Diabetes Prediction Based on Health Indicators

Abrar Rafid Noor	**170204059**
Md. Sakib Irtiza	**170204081**
Labib Abdullah	**170204114**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Spring 2021

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

Diabetes Prediction Based on Health Indicators

Submitted by

Abrar Rafid Noor	**170204059**
Md. Sakib Irtiza	**170204081**
Labib Abdullah	**170204114**

Submitted To

Faisal Muhammad Shah, Associate Professor

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

ABSTRACT

Machine Learning has a substantial influence on a variety of scientific and technological fields, including medical research and life sciences. Diabetes Mellitus, or diabetes, is a chronic condition characterized by unusually high glucose levels in blood cells and the ineffective use of insulin by the human body. Chronically high levels of sugar in the circulation are linked to complications including heart disease, eyesight loss, lower-limb amputation, and kidney illness in diabetics. While there is no cure for diabetes, many individuals can benefit from tactics such as decreasing weight, eating healthy, staying active, and obtaining medical treatment. Predictive models for diabetes risk are crucial tools for public and public health professionals since early diagnosis can lead to lifestyle modifications and more effective treatment. In this project, we have tried to preprocess the Behavioral Risk Factor Surveillance System (BRFSS) dataset on diabetes that focuses on general health indications and later used machine learning models to see which model performs better in terms of classifying the stages of diabetes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Diabetes is a long-term illness that inhibits the body's ability to convert food into energy. The majority of our meals is broken down into sugar (also known as glucose) and released into our bloodstream. When our blood sugar levels rise, our pancreas is prompted to release insulin. Insulin is a key that allows blood sugar to enter your body's cells and be used as energy. When a person develops diabetes, their body either doesn't produce enough insulin or can't process it as well as it should. Too much blood sugar persists in the circulation when there isn't enough insulin or when cells cease reacting to insulin. This can lead to major health issues such as heart disease, visual loss over time, and kidney disease.
The World Health Organisation (WHO) considers blood glucose levels of below 5.5 mmol/l to be normal. Those of 7 mmol/l and above are considered diabetic. Between these two cutoff points lies the prediabetic range: 5.5 to 7 mmol/l. If someone has prediabetes, they can make diet and lifestyle changes and bring blood sugar levels back to the normal range. This dramatically reduces the risk of developing Type 2 diabetes.
That is why in this project, along with classifying no diabetes and diabetes, more emphasis has been given towards predicting prediabetes stage as it can help people to be cautious of their current state and make appropriate changes in their day to day life to prevent it from becoming even serious.

Chapter 2

Literature Reviews

2.1 Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques

From the work [1] of Xie et. al were similar to ours as they also did their experiment with CDC's BRFSS dataset of 2014, but one of the main difference with their approach of ours was that they only predicted type-2 diabetes, while we focued on type-2 as well as type-1 diabetes.
We learnt about body mass index (BMI) ranges and discarded them throughout the pre-processing phase of the experiment. Important characteristics such as BMI, age, and other factors were also discussed in their paper. The best performing models were SVM, Random Forest, and Neural Networks. Neural Networks had the highest accuracy of 0.8241.

2.2 Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh

The work [2] of Pranto et. al were mainly focused on the women patients and were done on PIMA dataset, which had different characteristics than ours. Among their models, Random Forest fared the best. In order to find relevant characteristics, we learned basic preprocessing techniques, such as the use of heatmap. Random Forest had the highest accuracy of 0.779.

Chapter 3

Data Collection & Processing

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC). The study collects data on health-related risk behaviors, chronic health issues, and the utilization of preventative services from over 400,000 Americans each year. Since 1984, it has been held every year. A csv of the dataset is available on Kaggle for the year 2015 was utilized for this project.
Dataset link: https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset

3.0.1 Features & Class Distribution

Excluding the target feature, the dataset has 21 features. All of them are numerical values of floating type. No missing values were seen for any of the entries. Brief idea about the features are given below:
Diabetes_012 is our target feature where 0 stands for no diabetes, 1 for prediabetes and 2 for diabetes. *HighBP*- has high blood pressure or not. *HighChol*- has high cholesterol or not. *CholCheck*- has done cholesterol check in 5 years or not. *BMI*- shows the Body Mass Index. *Smoker*- have smoked at least 100 cigarettes in their entire life or not. *Stroke*- if had a stroke or not. *HeartDiseaseorAttack*- if had Coronary Heart Disease (CHD) or Myocardial Infarction (MI). *PhysActivity*- if had done physical activity in past 30 days - not including job. *Fruits*- if consumes fruit 1 or more times per day. *Veggies*- if consumes Vegetables 1 or more times per. *HvyAlcoholConsump*- If a heavy drinker. *AnyHealthcare*-If has any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. *NoDocbcCost*- Was there a time in the past 12 months when they needed to see a doctor but could not because of cost. *GenHlth*- In general their health is: (scale 1-5) 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor. *MentHlth*- Mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was their mental

health not good. *PhysHlth*- Physical health, which includes physical illness and injury, for how many days during the past 30 days was their physical health not good. *DiffWalk*- If they have serious difficulty walking or climbing stairs. *Sex*- Their sex, male or female. *Age*- Total of 13-level age category. 1 = 18-24, 2 = 25-31 and so on. *Education*- Education level (scale 1-6). 1 = Never attended school or only kindergarten, 2 = Grades 1 through 8 and so on. *Income*- Income(USD) scale.

The class distribution of our dataset is given below:

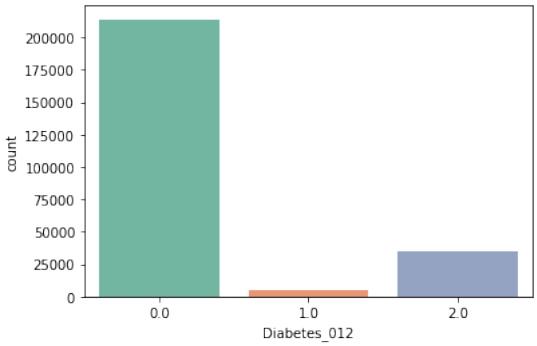


Figure 3.1: Class Distribution of Original Dataset

From the distribution, we can see that it is very much imbalanced. No Diabetes class is *84.24%*, Prediabetes is *1.82%* and Diabetes is *13.93%*.

When Decision Tree was applied, it got very poor results, specially in terms of classifying prediabetes and diabetes. Details of the results are shown in table "omuk".
In the following chapter, we have gone through undersampling for this dataset to obtain a good classification result.

3.1 Under Sampling

For our dataset, we have gone through 2-types of undersampling techniques. One is applying Random Undersampling and the other is Applying NearMiss Algorithm.

3.1.0.1 Applying Random Undersampling

Random undersampling involves randomly selecting examples from the majority class and deleting them from the training dataset. In the random undersampling, the majority class instances are discarded at random until a more balanced distribution is reached.

After applying random undersampling on the dataset, Decision Tree gave a more stable result, although it was not satisfactory. Details of the results are shown in table "omuk".

3.1.0.2 Applying NearMiss

NearMiss is an algorithm that can help in balancing an imbalanced dataset. When two points belonging to different classes are very close to each other in the distribution, this algorithm eliminates the datapoint of the larger class thereby trying to balance the distribution.

There are three versions of the technique, named NearMiss-1, NearMiss-2, and NearMiss-3. NearMiss-1 selects examples from the majority class that have the smallest average distance to the three closest examples from the minority class. NearMiss-2 selects examples from the majority class that have the smallest average distance to the three furthest examples from the minority class. NearMiss-3 involves selecting a given number of majority class examples for each example in the minority class that are closest.
Details of the results after applying NearMiss algorithm is shown in table "omuk".

Undersampling technique NearMiss-1 performed quite better than NearMiss-3, as well as the previous Random Undersampling. So we further proceeded with this undersampled dataset. And the class distribution became equally balanced.

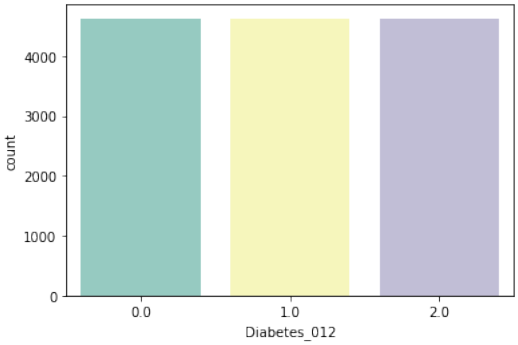


Figure 3.2: Class Distribution after Applying NearMiss-1

3.2 Feature Selection

3.2.1 Visualizing Data Distribution

Visualizing data distribution can give us meaningful insights. We visualized our features to see their distribution in the following.

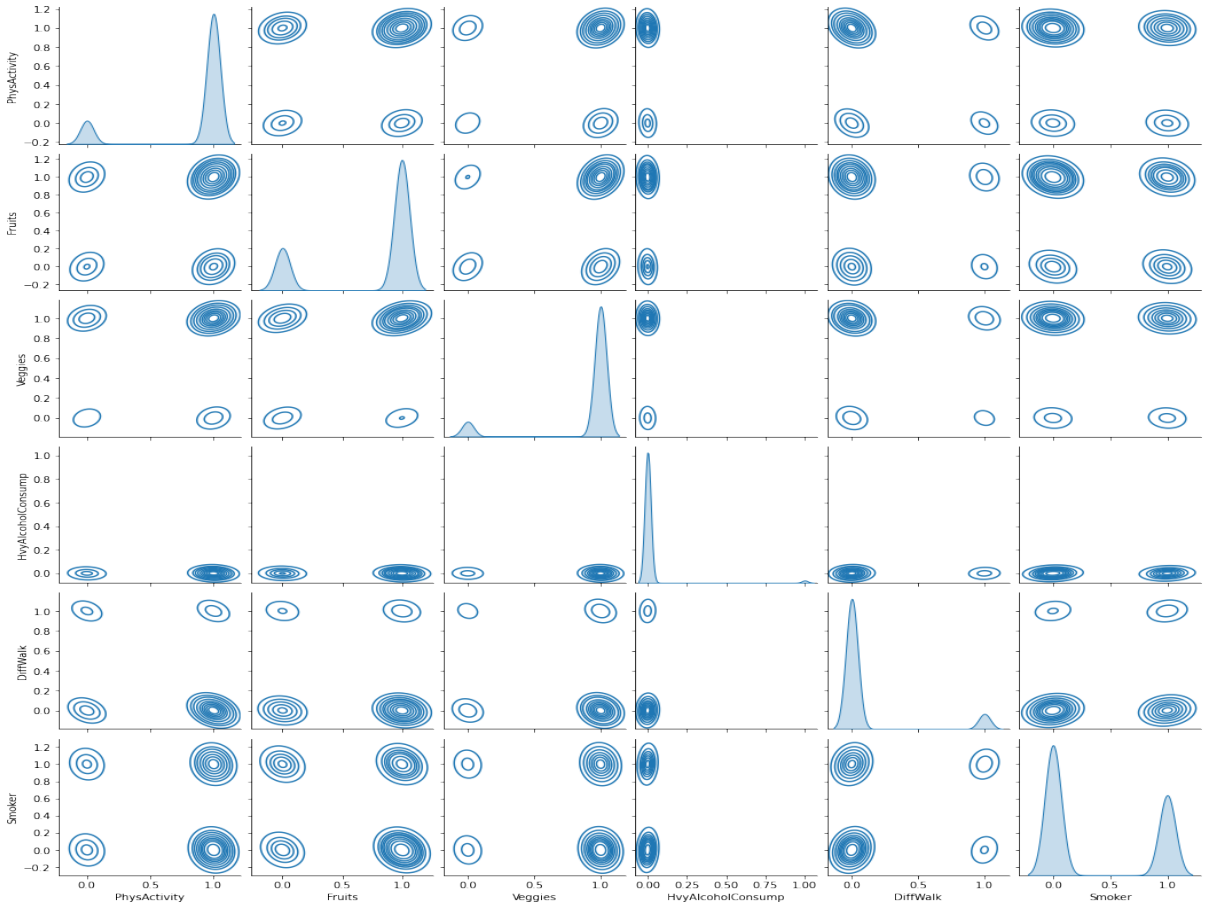


Figure 3.3: Data Distribution of Healthy Lifestyle

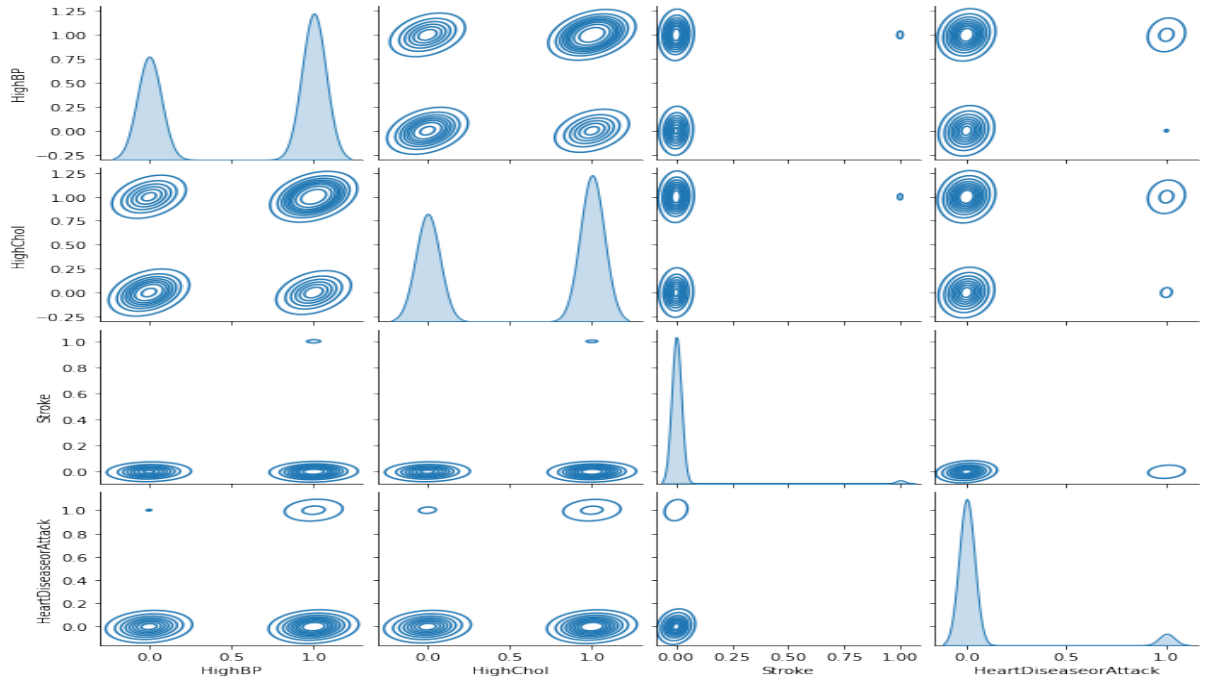


Figure 3.4: Data Distribution of Vulnerable Health State

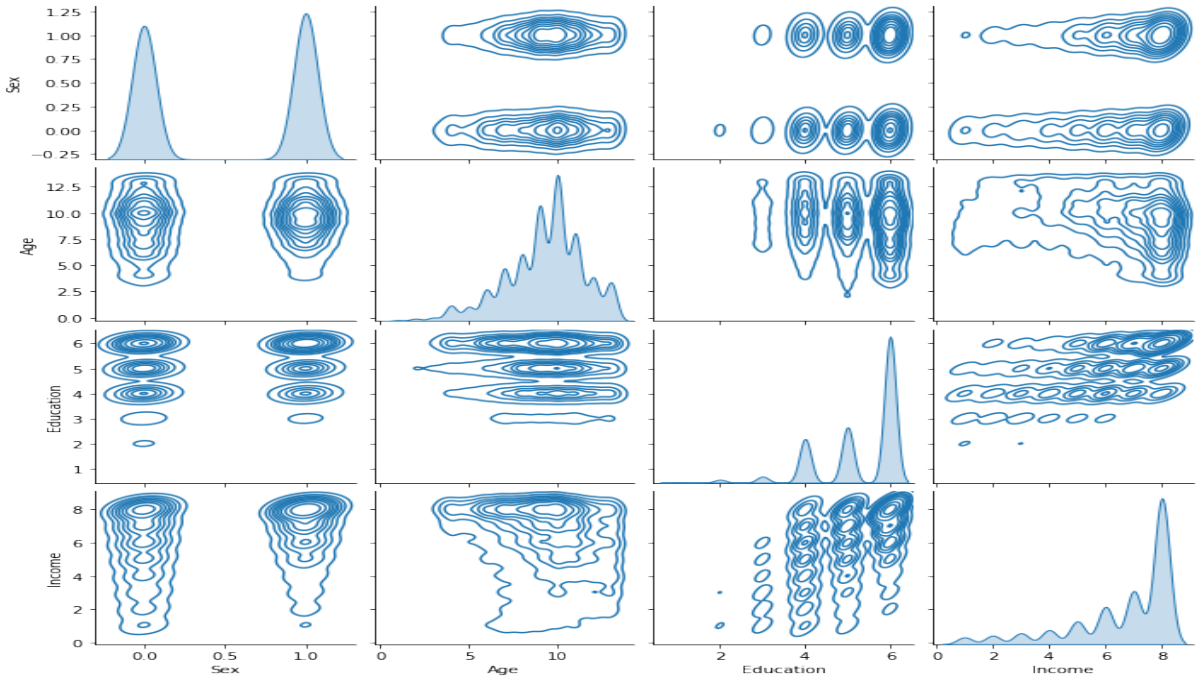


Figure 3.5: Data Distribution of General Lifestyle

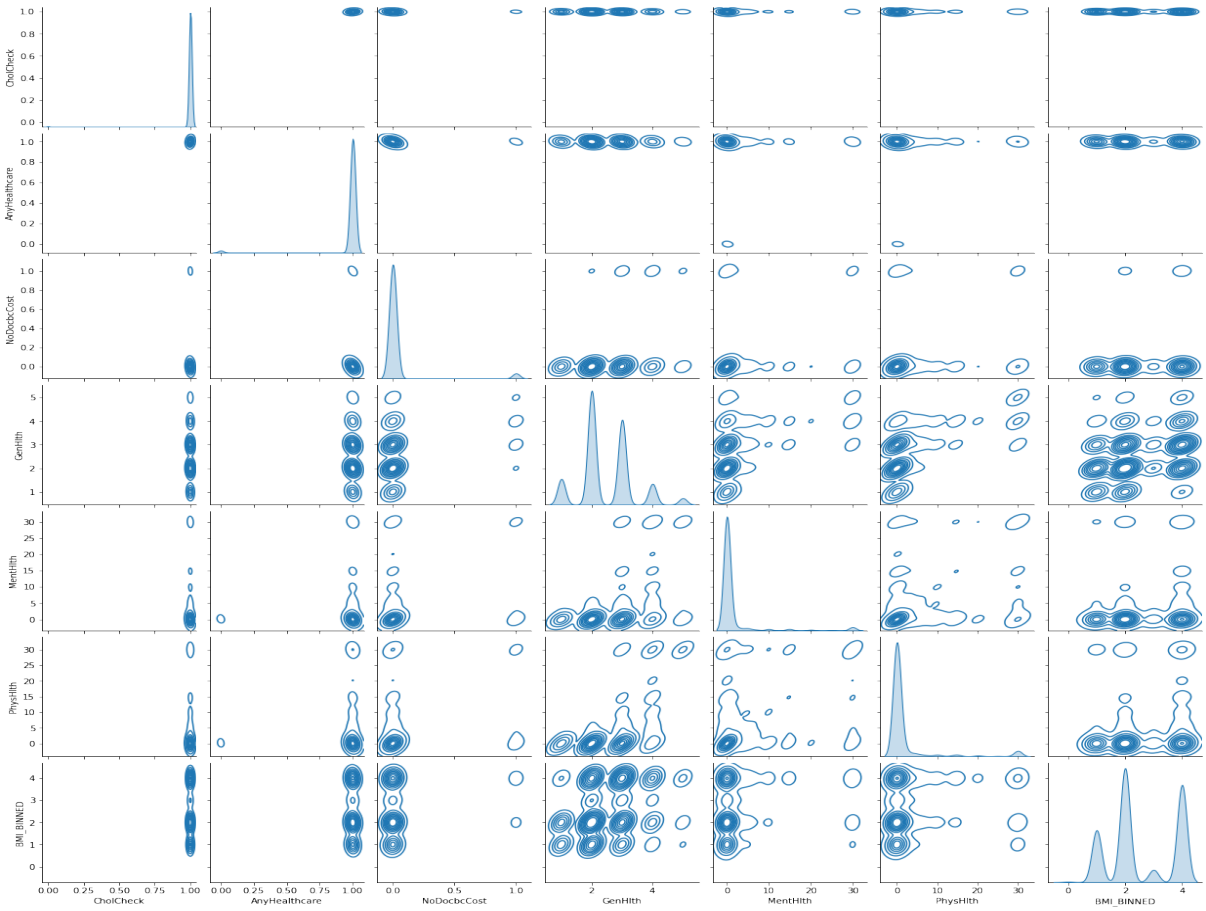


Figure 3.6: Data Distribution of General Health State

After analyzing the figures, it could be concluded that the following features' data distribution were poor.

Veggies	HvyAlcoholConsump	DiffWalk	Stroke	HeartDiseaseorAttck
CholCheck	AnyHealthcare	NoDocbcCost	MentHlth	PhysHlth

We removed these features and train-validated the dataset. The accuracy got was 0.67.

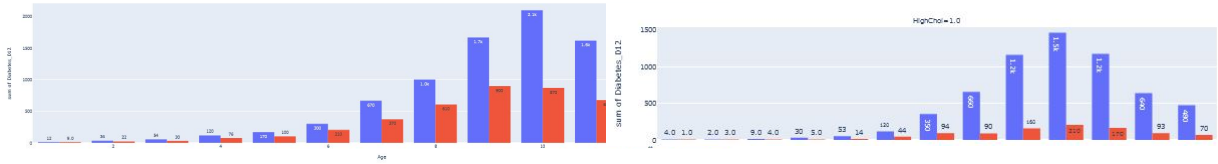
3.2.1.1 More Plotting Description

We furthermore plotted different types of graph such as barplots, piecharts etc. We found out that most of the people are involved in physical activity, eating veggies, fruits and checking cholesterol level. Therefore such features don't help us in anyway to distinguish between the classes. On doing some small research on diabetes we discovered that BMI has a large impact on diabetics and we confirmed it by plotting the data.

(a) Cholesterol Check Bar Diagram

(b) Physical Activity Bar Diagram

Figure 3.7: Cholesterol and Physical Activity Bar Diagrams



(a) Eating Fruits Bar Diagram

(b) Eating Veggies Bar Diagram

Figure 3.8: Eating Fruits and Vegetables Bar Diagrams

3.2.2 Correlation Among the Features and with Target Feature

We plotted correlation heatmap to see the feature's correlation among themselves and also plotted correlation heatmap with the target feature.

From the triangle correlation heatmap, we could clearly see that strong correlation(>0.8) among the features did not exists. So, we did not need to drop any columns.

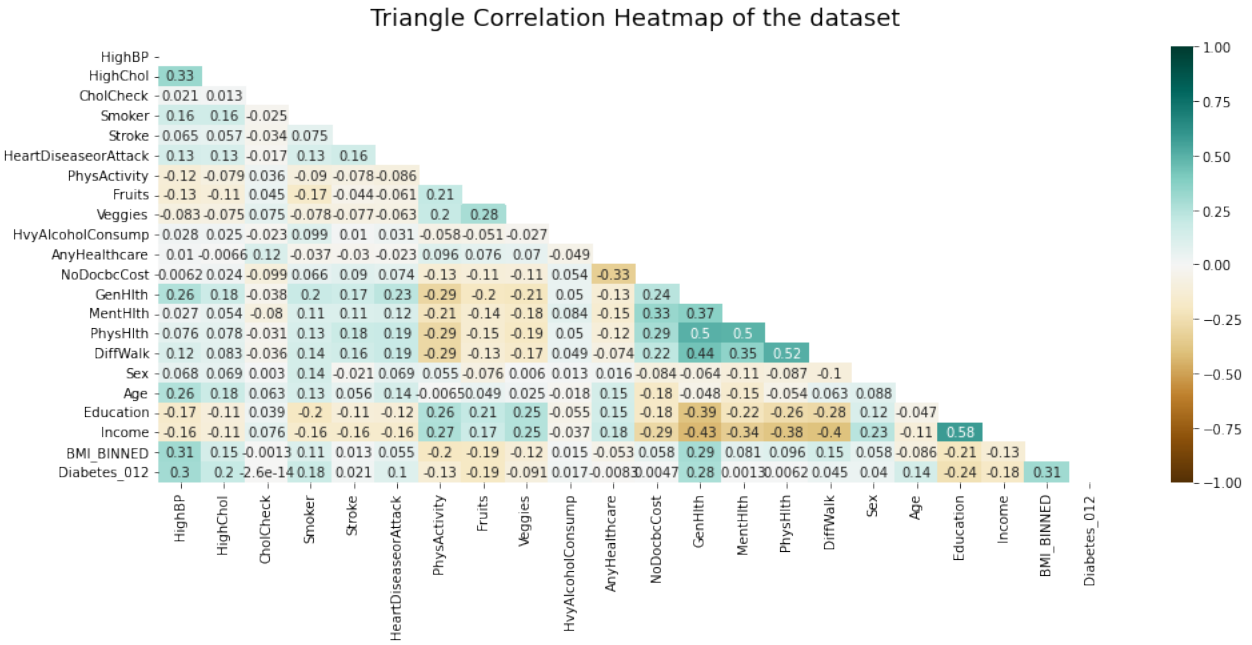


Figure 3.9: Correlation with Each Features

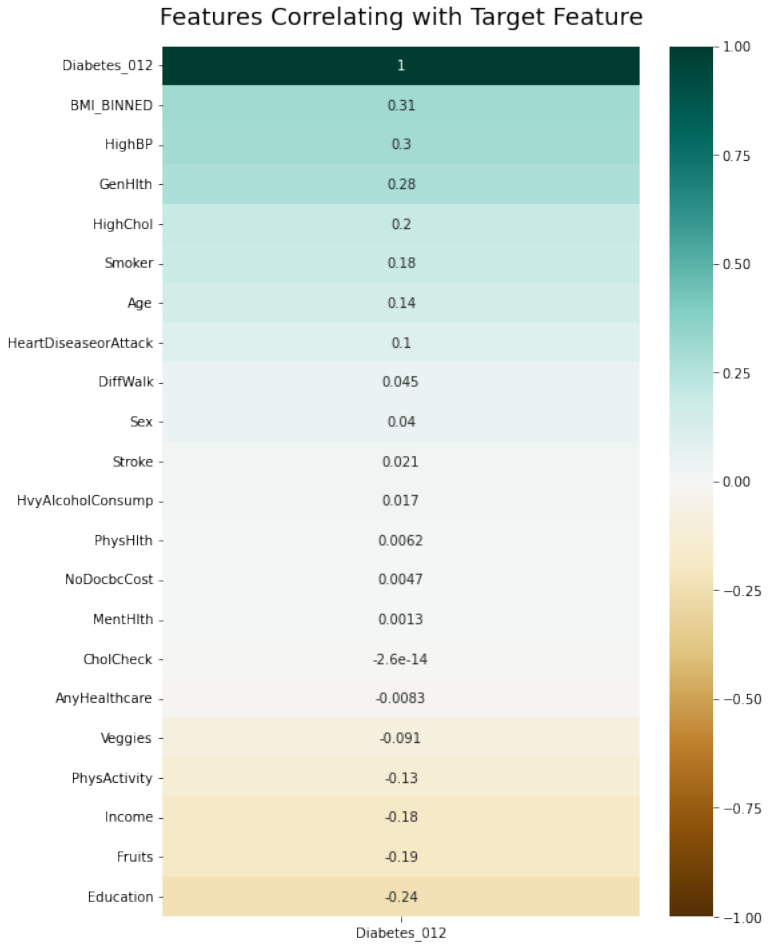


Figure 3.10: Correlation with Target Feature

And from the correlation heatmap with target feature, we could see some of the features had neutral correlation with it. They were:

DiffWalk	Sex	Stroke	HvyAlcoholConsump	PhysHlth
NoDocbcCost	MentHlth	CholCheck	AnyHealthcare	

We removed these features and train-validated the dataset. The accuracy got was 0.66.

3.2.3 Implementing Chi Squared Test

After implementing chi squared test, the features that scored less than 500 are:

HighChol	CholCheck	Stroke	PhysActivity	Fruits	Veggies
HvyAlcoholConsump	AnyHealthcare	Sex	Age	Education	

We removed these features and train-validated the dataset. The accuracy got was 0.71.

3.2.4 Implementing Extra Trees Classifier

After implementing extra tree classifier, the features that seemed less significant were:

CholCheck	AnyHealthcare	Stroke	HvyAlcoholConsump	NoDocbcCost
HeartDiseaseorAttck	Veggies	Sex	HighChol	Smoker

We removed these features and train-validated the dataset. The accuracy got was 0.71.

3.2.5 Implementing Voting System to Remove Features

As different feature selection methods suggested different features to remove, we tried to find the common features that were suggested by all the methods as not important.

We removed the features with most votes and train-validated the dataset. The accuracy got was 0.74. It was similar with our dataset before the starting of feature selection. And hence we successfully reduced the irrelevant features.

Table 3.1: Voting Table

Vote-1	Vote-2	Vote-3	Vote-4
PhysActivity	DiffWalk	Veggies	HvyAlcoholConsump
Fruits	HeartDiseaseorAttack	NoDocbcCost	Stroke
Age	MentHlth	Sex	CholCheck
Education	PhysHlth		AnyHealthCare
Smoker	HighChol		

3.3 Feature Extraction: Principal Component Analysis

We implemented PCA(Principal Component Analysis) as our feature extraction method.

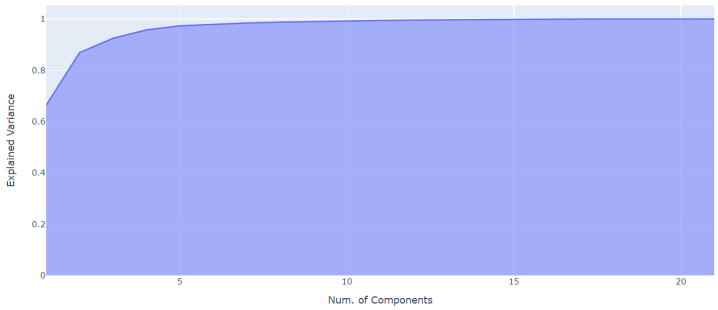


Figure 3.11: Expected Variance vs Number of components graph

From the above graph, we can observe that after number of components 7, the variance curve flattened. We also generated error rate vs number of components graph for our validation set and found out that pca performs better for number of components 15.

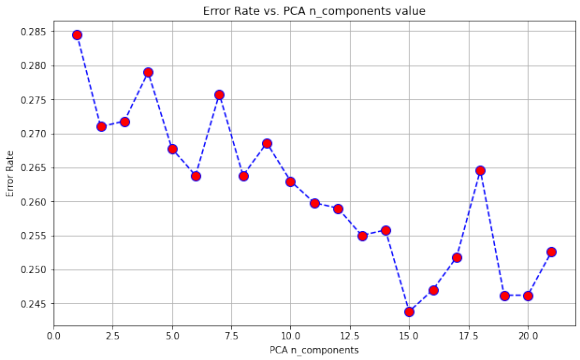


Figure 3.12: Error rate vs number of components graph

For number of components- both 7 and 15, we trained and validated our dataset and for 15, we got our best results where accuracy was 0.76%
And this was our final selected dataset to train-test our machine learning models.

Chapter 4

Methodology

We first collected our dataset, applied decision tree on it with train-validation set and saw that it performed poorly. Then we applied three types of undersampling techniques, compared them and chose the better under-sampled dataset for further proceeding. Then we applied five types of feature selection techniques, compared their accuracy and selected the voting system that gave better accuracy than the other ones.
After that we applied feature extraction method, pca and saw that for the number of components 15, it gave the better accuracy for the validation set than all the previous datasets. Then we applied Decision Tree, Random Forest, Support Vector Machine and K Nearest Neighbors on the test set and finally found Random Forest to be performing better than all the other models.
Our methodology flow diagram is briefly shown in the following.

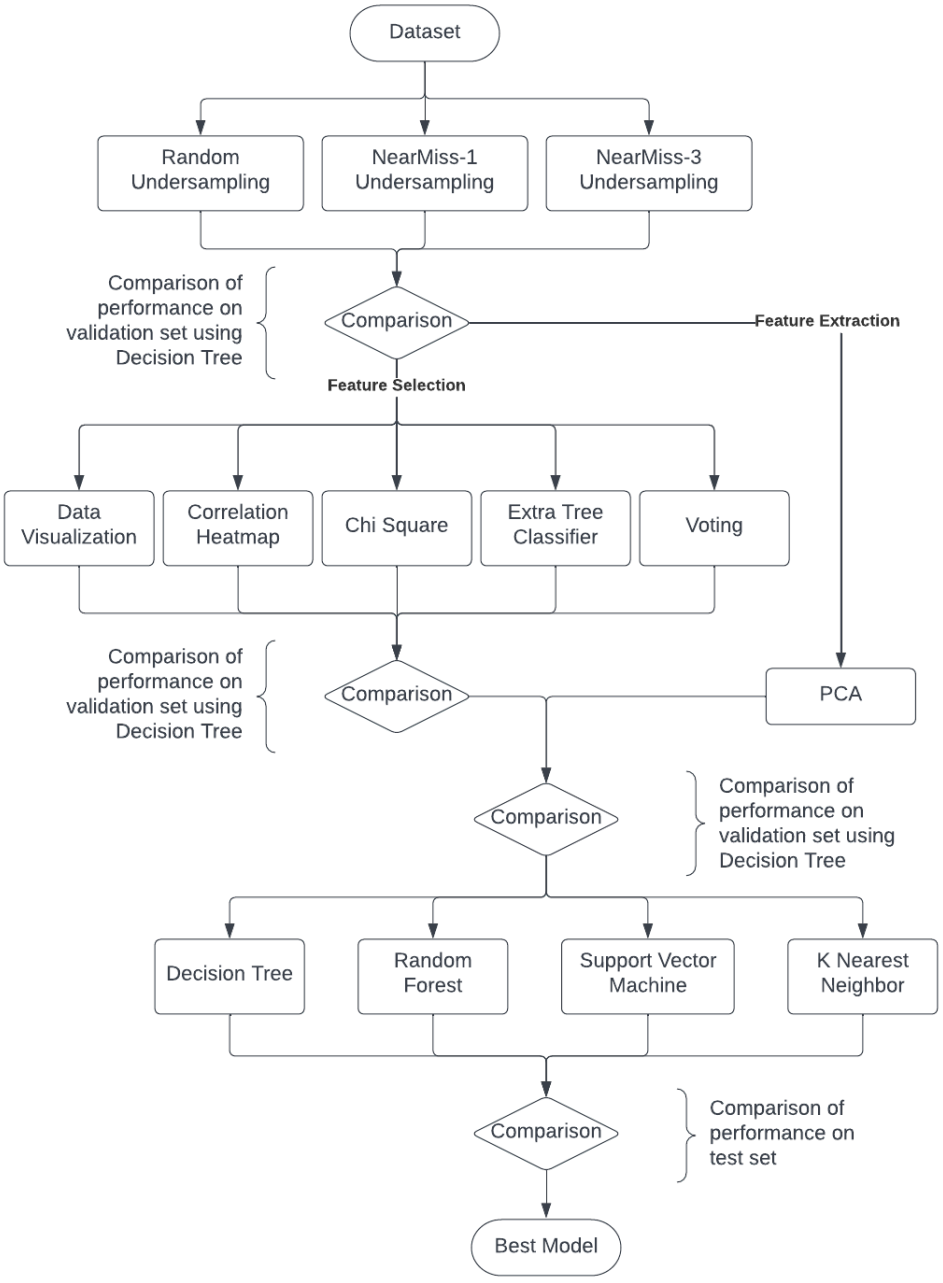


Figure 4.1: Flow Diagram of our Methodology

Chapter 5

Experiments and Results

5.1 Support Vector Machine Results and Discussion

Table 5.1: Support Vector Machine Results

Class	Accuracy	Precision	Recall	F1 Score
0.0	0.74	0.72	0.78	0.75
1.0		0.95	0.68	0.79
2.0		0.63	0.77	0.70

We used linear SVM - one vs rest.

5.2 Decision Tree Results and Discussion

Table 5.2: Decision Tree Results

Class	Accuracy	Precision	Recall	F1 Score
0.0	0.73	0.73	0.92	0.82
1.0		0.75	0.69	0.72
2.0		0.72	0.61	0.66

Obtained accuracy is lesser than SVM's.

5.3 Random Forest Results and Discussion

Table 5.3: Random Forest Results

Class	Accuracy	Precision	Recall	F1 Score
0.0		0.79	0.90	0.84
1.0	0.80	0.84	0.73	0.78
2.0		0.77	0.77	0.77

We put n-estimators value to 50 by tuning in respect to the validation set. We obtained a great increase in accuracy of 80%. Much higher than any other models we have used.

5.4 K Nearest Neighbors Results

Table 5.4: K Nearest Neighbors Results

Class	Accuracy	Precision	Recall	F1 Score
0.0		0.72	0.78	0.75
1.0	0.74	0.95	0.68	0.79
2.0		0.63	0.77	0.70

We defined 10% for test set and 10% for validation set. By tuning with respect to the performance of validation we obtained an optimal value of K. Which is 7. We plotted the errors for different values of K and observed the value 7 has the lowest. By using the value 7 we found accuracy of 74% in the test set.

5.5 Discussion of the Results

Random Forest performed the best among the machine learning models.The best accuracy we got is 0.80. Which is very satisfactory in terms of multi-class classification problem. The other accuracy we achieved were not that bad either.

Chapter 6

Future Work and Conclusion

Doing the experimentation we felt that we needed more features to have more accurate results. Therefore we are intending to get more features and combining with other datasets. Furthermore we have seen that Neural Networks have also performed well, which is described in the academic papers. We have used PCA(Principal Component Analysis) to reduce features, we did not try LDA(Linear Discriminant Analysis) so we will in future apply LDA to observe the results. In the academic papers we also observed the use of AUC/ROC curves, we would also like to illustrate such evaluation techniques in the future.

References

[1] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Peer reviewed: building risk prediction models for type 2 diabetes using machine learning techniques," *Preventing chronic disease*, vol. 16, 2019.

[2] B. Pranto, S. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, S. Momen, *et al.*, "Evaluating machine learning methods for predicting diabetes among female patients in bangladesh," *Information*, vol. 11, no. 8, p. 374, 2020.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Sunday 1st May, 2022 at 5:35pm.