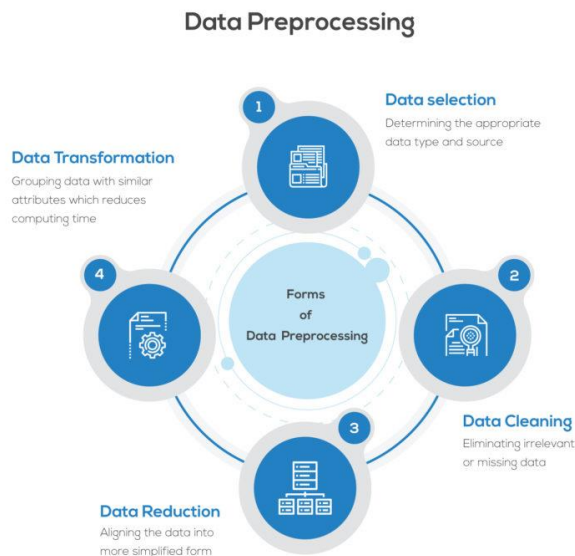
Following Column were the part of our dataset-

- 1.Favourite App
- 2.Daily user
- 3.Number of hours spend.
- 4.Like ads
- 5.Bored of Social Media

Names and Email id were also collected as a proof of identity of the user from which data is collected.

There were Several Steps in Pre-Processing a data-

- Importing the Data.
- Fix Missing Value in Data.
- Mapping the data.
- Dividing the data categorically.
- Dividing the data in Training set as well as Testing set.



A number of Python Libraries were imported in making this Data preprocessing possible-

- Numpy
- Pandas
- Scikit-Learn
- Matplotlib

as ask this data processing part was very crucial sowe also performed data analytics on that same data set using the tableausoftware.this will be discussed further in the later part of the paper.

In date preprocessing the data set is usually divided into training set and test set.  
Training set is used to train the dataset and test set is used to test the correction of the model.

### 3. Machine Learning Algorithms

As our Dataset is a Churn Modelling Example where we decide whether a certain thing will happen or not using a ML algorithm.

A lot of ML algorithms can be used to achieve this but we used a few and found the new best suited for it-

#### 1. KNN-

KNN is one of many algorithms used in data mining and machine learning, it's a classification algorithm where the learning is based how data is together patterned or connected and here data has a category.

**The KNN's steps are:**

1 — Receive an unclassified and unarranged data in csv or excel form;

2 — Measure the distance of the new data to everyothers data that is already classified using the

Model.

3 — Gets the K value so can decide it yourself.

4 — Check the list of classes had the shortest distance and count the amount of each class

5 — Takes as much as correct class the class that appeared the most times in our model

computations;

6 —Classifies the new data with the class that you took above in step 5;

Source-<https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d>

## 2. SVM-

A Support Vector Machine (SVM) is classification algorithm previously defined by a separating hyperplane the algorithm outputs an optimal hyperplane which divides new examples. In 2 dimensional space this hyperplane is a line separating a plane in two parts where in each class lay in either side of the plane.

### Kernel

The learning concept of the hyperplane in linear SVM is performed by transforming the problem using some linear algebra in our algorithms.

This is where the kernel plays an important role.

For the **linear kernel** the equation for prediction for a new input using the dot product between the input (x) and each support vector is measured as follows:

$$f(x) = B(0) + \sum(a_i * (x, x_i))$$

Given equation that involves calculate the inner products of a new input vector (x) with help all support vectors in training the data. The coefficients B0 and ai (for each input) must be estimated from the training data by our learning algorithm.

The **polynomial kernel** can be written as  $K(x, x_i) = 1 + \sum(x * x_i)^d$  and **exponential** as  $K(x, x_i) = \exp(-\gamma * \sum((x - x_i)^2))$ . [Source for this excerpt : <http://machinelearningmastery.com/>].

Polynomial and exponential kernels calculates separation line in higher dimension. This is called **kernel trick**

Source-<https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>

#### 4. Tableau

Data Analytics was also performed on the dataset using the software tableau which was provided to me in student category at free of cost.

Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data into the very easily understandable format.

Data analysis is very fast with Tableau and the visualizations created are in the form of dashboards and worksheets. The data that is created using Tableau can be understood by professional at any level in an organization. It even allows a non-technical user to create a customized dashboard.

The best feature Tableau are

- Data Blending
- Real time analysis
- Collaboration of data

The great thing about Tableau software is that it doesn't require any technical or any kind of programming skills to operate. The tool has garnered interest among the people from all sectors such as business, researchers.



#### 5. Artificial Neural Network-

An artificial neuron network (ANN) is a computational model often in deep learning based on the structure and functions of neural networks which is inspired by real biological human neural system of our brain. Information that flows through the network usually affects the complete structure of the ANN because a neural network changes in a sense - based on that input and output.

ANNs are considered nonlinear statistical data mode; where relationship between I/O is modelled and patterns are found among them. ANN is also known as a neural network.

An ANN has several advantages but one of the most recognized of these is the fact that it can actually learn from observing data sets that we pass in our neural network. In this way, ANN is used as a random function approximation tool in deep learning.. These types of tools help estimate the most cost-effective and ideal methods for arriving at solutions while defining computing functions or distributions which is the best choice here we can find the best solution by trying a lot of other ones. ANN takes data samples rather than entire data sets to arrive at solutions, which saves both money as well as time of the model and save our computational power. ANNs are considered fairly simple mathematical models to improve current data analysis technologies.

ANNs have three layers that are interconnected together. The first layer usually consists of input neurons. Those neurons send data on to the second layer, which then sends the output neurons to the third layer.

Training an artificial neural network involves choosing from allowed models for which there are several associated algorithms.

#### 6. Conclusions-

In my Project I made a csv file of the data collected named by **ProjectDataa.csv** a python file where my codes has been written. They both can be found at my github.

I preprocessed my data here dividing it into independent variables and dependent variables X and Y.

Then this X and Y is divided into training set and testing data.

A Artificial neural network has been made to train our model using sequential and dense and by following llibraries- and then our model was tested for a random value and in result it gave that our user will get bored of social media applications.

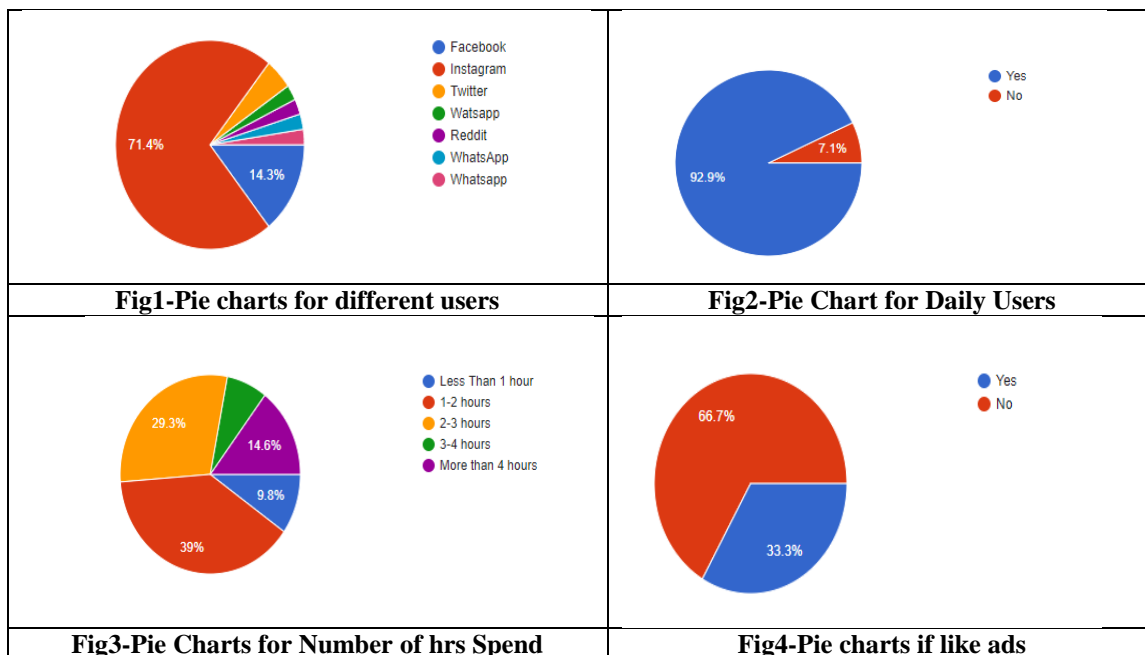
And then with the help of confusion matrix we checked the correction of our model and it was quite good.

And among SVM and KNN neighbour algorithm we found our SVM was most accurate having more correct values as compared to KNN.

I have decided to perform that data analytics on my dataset to show some visualization of my data using tableau and it will be

Available soon for study.....

Some of the Visualization from Google analytics has been pasted here..



## References

- Medium
- Udemy
- Geek For Geek
- Wikipedia
- Google Forms