

**PROFIT PREDICTION USING DATA MINING ALGORITHMS**

**By**

**FATIMA AKTER MUN**  
**ID: 151-15-5142**

**SAYED MAHMUD**  
**ID: 152-15-5607**

**TABASSUM MARJIA**  
**ID: 152-15-5988**

**MD. TAREKUL ISLAM**  
**ID: 152-15-5989**

This Report Presented in Partial Fulfillment of the Requirements for the  
Degree of Bachelor of Science in Computer Science and Engineering.

Supervised by

**RUBAIYA HAFIZ**  
Lecturer  
Department of CSE  
Daffodil International University

Co-Supervised by

**MD. Zahid Hasan**  
Assistant Professor  
Department of CSE  
Daffodil International University




**DAFFODIL INTERNATIONAL UNIVERSITY**  
**DHAKA, BANGLADESH**  
**MAY 2019**

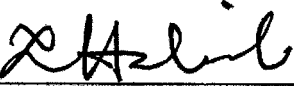
## **APPROVAL**

This Project titled “**Profit Prediction using Data Mining**”, submitted by Fatima Akter Moon 151-15-5142, Sayed Mahmud 152-15-5607, Tabassum Marjia 152-15-5988 and MD. Tarekul Islam 152-15-5989 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents.


## **BOARD OF EXAMINERS**

  
\_\_\_\_\_  
**Dr. Syed Akhter Hossain**  
**Professor and Head**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University


**Chairman**

  
\_\_\_\_\_  
**Md. Tarek Habib**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**

  
\_\_\_\_\_  
**Moushumi Zaman Bonny**  
**Senior Lecturer**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**

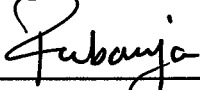
  
\_\_\_\_\_  
**Dr. Swakkhar Shatabda**  
**Associate Professor**  
Department of Computer Science and Engineering  
United International University

**External Examiner**

## DECLARATION

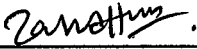
We hereby declare that, this project has been done by us under the supervision of **Rubaiya Hafiz, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



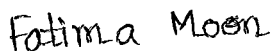
**Rubaiya Hafiz**  
Lecturer  
Department of CSE  
Daffodil International University

Co-supervised by:

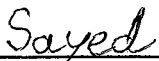


**MD. Zahid Hasan**  
Assistant Professor & Associate Head  
Department of CSE  
Daffodil International University

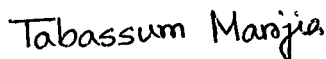
Submitted by:



**Fatima AkterMun**  
ID: 151-15-5142  
Department of CSE



**Sayed Mahmud**  
ID: 152-15-5607  
Department of CSE



**TabassumMarjia**  
ID: 152-15-5988  
Department of CSE



**MD. Tarekul Islam**  
ID: 152-15-5989  
Department of CSE

## ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Rubaiya Hafiz, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Data Mining*” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Syed Akhter Hossain, Professor and Head**, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

## **ABSTRACT**

Now a days everyone is trying out their luck with e-commerce websites and business. But they are not getting their expected results because their sell is sometimes not up to the mark. In an e-commerce business stocking up the right materials is everything. So, our solution to this problem is running data mining methods to figure out which products are likely to sell the most and result in better profit making. We will use different classifiers to build our models and compare their results to better determine which classifier gives the most accurate prediction so that we can use that classifier to build our final system in prediction making process.

# TABLE OF CONTENTS

<b>CONTENS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>8-10</b>
1.1 Introduction	8
1.2 Motivation	8
1.3 Objectives	9
1.4 Expected Outcome	9
1.5 Report Lawet	9-10
<b>CHAPTER 2: BACKGROUND</b>	<b>11-13</b>
2.1 Introduction	11
2.2 Related works	11-12
2.3 Research summary	12
2.4 Scope of the problem	13
2.5 Challenges	13

<b>CHAPTER 3: REQUIREMENTS ANALYSIS FOR THE PROPOSED SYSTEM</b>	<b>15-24</b>
3.1 Introduction	15
3.2 Proposed System Architecture	15
3.3 Data Sets	15-17
3.4 Data Mining Tools	17
3.4.1 Python	17
3.4.2 Sci-Kit Learn	17-18
3.4.3 NumPy	18-19
3.4.4 Pandas	19
3.5 Data Mining Classifiers	19-32
<b>CHAPTER 4: IMPLEMENTATION AND TESTING</b>	<b>33-36</b>
4.1 Implementation	33-34
4.2 Experimental Result	34-35
4.3 Comparison	35-36
<b>CHAPTER 5: CONCLUSION</b>	<b>37</b>
5.1 Discussion	37
5.2 Limitation	37
5.3 Scope for Future Works	37
<b>REFERENCES</b>	<b>38</b>
<b>APPENDIX</b>	<b>39</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.2: System Process	14
Figure 3.4.2: Sci-Kit for classification and regression	18
Figure 3.4.3: Binary Classification with NumPy	18
Figure 3.5.1: K-nearest neighbors Algorithm	19
Figure 3.5.2: Multiple decision boundaries	21
Figure 3.5.3: Decision boundary with support vector	22
Figure 3.5.4: Decision tree simplified	23
Figure 4.2.1: Website sign in page	30
Figure 4.2.2: website homepage	30
Figure 4.2.3: specification page	31
Figure 4.2.4: best sellers	31
Figure 4.2.5: website cart	32
Figure 4.2.6: Accuracy rates for all possible data	32
Figure 4.2.7: Accurate rates for original data	33



## LIST OF TABLES

### TABLES

### PAGENO

Table 3.1: Attribute List

4

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

We are living in the age of modern science and technology. Technology has become an integral part of our day to day life. One of those technological advancement is smartphones. Everybody uses one. There are various types of smartphones out there consisting of different price, quality and production. The options are limitless. But not all of them are profitable to sell because not all the smartphones have the same kind of demand. If we think from the point of view of a shopkeeper or an e-commerce business owner, he or she would only want to keep the profitable smartphones in stock. But how would they know which phone would be profitable to stock or which won't even before the sell starts?

We are focusing on creating a prediction system with which we would be able to determine if a phone would sell more and result in greater profit for an online business. The business would only stock the profitable smartphones which will result in less amount of loss and greater profit.

### 1.2 Motivation

Using modern science to better help out the economic situation of our country played a big role as our motivation. If this prediction process actually becomes fruitful then it will boost up the business of existing e-commerce owners as well as motivate other people to come in the business, make a name for themselves and help the country economy in the process.

### **1.3 Objectives**

Bangladesh has achieved all 3 conditions to overcome as a least developing country. Bangladesh has been recognized both economically and socially as a developing country in the last few years [1]. Information and Communication Technology or ICT has played a big role in this journey. Step by step both government and non-government organizations or NGO also private sector are growing rapidly using various technologies. Apart from ICT development our style has also changed. We are trying to incorporate tech in all the phases of our day to day things.

By using profit prediction in e-commerce sites, we will be able to contribute in the ongoing online boom. Everybody is trying their luck with online business platforms. But having no luck for lack of knowledge and proper guidance. Our objective is to make that easier for them and contribute to the financial development of Bangladesh.

### **1.4 Expected Outcome**

Our goal is to come up with a system that provides predictions on which smartphone is going to sell more on that particular time-period in real time. We would help hundreds and thousands of entrepreneurs out there trying hard to become self-dependent. With the help of this study we will be able to conclude which classifier gives the most accurate predictions when it comes to e-commerce. As a result, with the help of this paper, future researchers can also upgrade and modify their own system.

## **1.5 Report Layout**

There are five chapters in this paper and this phase will let we know which chapter has which aspect of the topic.

1. Chapter one consists of introduction, motivation and expected outcome.
2. Chapter two consists of background study.
3. In chapter three, necessary tools and systems needed for the study is discussed.
4. Chapter four of this report has experimental results, comparison and outcome.
5. Last chapter is on conclusion, limitations and future study.

## **CHAPTER 2**

### **BACKGROUND**

#### **2.1 Introduction**

Online shopping experience has changed vastly over the last two decades. Small and rookie websites selling products over an auction are now the big shot callers of the e-commerce industry and it is flourishing as we speak.

e-commerce websites rely a lot on data mining now a days. Converting data into information in a systematic manner is known to be data mining. For example- pattern mining, trend discovery, and prediction. We use data mining for fraud detection, suggestion and search in the e-commerce industry. [2]

#### **2.2 Related Works**

This is not a new concept. After the reveal of the idea of data mining there have been several implementations of this concept. It's just a new approach for profit prediction at basic level with real time visualizations.

The paper [3] describes, Inventory intelligence requires us to use data mining to process items and map them to the correct product category. This involves text mining, natural language understanding, and machine learning techniques. Successful inventory classification also helps us provide a better search experience and gives a user the most relevant product.

In the paper [4], today web is the best medium of communication in modern business. In order to improve output many companies are adopting new strategies. Business over internet provides the opportunity to customers and partners where their products and specific business can be found. Online business is not bound by time or space unlike physical stores. Many multinational organizations are now understanding the fact that e-commerce is not just buy or sell, rather it improves the efficiency to compete. For this purpose, data mining sometimes

called as knowledge discovery is used. Web mining is data mining technique that is applied to the WWW. There are vast quantities of information available over the Internet. [5], they created a model to try to help the investors in the stock market to decide the best timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks. The decision taken will be based on one of the data mining techniques; the decision tree classifiers. In the paper [6], it is said that The application of Web data mining technology is data mining in e-commerce site, the English name is “Web Data Mining”, a technique that is developed based on a Web environment. It is potentially useful model or information able to collect from complicated Web documents and sites. Web Data Mining technology is an integrated technology, not only related to the computer network technology and artificial intelligence technology, and also involves the discipline of computational linguistics, information science and statistics. Web data mining technology were used for three types of Web data forms: Web content mining, Web structure mining and Web usage mining methods. E-commerce is a wide range of business types now, and the orderly conduct of electronic commerce on the Internet can't be separated from the support of data mining technology. From the point of view of data mining technology, e-commerce has a sufficient condition for data mining (e.g.: the data source richly and reliable data automatically collected and other conditions).

### **2.3 Research Summary**

We have created a website of our own to do the research, so that we can have unlimited access to all the features before the system is ready for open trial. We have used JavaScript for website and python for mining. Our database can be both read and write by JavaScript and python. So, at real time, our system will read and write predictions on the website. We are also categorizing all the products, giving the system a chance to pin point and narrow down the search to be more accurate.

## 2.4 Scope of the Problem

We are creating a profit prediction system which will be applied in e-commerce website. By applying the system, only profitable and most likely to be sold more products will be shown to the customers which will result in greater profit.

### **Financial growth:**

If our system is being used it will certainly result in greater sells and less loss. So they will earn more and contribute in the financial growth of the country.

### **Less wastage:**

A lot of times as owners have no idea which product will sell and which will not, they buy a lot of unnecessary stuff which never get out of the shelf. It wastes precious shelf space and also as it does not sell, it's nothing but waste of invest money.

### **Better experience:**

It's not only a profit prediction system but also works as a recommendation model. Based upon the client it will recommend the smartphone he or she is most likely to buy. It makes the user experience much better.

## 2.5 Challenges

There aren't many difficult challenges in the process, but if we were to make a list it would be as follows:

**Lack of previous data:** as of right now we do not possess as much previous data as we want. Let's be clear. We have enough data. But more data means more information for the system to go through which will result in even more accurate results.

**Awareness:** Most of the new comer e-com owners are not aware of the data mining aspect. Getting them on board might be a difficult task.

**Real time testing:** we would need to apply the system on an actual selling website with actual customers to see if the system is ready. But to do that we would need access and admin power of that said website. Most people would not be up for it.

# CHAPTER 3

## REQUIREMENTS ANALYSIS FOR THE PROPOSED SYSTEM

### 3.1 Introduction

approaches taken to work out specific options, demands, expectation by communication with users is requirement. It needs combination of models, classifiers.

### 3.2 System Architecture

All the prices and smartphone specifications will be stored on our database. Website will read from it. Our model will have the access to write on our database. So that the update can be made in real time.



Fig-3.2: System Process



### 3.3 Requirement Collection

We are keeping it as simple as possible. First, we would need data of different smartphones of various range and kinds. These are our data sets. We would divide them into 6 kinds of groups. Feed that information through numerous kinds of classifiers to determine the results. Compare the results to determine which classifier is giving the best results. Use that classifier to determine which smartphones would sell more, which will be best to keep in stock and which won't.

### 3.3 Data Sets

Our data sets were mainly components of smartphones, that determine how good or bad a smartphone is. People identify a smartphone as good or bad depending upon these factors. They are: Processor, Ram, Rom, Primary camera, Secondary camera, Design and Display. These six things were divided into various groups to judge them. Such as:

Table-3.1: Attribute List

Attribute	Description	Possible Values
Processor	GPU power	1 to 5
Ram	Speed	1 to 5
Primary Camera	Back shooter	1 to 5
Secondary Camera	Selfie/front facing	1 to 5
Design	Up to date outlook	1 to 5
Display	Panel and DPI	1 to 5

**Processor**

Very low – below Snapdragon 400 series – 1

Low – Snapdragon 400 series – 2

Medium – Snapdragon 600 series – 3

High – snapdragon 700 series – 4

Very high – snapdragon 800 series – 5

**Ram**

Very low – 1 GB – 1

Low – 2 GB – 2

Medium – 3 to 4 GB – 3

High – 6 GB – 4

Very High – 8 GB – 5

**Primary Camera**

Very low – 1/2 MP – 1

Low – 8 MP – 2

Medium – 15MP – 3

High – 25MP – 4

Very High – 30MP – 5

**Secondary Camera**

Very low – 1/2 MP – 1

Low – 8 MP – 2

Medium – 15MP – 3

High – 25MP – 4

Very High – 30MP – 5

## Display

Very low – LCD, below 720p, ppi count below 300 – 1

Low – ppi count 300 to 350 – 2

Medium – ppi count above 400 – 3

High – IPS LCD 2K ppi count above 450 – 4

Very high – IPS LCD 4K ppi count above 500 – 5

	Name	ProcessUnit	Ram	Primary Camera	Secondray Camera	Design	Display	price	Average	UserDemand	Prediction
0	Samsung galaxy j6+	3	1	3	2	3	2	16999	2.333333	1	0
1	Samsung Galaxy A7	4	2	4	5	4	3	24990	3.666667	5	2
2	Samsung Galaxy Note 9	5	5	5	4	5	5	94900	4.833333	2	1
3	Samsung Galaxy J2 Core	4	1	2	1	2	2	8290	2.000000	1	0
4	Samsung Galaxy J8	1	3	3	3	2	2	21990	2.333333	1	0
5	Samsung Galaxy A6+	1	3	3	4	3	4	26900	3.000000	2	1
6	Samsung Galaxy J4	2	2	2	1	2	2	11990	1.833333	1	0
7	Samsung Galaxy J7 prime	4	3	2	2	3	2	20900	2.666667	1	0
8	Samsung Galaxy S9+	5	4	4	4	4	5	70900	4.333333	2	1
9	Samsung Galaxy J7 Pro	4	3	3	4	4	3	29490	3.500000	1	0
10	Xiaomi Mi8 Lite	3	4	4	4	4	4	23599	3.833333	5	2

Fig–3.3.1: data set

We gathered near about 500+ real public input as our data set which was created based upon their ratings and comments.

## 3.4 Data Mining Tools

- Python
- Sci-kit Learn
- Numpy
- Pandas
- Firebase
- Google Cloud

### 3.4.1 Python

Python is an interpreter, high-level programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. clear programming on both large and small scales is possible with the help of python as it provides forge.

### 3.4.2 Sci-kit Learn

Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN,

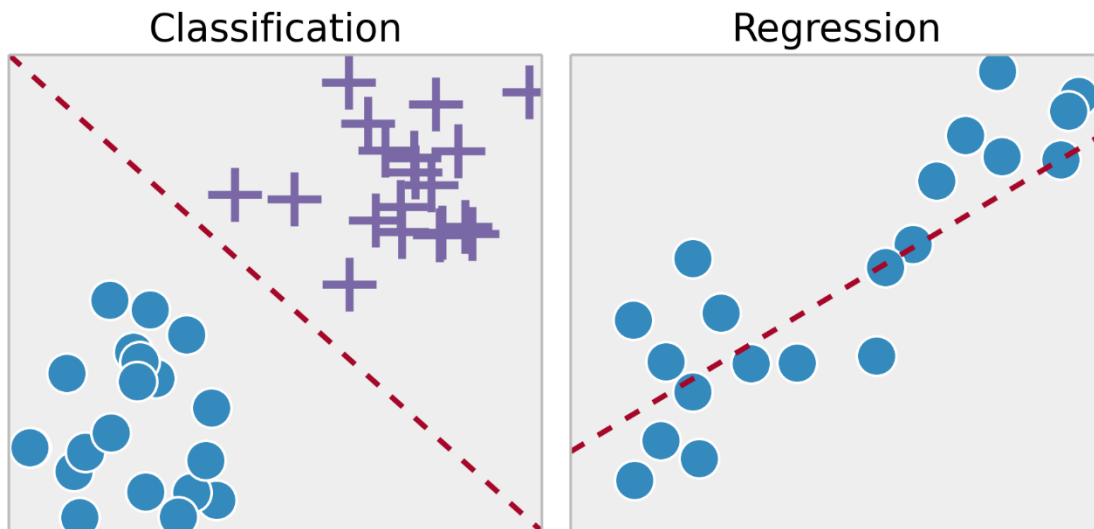


Fig-3.4.2.1: Sci-kit for classification and regression

and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

### 3.4.3 NumPy

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.

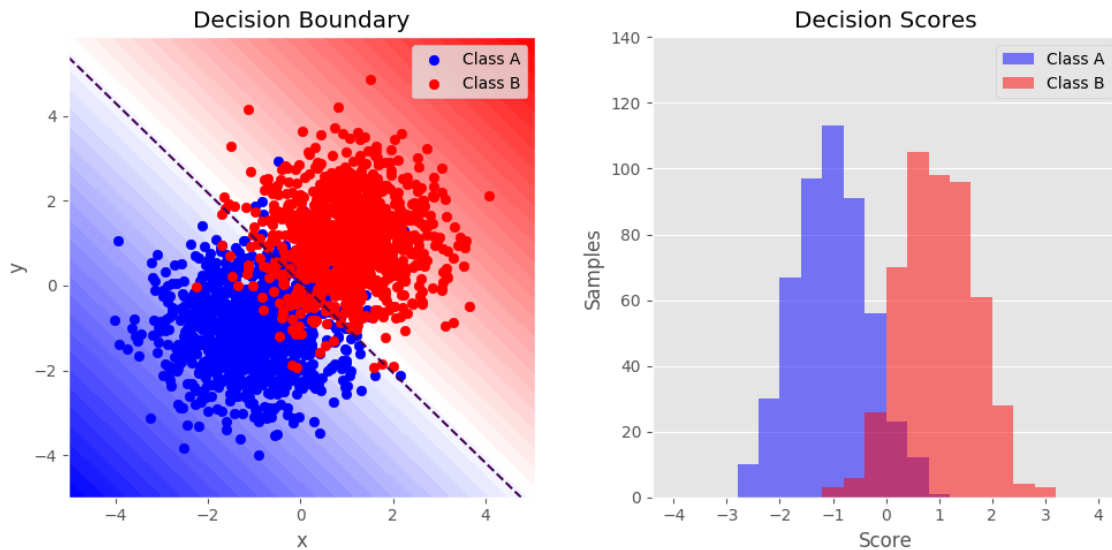


Fig-3.4.3.1: Binary classification with NumPy

The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

### 3.4.4 Pandas

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

## 3.5 Data mining classifiers

### KNeighbors

KNN has some nice properties: it is automatically non-linear, it can detect linear or non-linear

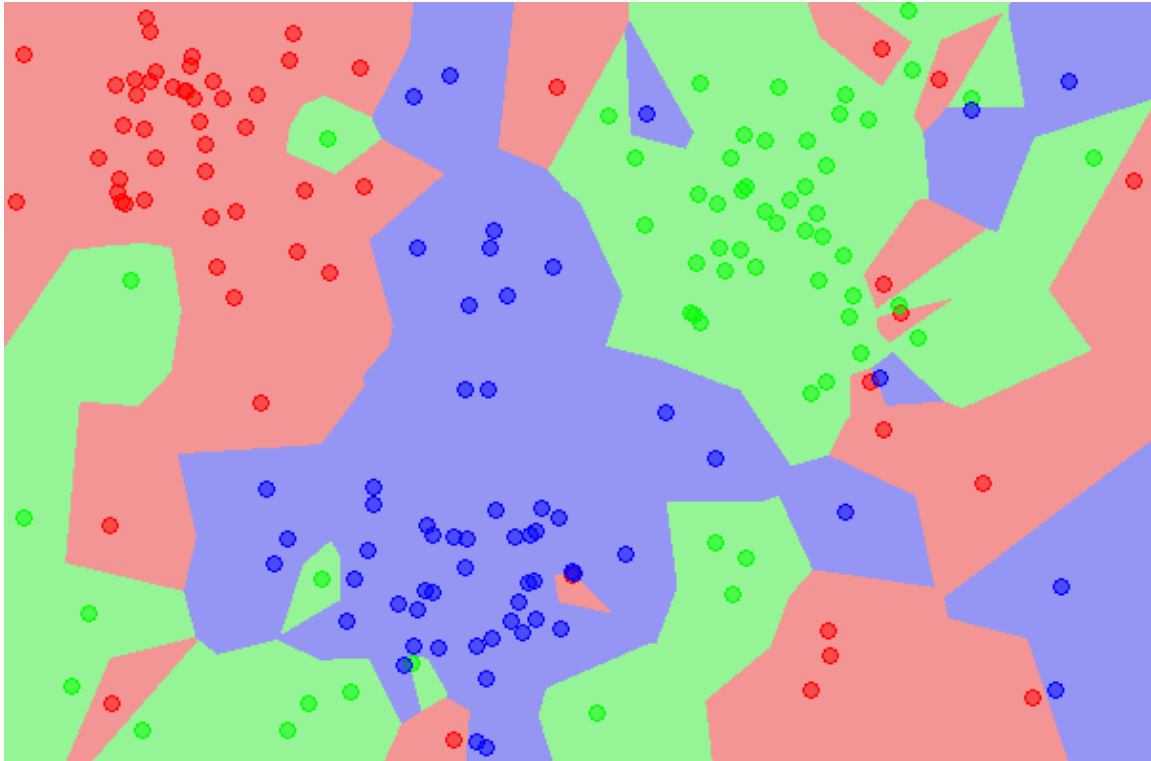


Fig-3.5.1: K-nearest Neighbors Algorithm

distributed data, and it tends to perform very well with a lot of data points. On the minus side KNN needs to be carefully tuned, the choice of  $K$  and the metric (distance) to be used are critical. As Michal Illich mentioned for many data points KNN has performance problems.

If we are in a very low dimensional space we can use a RP-Tree or KD-Tree to improve performance, if we have a higher number of dimensions then we need an approximation to the nearest neighbor problems and whenever we use an approximation we have to think if KNN with the NN approximation is still better than other algorithms. KNN is also very sensitive to bad features (attributes) so feature selection is also important. KNN is also sensitive to outliers and removing them before using KNN tends to improve results.

## **Pros**

1. It is extremely easy to implement
2. As said earlier, it is lazy learning algorithm and therefore requires no training prior to making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g SVM, linear regression, etc.
3. Since the algorithm requires no training before making predictions, new data can be added seamlessly.
4. There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)

## **Cons**

1. The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate distance in each dimension.
2. The KNN algorithm has a high prediction cost for large datasets. This is because in large datasets the cost of calculating distance between new point and each existing point becomes higher.
3. Finally, the KNN algorithm doesn't work well with categorical features since it is difficult to find the distance between dimensions with categorical features.

## **Support Vector Machine**

SVM can be used in linear or non-linear ways with the use of a Kernel, when we have a limited set of points in many dimensions SVM tends to be very good because it should be able to find the linear separation that should exist. SVM is good with outliers as it will only use the most relevant points to find a linear separation (support vectors).

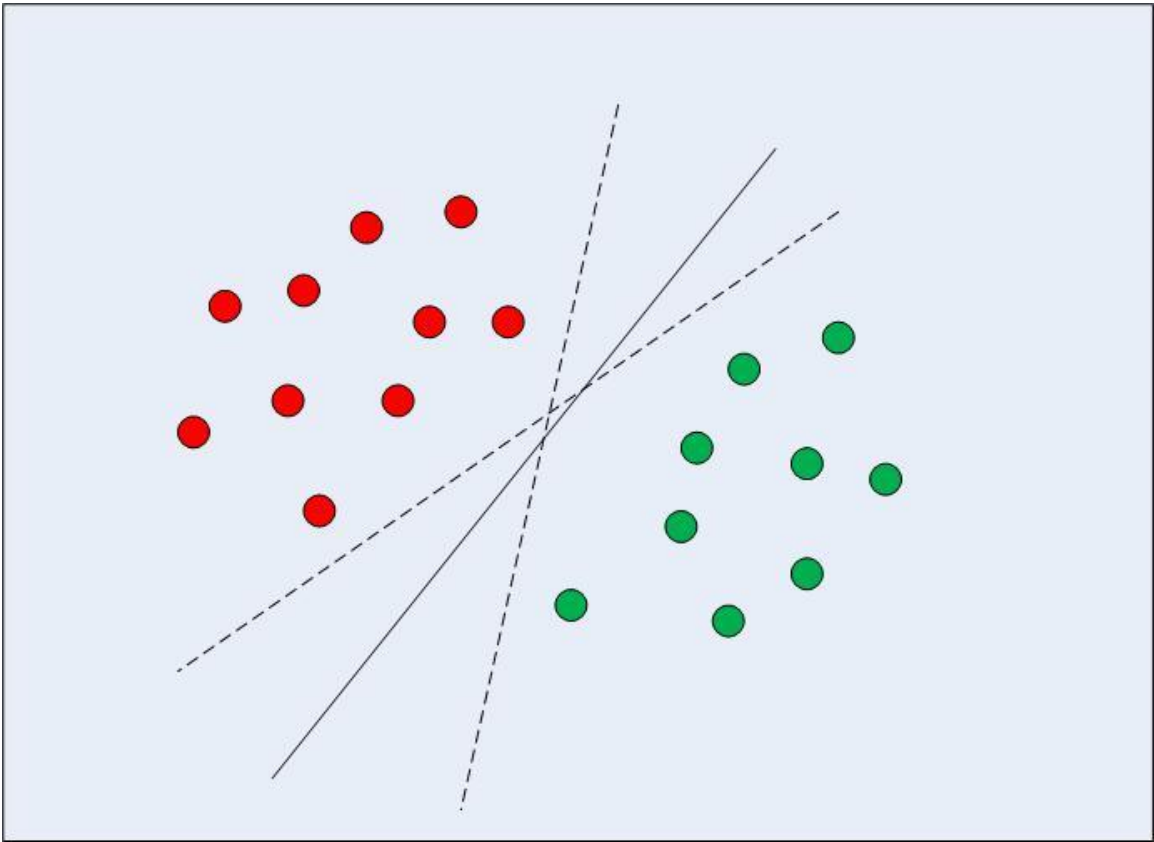


Fig-3.5.2: Multiple Decision Boundaries



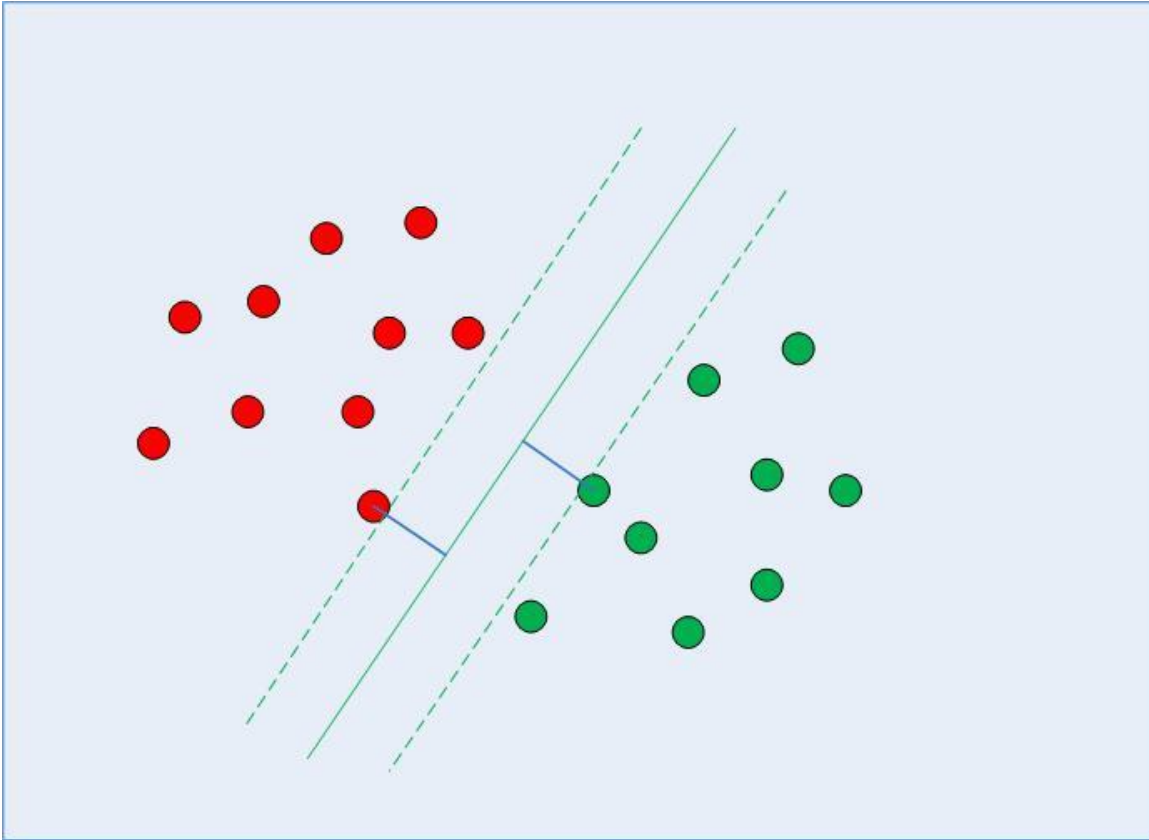


Fig-3.5.3: Decision Boundary with Support Vectors

SVM needs to be tuned, the cost "C" and the use of a kernel and its parameters are critical hyper-parameters to the algorithm.

### **Random Forest**

Random forests or random decision forests as a whole are learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

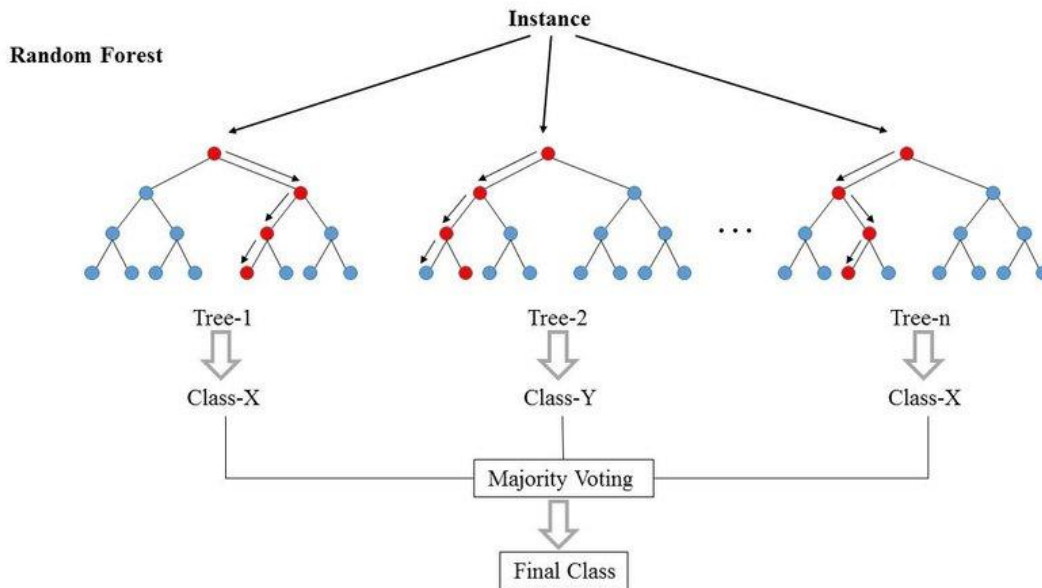


Fig-3.5.4: Decision Tree Simplified

## How the Random Forest Algorithm Works

The following are the basic steps involved in performing the random forest algorithm:

1. N number of random records are picked from the dataset.
2. Decision tree is built based on that N number of records.
3. Number of trees are decided and step 1-2 is repeated.
4. In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

## **Advantages of using Random Forest**

As with any algorithm, there are advantages and disadvantages to using it. In the next two sections we'll take a look at the pros and cons of using random forest for classification and regression.

1. Bias is not one of the qualities of RF algorithm, since, there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore, the overall biasedness of the algorithm is reduced.
2. This algorithm is very stable. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees.
3. The random forest algorithm works well when we have both categorical and numerical features.
4. The random forest algorithm also works well when data has missing values or it has not been scaled well (although we have performed feature scaling in this article just for the purpose of demonstration).

## **Disadvantages of using Random Forest**

1. Complexity is random forests big bad side. They join large number of decision trees together so. So massive computational resources are required.
2. They train longer.

## CHAPTER 4

### IMPLEMENTATION AND TESTING

#### 4.1 Implementation

We used python sci-kit library to perform data mining operations. There were also some other libraries and tools used along the way such as pandas and numpy. To understand our procedure easily we can divide our full operation into 8 easy steps. They are as follows:

Step 1: first we import pandas, different classifiers of sci-kit learn machine learning library such as SVC, KNeighbours, and Random forest.

Step 2: then we read the data set using pandas and also fill up any null data.

Step 3: we calculate the average of their quantity and query price & average quantity to figure out customer demand.

Step 4: using label enabler we transform null attributes to natural numbers.

Step 5: we label prediction attribute as 'y' and all the other attributes as 'x' from the total data set.

Step 6: we divide 'x' into test and train data using train\_test\_split sci-kit learn function.

Step 7: using different classifier and fit method we build a model with the train data to predict our needed result.

Step 8: then using 'y' test data compare predict data to create accuracy score.

$$average = \frac{\text{score given to all attribute}}{6}$$

$$PQ\ Rate = \frac{\text{price}}{\text{average}}$$

$$prediction = PQ\ Rate \times DemandUse$$

Following table shows the user demand calculation method

User Demand	Quality	Price
0% - 5%	2.5 – 2.8	null
5% - 10%	2.8 – 3.0	null
10% - 20%	3.0 – 3.5	20,000 – 22,000
10% - 20%	4.0 – 4.5	33,000 – 40,000
20% - 25%	4.8	40,000 – 55,000
25% - 30%	3.5 – 4.0	25,000 – 30,000
25% - 30%	4.8 – 5.0	50,000 – 80,000
30% - 40%	4.0 – 4.2	25,000 – 35,000
30% - 40%	3.5 – 4.0	25,000 – 27,000
30% - 40%	4.3 – 4.5	25,000 – 30,000
40% - 50%	4.5 – 4.7	35,000 – 40,000
50% - 60%	3.0 – 3.5	15,000 – 20,000
50% - 60%	3.5 – 4.0	20,000 – 25,000
60% - 70%	4.0 – 4.2	20,000 – 25,000
60% - 70%	4.3 – 4.8	28,000 – 33,000
60% - 70%	3.0 – 3.5	10,000 – 15,000

## 4.2 Experimental Result

After following our steps and running all the classifiers we got accuracy rates. First we inserted our original 100 data inputs. Which were raw and public generated. It gave us that SVC was the best classifier with highest accuracy.

To improve the accuracy even more and to find out if the accuracy rate holds up even against a higher number of data inputs we used python function ‘range’ to duplicate data sets up to 2000 instances.

What we did was – using the range function we defined a range such as 2.5 to be the threshold and calculate all possible outcomes or combinations from 15000 to 100000. Upon doing this we got the same result.

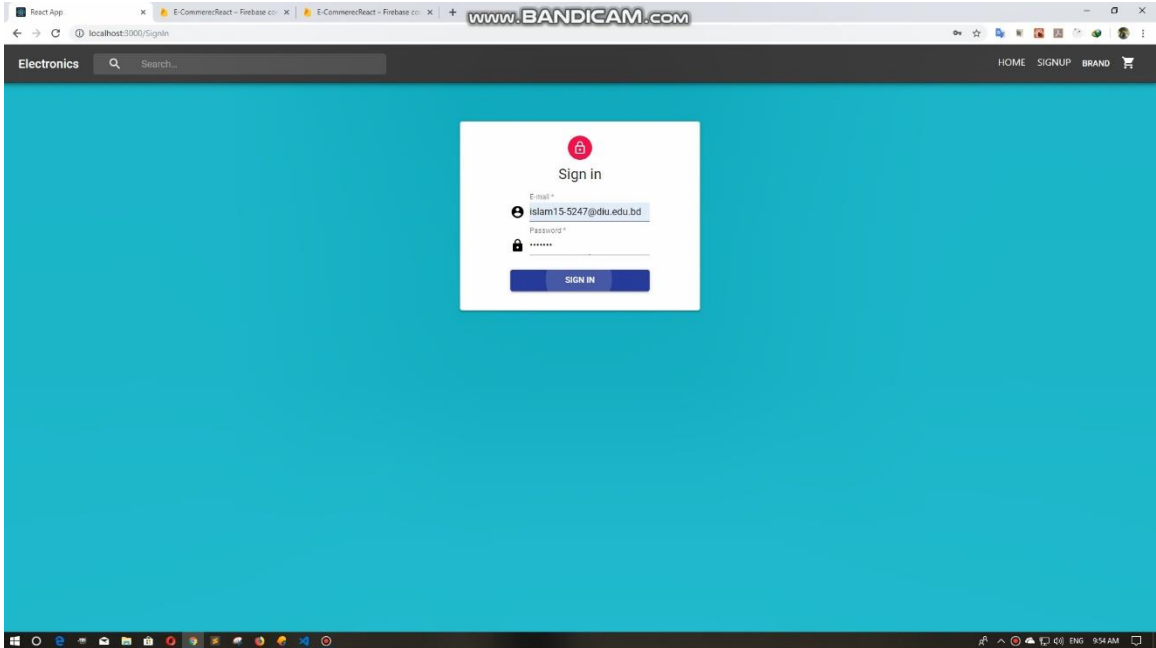


Fig-4.2.1: website signin page

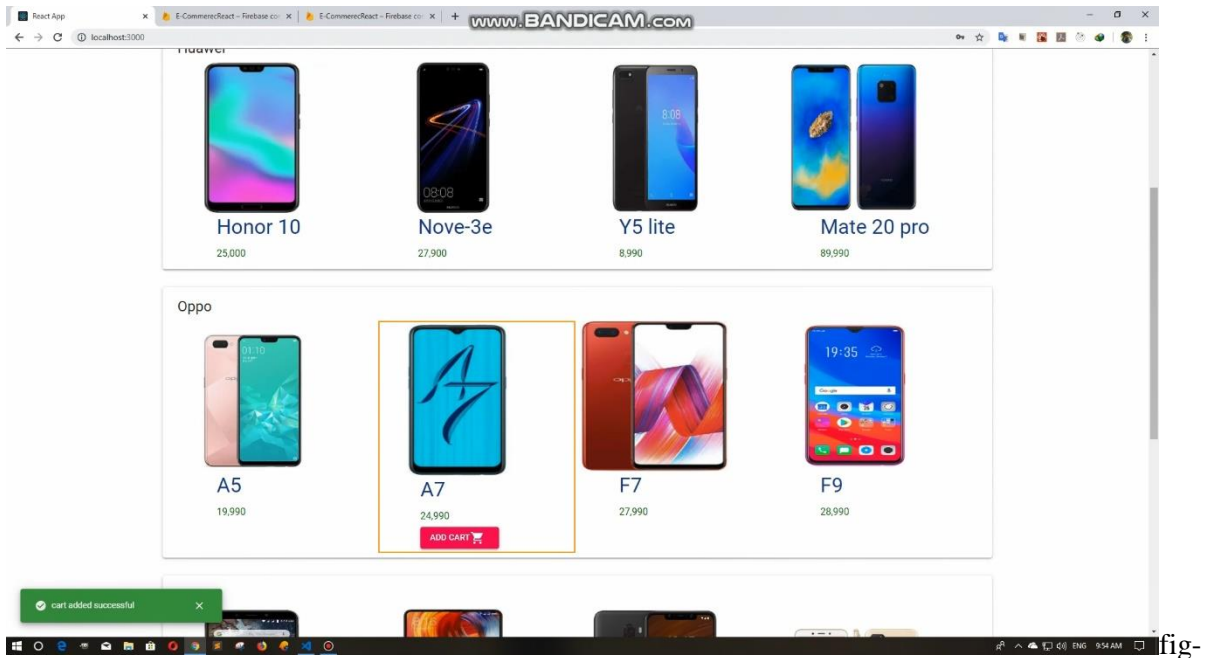


Fig-4.2.2: website homepage

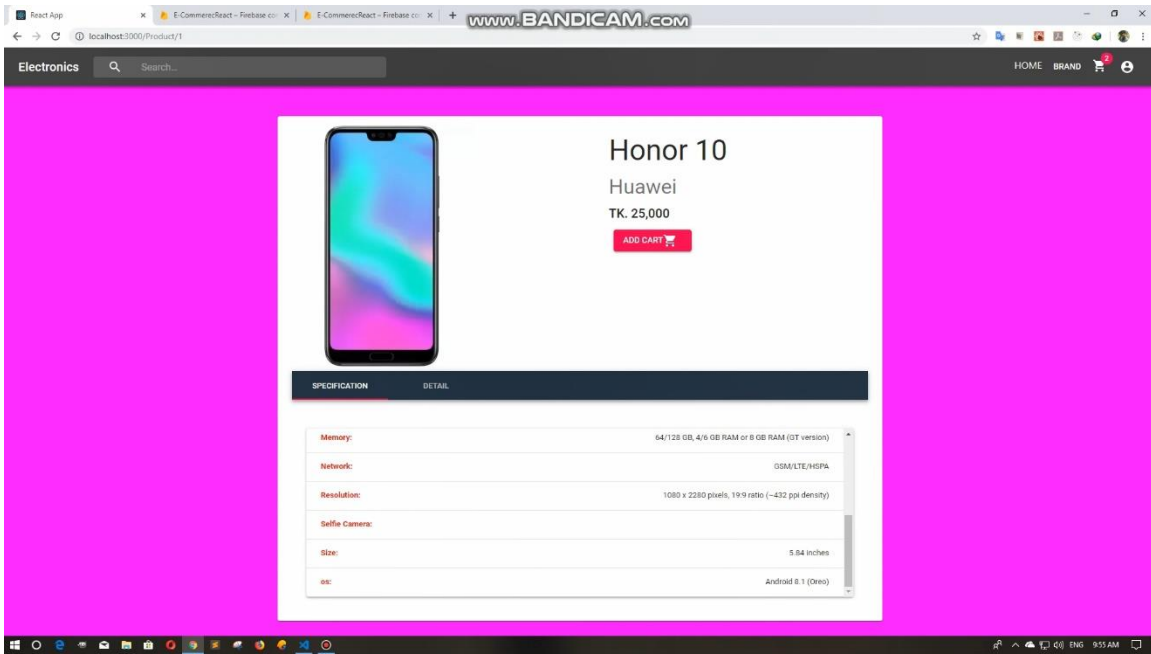


Fig-4.2.3: specification page

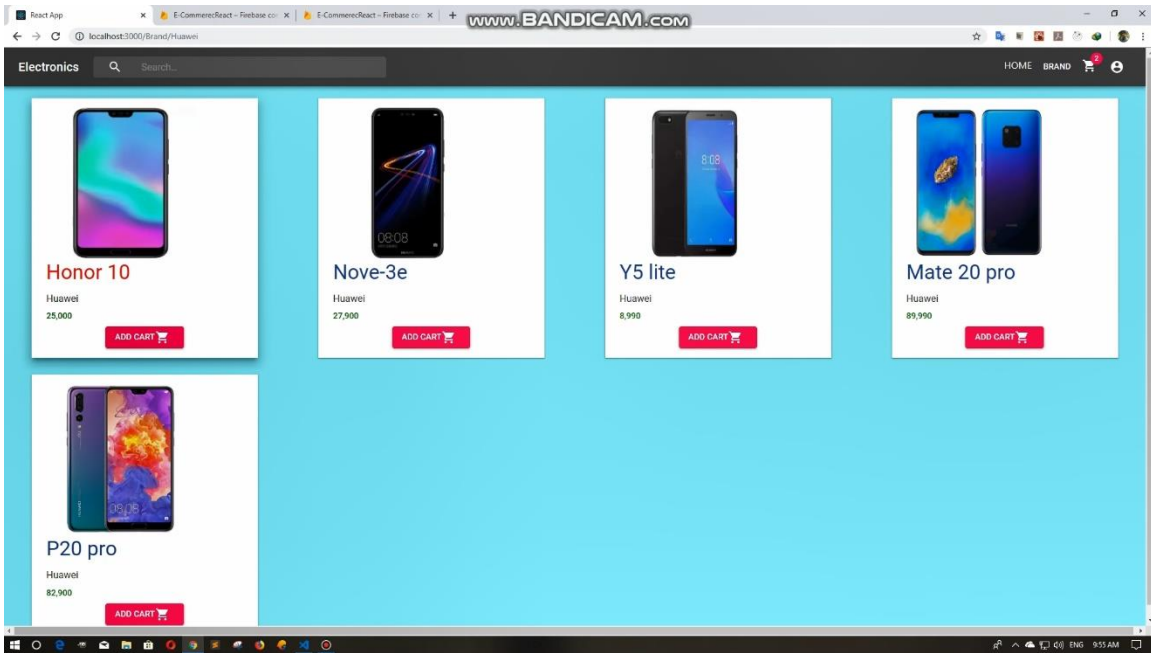


Fig-4.2.4: bestsellers

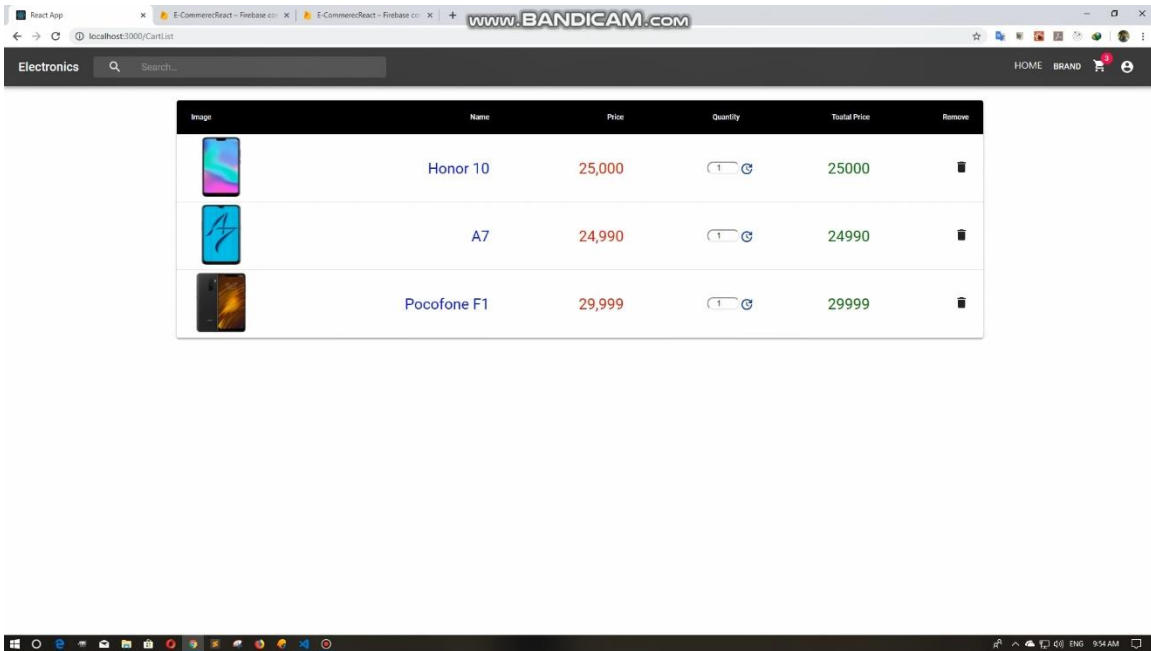


Fig-4.2.5: Website Cart

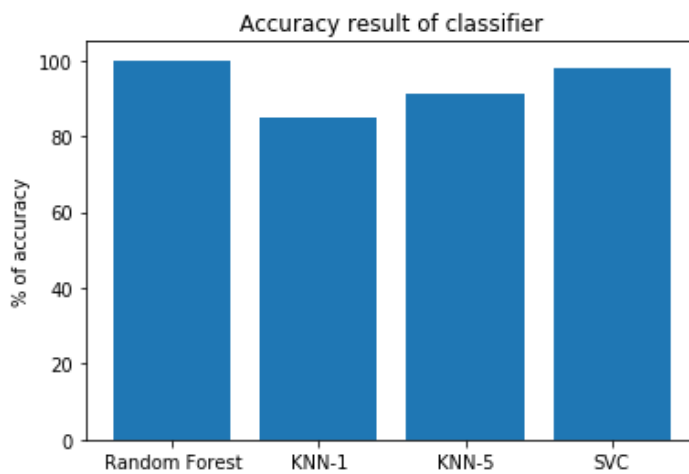


Fig-4.2.6: accuracy rates for original data

From the figure we can see that random forest has accuracy of 100%, SVC 94%, KNeighbors 82% when n is 1 and 90 when n is 5.



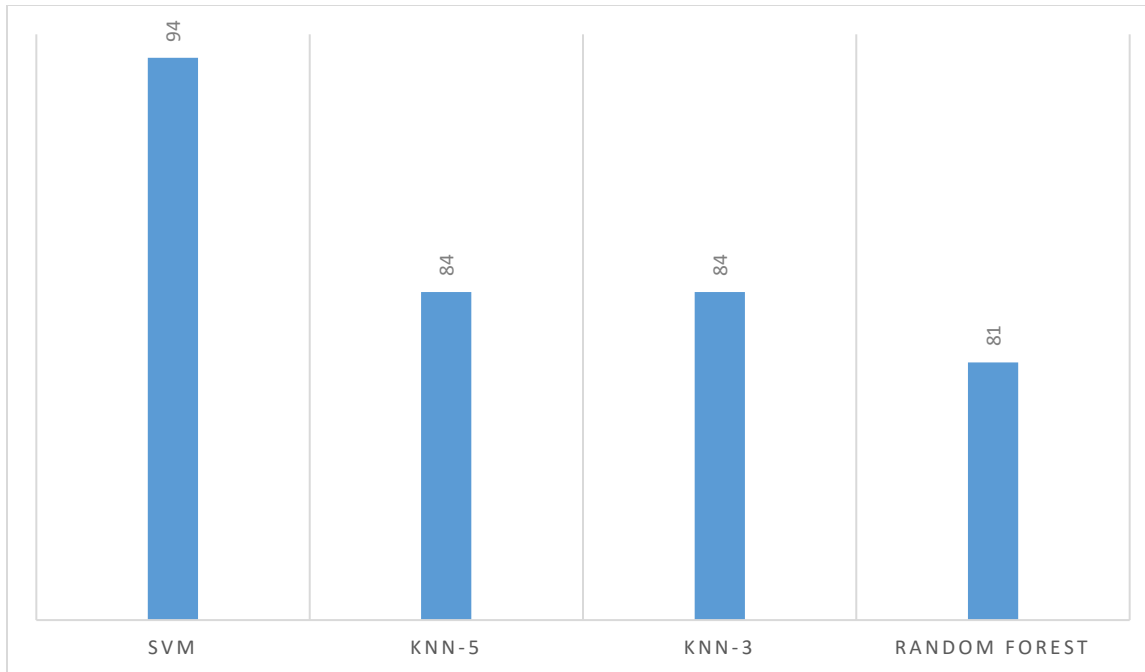


Fig-4.2.7: accuracy with all possible data

As we can see here, SVM has the accuracy of 94%, KNeighbors 84% when n is 5 and 91% when n is 3.

### 4.3 Comparison

The basic steps to decide which algorithm to use will depend on a number of factors. Few factors which one can look for are listed below:

- Training set numbers.
- Space dimensions.
- Do we have corresponding interactions?
- Is over fitting a problem?

These are just few factors on which the selection of algorithm may depend. Once we have the answers for all these questions, we can move ahead to decide the algorithm.

## **SVM**

The main reason to use an SVM instead is because the problem might not be linearly separable. In that case, we will have to use an SVM with a nonlinear kernel (e.g. RBF). Another related reason to use SVMs is if we are in a highly dimensional space. For example, SVMs have been reported to work better for text classification. But it requires a lot of time for training. So, it is not recommended when we have a large number of training examples.

## **KNeighbors**

It is robust to noisy training data and is effective in case of large number of training examples. But for this algorithm, we have to determine the value of parameter K (number of nearest neighbors) and the type of distance to be used. The computation time is also very much as we need to compute distance of each query instance to all training samples.

## **Random Forest**

Random Forest is nothing more than a bunch of Decision Trees combined. They can handle categorical features very well. This algorithm can handle high dimensional spaces as well as large number of training examples. Random Forests can almost work out of the box and that is one reason why they are very popular.

## **CHAPTER 5**

### **CONCLUSION**

#### **5.1 Discussion**

In this study, we proposed and implemented a prediction system for e-commerce websites. This system will ensure profit and save time, money & effort. It will also provide an easy pick up system by providing notification. After all, we may hope for a better solution of e-commerce system. If e-commerce sites use this method to stock their smartphones, they business will become more profitable. They will be able to predict accurately which product is for which demographic and how much money will they be able to earn on each smartphone.

#### **5.2 Limitation**

In this study we have only worked with smartphones. But there are many other products that need to be included. For the lack of time and data we could not do so. Our system has not been trained for any other products. It can only predict smartphone profits. We need to train our system for all the sellable products in existence.

#### **5.3 Future Work**

In this study we have tried to work with smartphones. In future all consumer products can be included. As for research, one can try to find ways how to incorporate product suggestions depending on present cart for better results.



## REFERENCES

1. Zhao, Liang, Nai-Jing Hu, and Shou-Zhi Zhang. "Algorithm design for personalization recommendation systems." *Journal of computer research and development* 39.8 (2002): 986-991.
2. Siddiqui, Ahmad Tasnim, and Sultan Aljahdali. "Web mining techniques in e-commerce applications." *arXiv preprint arXiv:1311.7388* (2013).
3. Al-Radaideh, Qasem A., Adel Abu Assaf, and Eman Alnagi. "Predicting stock prices using data mining techniques." *The International Arab Conference on Information Technology (ACIT'2013)*. 2013.
4. Hongjiu, Gu. "Data mining in the application of e-commerce website." *Intelligence Computation and Evolutionary Computation*. Springer, Berlin, Heidelberg, 2013. 493-497.
5. Analytics Vidhya. Which one to use – RandomForest vs SVM vs KNN?  
Available at : <https://discuss.analyticsvidhya.com/t/which-one-to-use-randomforest-vs-svm-vs-knn/2897/3> [last access date: 24 February, 2019, 8:20PM]
6. Wikipedia. Statistical classification.  
Available at : [https://en.wikipedia.org/wiki/Statistical\\_classification](https://en.wikipedia.org/wiki/Statistical_classification) [last access date : 24 february, 2019, 7:41PM]
7. Quora. What is better, k-nearest neighbors' algorithm (k-NN) or Support Vector Machine (SVM) classifier? Which algorithm is mostly used practically? Which algorithm guarantees reliable detection in unpredictable situations?  
Available at : <https://www.quora.com/What-is-better-k-nearest-neighbors-algorithm-k-NN-or-Support-Vector-Machine-SVM-classifier-Which-algorithm-is-mostly-used-practically-Which-algorithm-guarantees-reliable-detection-in-unpredictable-situations> [last access date : 2 march, 2019, 3:27PM]
8. Stoica, Eduard Alexandru, and Esra Kahya Özyirmidokuz. "Mining customer feedback documents." *International Journal of Knowledge Engineering* 1.1 (2015): 68-71.
9. Gamon, Michael, et al. "Pulse: Mining customer opinions from free text." *international symposium on intelligent data analysis*. Springer, Berlin, Heidelberg, 2005.
10. ORALHAN, Burcu, U. Y. A. R. Kumru, and Zeki ORALHAN. "Customer satisfaction using data mining approach." *International Journal of Intelligent Systems and Applications in Engineering* (2016): 63-66.
11. Osmanbegovic, Edin, and Mirza Suljic. "Data mining approach for predicting student performance." *Economic Review: Journal of Economics and Business* 10.1 (2012): 3-12.
12. Huang, Wenjie, et al. "A novel trigger model for sales prediction with data mining techniques." *Data Science Journal* 14 (2015).

## APPENDIX

### APPENDIX A: CODES

#### **Importing necessary files:**

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from matplotlib import pyplot as plt
df = pd.read_csv('MobileData.csv', usecols=['Name', 'ProcessUnit', 'Ram', 'Primary
Camera', 'Secondary Camera', 'Design', 'Display', 'price'])
```

#### **Printing accuracy**

```
clf = SVC()
clf.fit(X_train, y_train)
pred_clf = clf.predict(X_test)
print(accuracy_score(y_test, pred_clf)*100, '%')
```

# Turnitin Originality Report

Processed on: 17-Apr-2019 10:10 +06  
ID: 1114075326  
Word Count: 4366  
Submitted: 1

Similarity Index <b>29%</b>	<b>Similarity by Source</b> Internet Sources: N/A Publications: N/A Student Papers: 29%
--------------------------------	--

profit prediction using data mining algorithm By Sayed Mahmud