Pattern Recognition on Given Dataset

Sarwar Saif [1*] **and Samiur Rahman**[1+]

[1]AUST, Computer Science and Engineering, Dhaka, Bangladesh
[*]15.01.04.091
[+]15.01.04.092
[1]these authors contributed equally to this work

ABSTRACT

Depending on human choices and approach to some specific things we can classify their traits. This report discusses mainly about two aspects related to human life. That is whether a person will choose an expensive or fancy restaurant over cheap ones just for social status and the relationship of his career choice with CGPA and current state of mind etc.

1 Problem

1.1 What is the Problem?

In our project, we'll talk about two problems.

1.1.1 Problem 1

Most of the students don't have a proper career goal and their future. Most of them choose the wrong profession. They choose the wrong career path and most of them are shy about to tell what they want to be.
We want to provide each student with a proper career plan and what they need to do reach their goal.

1.1.2 Problem 2

In this day and age the influence of social media has become quite prominent in most sectors of our daily life. Our problem discusses about this tendency of selection of expensive restaurants/coffee shops over cheap ones only for social status.

2 Motivation

2.1 Problem 1

Most of the students choose the wrong profession. To make a better future they need a proper guideline. By solving this problem we can make a better generation who knows how to reach their goal.

2.2 Problem 2

The main motivation behind this problem is to show how this young generation is getting influenced even in the smallest things like selection of a restaurant to eat or hangout which is very alarming.This indicates us that social media might be a source of bad influence for this and the future generations.

2.3 Feasibility

Our solution predicts results with more than 70 percent accuracy. If we collect enough data our method can predict more optimal results.Thus we can provide the users with enough information.

3 Background

3.1 Supervised techniques

Supervised learning is the Data mining task of inferring a function from labeled training data.The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a “reasonable” way.

A. Support Vector Machines(SVM)

Here we plot the data as points in n-dimensional space where n is the number of features with the value of each feature being the value of a particular coordinate. These values are called support vectors. Classification of the tracts is performed by finding the hyper-plane that differentiates the two classes. We take the hyper-plane that leaves the maximum margin from both classes. This is how our prediction is done.

Advantage: SVM is good as compared to other classifiers as the computational complexity is reduced classification efficiency is increased when compared to any other non linear classifier

Disadvantage: Slow training process for large data sets. Hard to find features.

B. K-Nearest Neighbour(K-NN)

An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

Here a streamline is compared with its nearest streamline and they are declared same class upon similarities of their properties. Choosing “K” factor is an important task because the output is largely influenced by it.

Advantage:
Easy to implement.

Disadvantage:
Suffer due to impact of noisy examples and needs very large datasets.

3.2 Tools

We used python as our language for coding and some of its libraries like Numpy, Pandas, Matplotlib, sklearn. We used Anaconda for creating our environment to work in.

4 Solution

4.1 Pseudo Code

```
[raster columns=1,raster equal height] [nobeforeafter, title=K-Nearest Neighbors] Df ← $LoadDataset.csv$
$Fd \leftarrow NormalizedDataset$
$TrainSet \leftarrow X : FD(i), i = 0,1,2,3; Y : FD(i), i = 4$
$TestSet \leftarrow X' : FD(i) - X, i = 0,1,2,3; Y' : FD(i) - Y, i = 4$
$clf \leftarrow KNeighborsClassifier(neighbors = 4)$
$clf.fit(xTrain, yTrain)$
$yPred \leftarrow clf.predict(xTest)$
$Acc \leftarrow accuracy_score(yTest, yPred)$

    [nobeforeafter, title=Support Vector Machine] Df ← $LoadDataset.csv$
$Fd \leftarrow NormalizedDataset$
$TrainSet \leftarrow X : FD(i), i = 0,1,2,3; Y : FD(i), i = 4$
$TestSet \leftarrow X' : FD(i) - X, i = 0,1,2,3; Y' : FD(i) - Y, i = 4$
$clf \leftarrow svm.SVC(kernel = `kernel')$
$clf.fit(xTrain, yTrain)$
$yPred \leftarrow clf.predict(xTest)$
$Acc \leftarrow accuracy_score(yTest, yPred)$
```

4.2 Flowchart

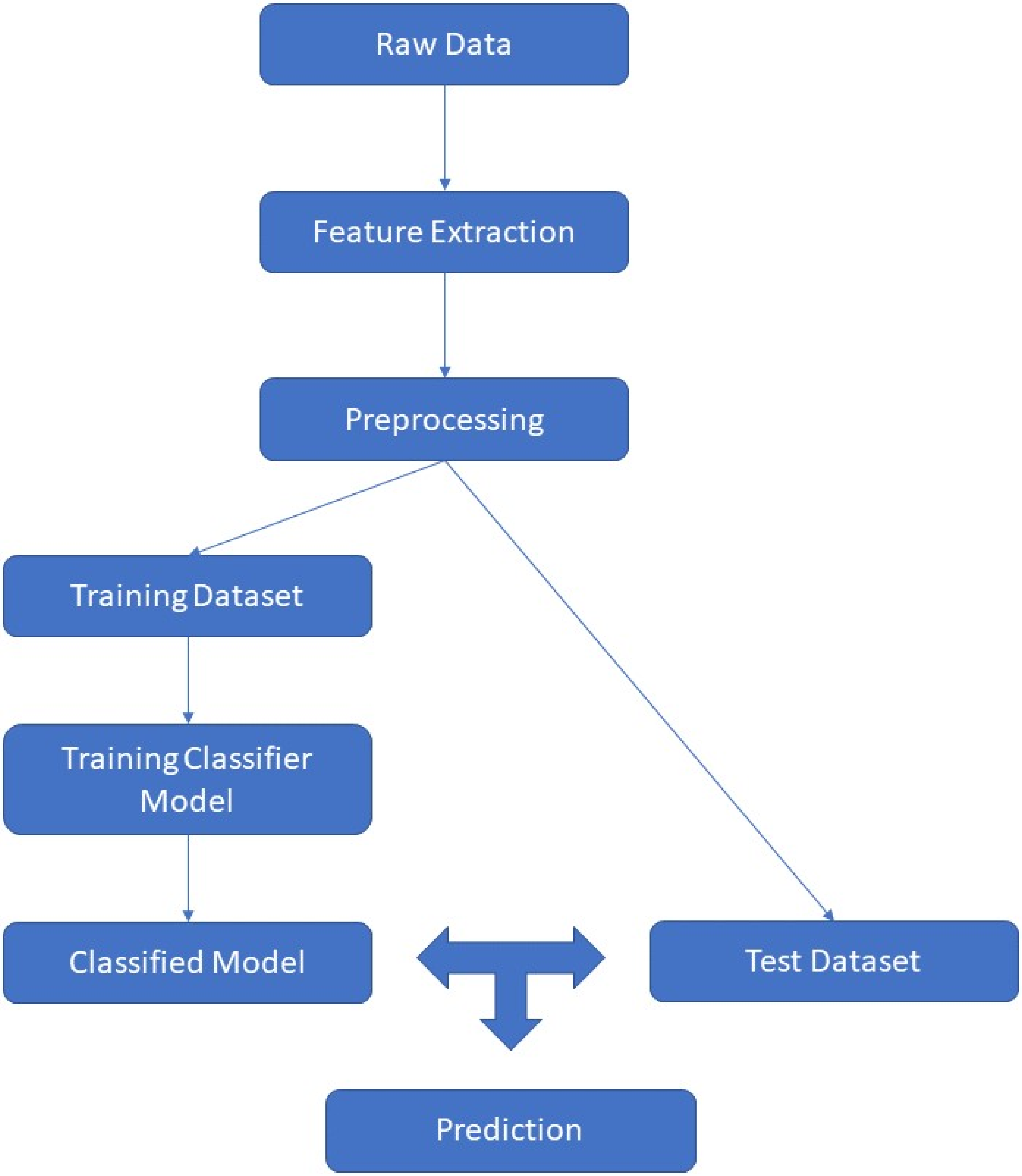

Figure 1. Flowchart.

5 Evaluation

5.1 Experiment Setup

We used data of 95 different students for prediction.

5.2 Experimental results

Problem	**KNN Classifier**	**SVM Classifier (kernel=linear)**	**SVM Classifier (kernel=polynomial)**	**SVM Classifier (kernel=rbf)**
Problem 1	70.213%	59.574%	53.191%	70.213
Problem 2	91.176%	91.176%	94.118%	91.176%

Table 1. Experimental result analysis for applied techniques

5.3 Visualization of Experimental results

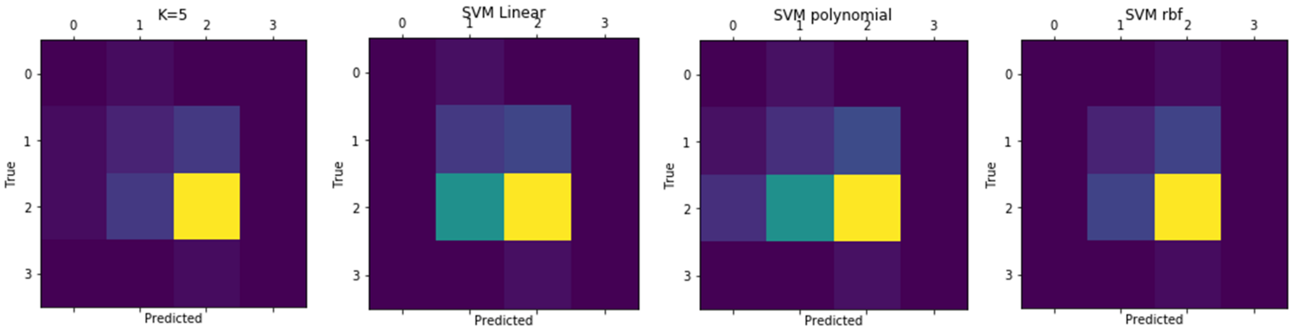

Figure 2. Confusion matrix for the implemented techniques in problem 1.

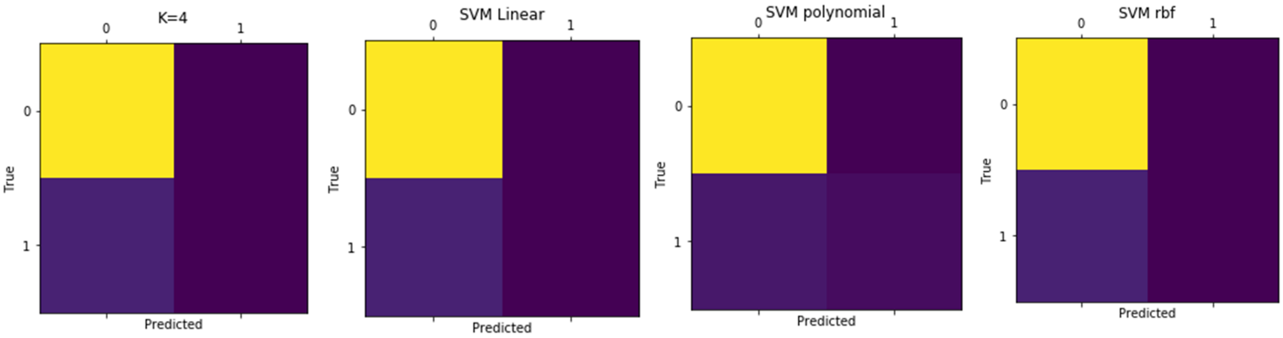

Figure 3. Confusion matrix for the implemented techniques in problem 2.

6 Discussion

The obtained results are pretty much what we expected. In the case of accuracy, KNN performs well for both problems whereas SVM varies from kernel to kernel. The raw dataset was poor for the classification. Some students didn't even bother to fill up the form correctly. We have to normalize the dataset before using it. For a classification problem the more the data, the better the result. If we are able to collect enough dataset, our methods will work much better and will produce more optimal results. The main limitations of our solutions are the lack of data.

7 Conclusion

The sector of pattern recognition is getting popular in our country day by day due to the digitization of our country. Many well-known companies are currently exercising the trend enormously by analyzing user data. If we can use our models for any learning management website or cafes, we can add a huge value commercially. We can also collect enough user data from the clients and implement a more robust model.