Diabetes Type Prediction using Data Mining

Abstract

Diabetes has been a prevalent topic of study by various medical and data researchers. Research has shown results that there are various types of diabetes of which two have been described as follows. Type 1 Diabetes is insulin-based diabetes, caused by autoimmune destruction of β-cells of pancreas and Type 2 is considered as non-insulin-based diabetes, caused by impaired insulin secretion and resistance to insulin action. From these research, various datapoints were collected and a dataset was created which includes factors which affect a diabetic patient and their diabetic state, i.e. whether they have diabetes or not and what kind of diabetes they are suffering from. The dataset used is a very detailed dataset which includes factors such as Age, Fasting Blood Sugar, Post-prandial Blood Sugar, Random Plasma Count, Fasting Plasma Count and HbA1c for predicting diabetes. It also involves medical details of healthy, non-diabetic patients for the model to compare with and determine the state of a healthy patient as well. We have worked on several models that can predict diabetes from the given medical factors and also predict the kind of diabetes. We saw that almost every model we used, could predict, according to the given medical factors, with atleast 89% accuracy and upto a 100% accuracy for the given dataset. With thorough research on the given models and the feature importances for each model, we have chosen the model which provides us with most accuracy and the best results. We also focused on the rising trend of diabetes amongst young adults and prepared the model to return the percentage of youngsters suffering from diabetes. This research showed us results on 1) The best Data Mining Algorithm we can use to predict diabetes from the given features and 2) The percentage of Young Adults population suffering from diabetes. This research is focused on increasing the speed of diabetes diagnosis process at the patient-level and predict diabetes at an early stage, before it can affect a healthy patient.

Keywords – Data Mining, Machine Learning, Diabetes, Feature Importance, Diagnosis

Index

1. Introduction

Diabetes has been present since ages. Ancient doctors and scientists have also quoted down the presence of diabetes as a disease of urine. Egyptian scientists have quoted it as 'too great emptying of urine' and Indian researchers termed it as *Madhumeha* (Madhu=sweet + Meha=urine). [1] They believed the urine to be tasting sweet because it attracted ants. Diabetes was first coined by Greek

scientist Aretaeus the Cappadocian. The term mellitus was coined by was coined by British General Surgeon John Rollo in 1798 to distinguish the insulin-based diabetes from non-insulin-based diabetes. [2]

In their article, Kerner and Brückel define diabetes as 'general term for heterogenous disturbances of metabolism for which the main finding is chronic hyperglycaemia. The cause is either impaired insulin secretion or impaired insulin action or both.' In easy words, diabetes is a disease caused due to improper insulin action or secretion in the body or both. [3]

With the advent of Diabetes into a mainstream disease, the doctors have found prevention routines and diagnostic methods for the disease. As a group of data scientists, our goal is to make it as easy as possible for the patients to diagnose diabetes at a home level, using Data Mining Algorithms and models. The ADA and World Health Organisation defined the diagnostic criteria for diabetes in terms of Glucose Levels and Glycated Haemoglobin. After going through online repositories, we found a reliable dataset consisting of the various medical records recorded for various people and whether they have diabetes or not. The medical records included in the dataset were Age, Post Prandial Blood Sugar, Fasting Blood Sugar, Random Sugar Level in Plasma, Fasting Sugar Level in Plasma and Glycated Haemoglobin. These medical records are considered vital by ADA and WHO in proper diagnosis of diabetes. [4] The dataset also involves the type of diabetes the diabetic person is suffering from. Although there are various types of diabetes, which have been researched upon and discovered, we have focused majorly on two kinds of diabetes: Type 1 and Type 2. Type 1 Diabetes is insulin-based diabetes, caused by autoimmune destruction of β-cells of pancreas [5] and Type 2 is considered as non-insulin-based diabetes, caused by impaired insulin secretion and resistance to insulin action. [6]

1.1 Problem Statement

With the problems of diabetes in mind, we have used Data Mining Algorithms on the given dataset to make quick and accurate predictions on whether a person has diabetes or not. Our goal is to analyse diabetes as quickly as possible at a patient-level so that they can predict diabetes themselves or prediabetes themselves and can start maintaining prevention routines, before the disease actually starts its affect. As always said, Prevention is better than cure. Being Data Scientists, we have rather focused on the Data Mining Aspects of the problem rather than the medical aspects. As a result, our first Research Question is framed as:

RQ 1 Which Data Mining Classification Algorithm will be best suited in predicting the type of diabetes for a patient?

One of the biggest societal impacts diabetes has been on Young Adults. Obesity and High Calorie Intake has been a featuring trend in today's generation. This has led to a lot of Young Adults being diabetic patients due to not taking proper care of their health. Here we have described Young Adults as population between the ages of 20 and 25 (both included). As a result, Educational Camps against Diabetes have been organised to raise awareness amongst the Young Adults about the disease and its side effects. Here we have used our model to find out the percentage of population, that belongs to Young Adults subset and are suffering from diabetes. So, our second Research Question is framed as:

RQ 2 From a given dataset of population, what is the percentage of Young Adults, as predicted by our model, suffering from diabetes?

2. Dataset

The dataset used in the research is a detailed dataset that includes various medical records of different people that are useful in prediction of diabetes and set as standards by WHO and ADA. The included medical records are as follows:

1) Age: Age (in Years) of the person in account on the day of making of dataset

2) Post Prandial Blood Sugar: It is the concentration of sugar in blood (mmol/L) exactly after 2 hours of a full meal (Blood Glucose Test | Michigan Medicine (uofmhealth.org))

3) Fasting Blood Sugar: It is the concentration of sugar in blood (mmol/L) after the patient in account has fasted for at least 8 hours before the test. (What is a fasting blood sugar level or fasting glucose? | Dexcom)

4) Random Plasma Sugar: It refers to the amount of glucose (mmol/L) circulating in the patient's blood at any random given time (Diagnosing Diabetes :: Diabetes Education Online (ucsf.edu))

5) Fasting Plasma Sugar: It refers to the amount of glucose (mmol/L) circulating in the patient's blood after he has fasted for at least 8 hours (Mean fasting blood glucose (who.int))

6) Haemoglobin A1c: Haemoglobin A1c test refers to the blood sugar (mmol/mol) attached to the patient's haemoglobin. (Hemoglobin A1C (HbA1c) Test: MedlinePlus Medical Test)

The target of the dataset is given through two columns, that represents its names and categorical representation:

1) Type: This column includes the Diabetic state of the patient in account. It states whether the patient is diabetes-free or suffers from Type1 or Type2 diabetes

2) Class: This column is a categorical representation of the column Type. Normal Patients are categorised as Class 0, Type 1 Patients as Class 1 and Type 2 patients as Class 2.

As a general trend in datasets, our dataset also had some noisy data and needed pre-processing, which is further explained in Unit 3. The dataset had some missing values and some extra values which weren't required. It had missing values in columns of Fasting Blood Sugar and Fasting Plasma Sugar which had to be filled in using data pre-processing. Also, there was no proper classification of diabetes in the Class column, i.e. there were only two class representations for three types of patients in the Type column. All the inaccuracies and missing values were handled by pre-processing techniques such as type-classification and mean-filling. The dataset was cleaned thoroughly before any of the Data Mining Algorithm could work on it and as a result, we got very accurate models that could predict with an accuracy of almost 100%.

The raw dataset could be accessed using the following link:

https://drive.google.com/file/d/1W7I3D1fCaZhxHAPOmU0WzL2jux-cN3oi/view?usp=sharing

3. Literature Review

3.1 Literature on Diabetes

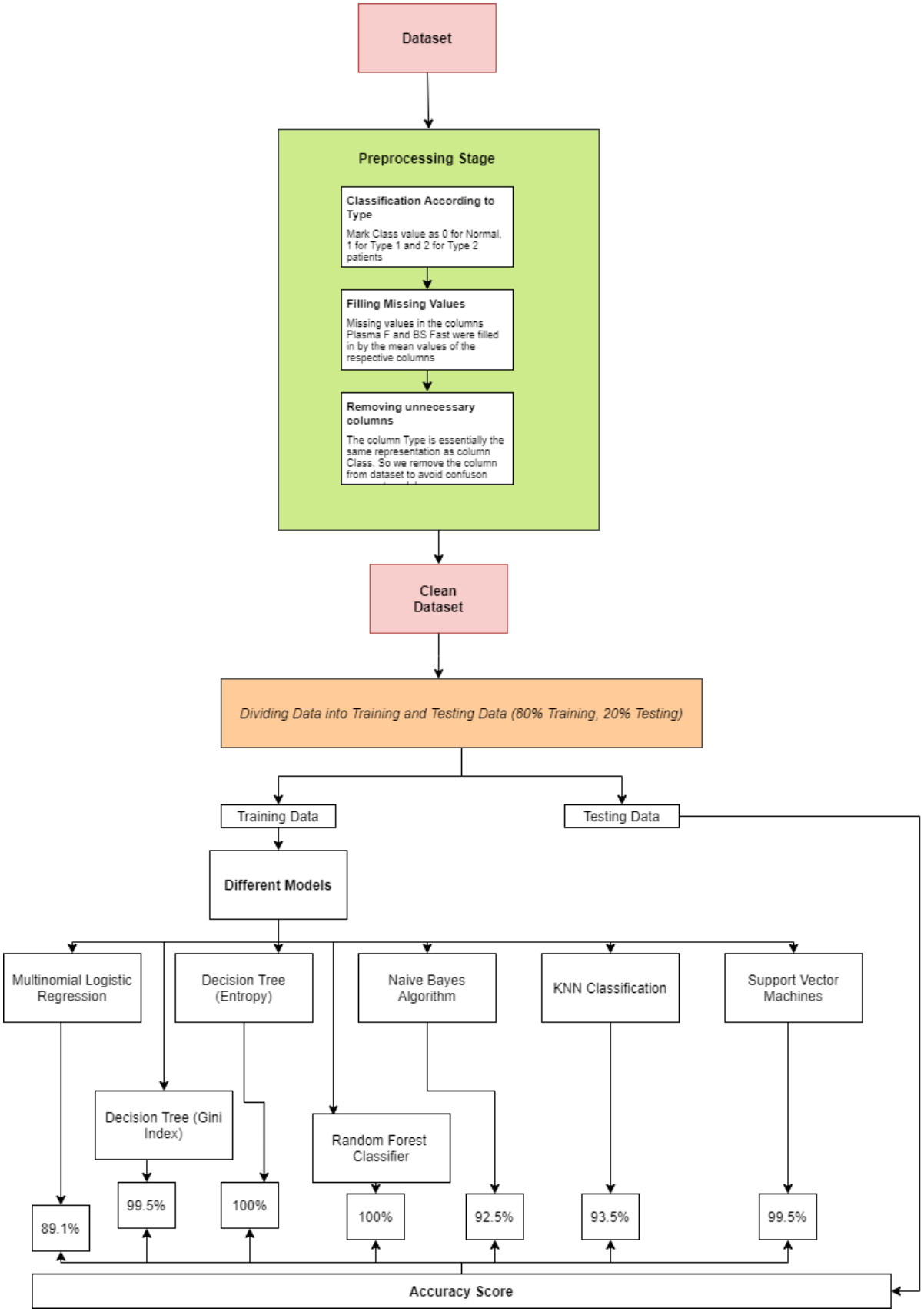
Diabetes as defined by WHO is a serious, chronic disease when the pancreas either doesn't produce enough insulin or body doesn't effectively use the insulin produced by pancreas. [7] Diabetes can be

classified into two different categories: Type 1 Diabetes and Type 2 Diabetes. Atkinson, Eisenbarth and Michels et al. described Type 1 Diabetes as insulin-based diabetes wherein autoimmune destruction of β-cells takes place by pancreas [5]. Chatterjee, Khunti and Davies et. al. described Type 2 Diabetes as a non-insulin-based diabetes caused by impaired insulin secretion and resistance to insulin action [6]. There were many changes recorded in diabetes diagnosis by doctors in 1997, after re-examination of National Diabetes Data Group and WHO study group. This resulted in introduction of new factors or improved factors such as Fasting Plasma Glucose and Haemoglobin A1c. [8]

3.2 Literature on Diabetes Prediction using Machine Learning

Vaidehi and Majumdar, in their work have taken into consideration the PIMA Indian Diabetes Dataset, but have also added other features which they feel were important in Diabetes Prediction such as BMI and Blood Glucose. They have been successful in improving the accuracy from the existing methods upto 96% [9]. JayaMalini and Sonar, in their work used different diabetes dataset and were able to achieve 85% accuracy in diabetes prediction with Decision Tree model. [10] Sisodia and Sisodia et al. used PIMA Indians Diabetes Database and were able to bring out 76.33 % accuracy as highest accuracy with Naïve Bayes Classifier. [11]

4. Architecture



4.1 Models used in Experiment

We have used 7 different learning models for our experiment and received almost 100% accuracy score in 4 different models. Each of the model used is described below:

1) Multinomial Logistic Regression: Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale) [12]

2) Decision Tree using Gini Index: Gini index is an attribute selection method in classical Decision Trees where it returns the best attribute for splitting the node using formula:

$Gini\ (D) = 1-\Sigma_{i=1}^{m} P_i^2$

Where D is a tuple, pi is the probability of D belonging to the class Ci, m is the total number of classes. When the number of classes are large, and the biases are increased, the Gini-based decision tree method is modified to overcome the known problems, by normalizing the Gini indexes by taking into account information about the splitting status of all attributes. [13]

3) Decision Tree using Entropy: Entropy is another attribute selection method in Decision Tree. It returns the best attribute for splitting the node by:
$Info(D)=-\Sigma_{i=1}^{m} P_i \log_2 P_i$
Here D is a tuple, pi is the probability of D belonging to the class Ci, m is the total number of classes. [13]

4) Random Forest Classification: Random Forest is a classifier that grows many classification trees. Each tree is trained on a bootstrapped sample of the training data, and at each node the algorithm only searches across a random subset of the variables to determine a split. To classify an input vector in random forest, the vector is submitted as an input to each of the trees in the forest, and the classification is then determined by a majority vote. [14]

5) Gaussian Naïve Bayes: Gaussian Naive Bayes is a generalization of Naive Bayes Networks, which are a special case of probabilistic networks that allows treating continuous variables [15]

6) K Nearest Neighbors: K Nearest Neighbor (kNN) method is a popular classification method in data mining and statistics because of its simple implementation and significant classification performance. However, it is impractical for traditional kNN methods to assign a fixed k value (even though set by experts) to all test samples. [16]

7) Support Vector Machines: A support vector machine (SVM) is a computer algorithm that learns by example to assign labels to objects [17]

5. Experiment

5.1 Libraries used

We have used various Python Libraries already installed in Google Colab. A python library is a collection of all the accessories required to complete a task under one roof. Here is a complete list of all the libraries and sub-libraries we have used and where we have used:

1) Numpy: For doing complex numerical calculations
2) Pandas: For creating and maintaining Dataframe
3) Seaborn: For data and model visualisation
4) Matplotlib: For data visualisation

5) train_test_split: For splitting data into 80% training and 20% testing parts
6) LogisticRegression: To create Multinomial Logistic Regression model
7) DecisionTreeClassifier: To create Decision Tree Models with both, Gini index and Entropy
8) Graphviz: To visualise decision trees
9) pydotplus: Same as Graphviz
10) Image: To output decision tree as an image
11) StringIO: To save decision tree output as dot file
12) RandomForestClassifier: To create Random Forest model
13) GaussianNB: To create the Gaussian Naïve Bayes model
14) confusion_matrix: To generate the heatmaps for certain models to display accuracy
15) Graph Objects: For various uses in different graphs
16) KNeigborsClassifier: To create the KNN model
17) svm: To create the SVM model

5.2 Pre-processing Stage

There were three steps used in pre processing the data. They are mentioned below and how they were performed:

1) Proper sorting of Class column: Class column had only two values: 0 and 1. While the corresponding Type column had 3 values: Type 1 and Type 2. So, we made a copy of both Class and Type columns and edited the list of Class copy, according to corresponding Type values. Then we replaced the Class column with the edited copy.

2) Missing Values Handling: Two columns; BS Fast and Plasma F had missing values. We tried to check their dependencies on other factors as first step. But when they turned out to be independent values, we filled the missing values with the respective column's mean and fillna method.

3) Column Deletion: Two columns, namely Type and Class, showed the same data, that is the type of patient. As Class represented it in numerical form, which is more necessary in Data Mining algorithms, we removed and stored the Type column in a dummy variable for future reference.

After cleaning up the dataset, the clean data was split into 80% Training and 20% Testing data. The models were all trained on the Training Dataset and provided accuracies back with the Testing data. We got a range of accuracies from just above 80% to 100% accurate models.

6. Observations

6.1 Major Observations:

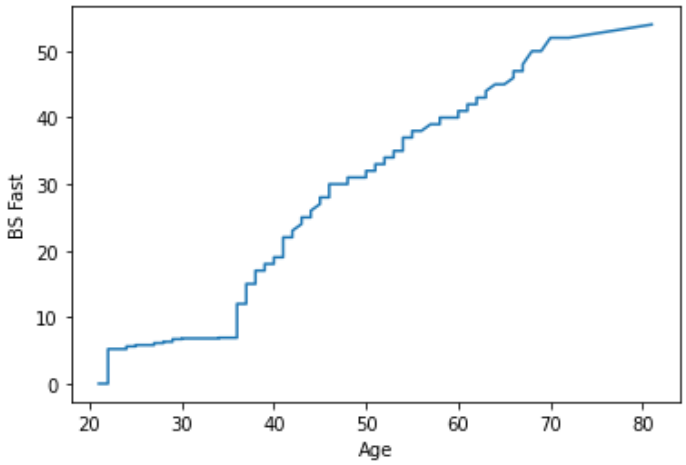
We observed that with changing of the Data Mining Algorithm, the accuracy of prediction equally changed. Some of the data mining models failed in accuracy giving us an accuracy of prediction in range of 82%-94% while the more accurate models maintained an average accuracy in between 98%-100%. There are causes of why the models failed and models succeeded which are explained in further sections. By measuring the Feature Importances of each and every model, we found that Random Plasma Sugar carries a very prominent role in measuring diabetes while factors such as Age and play a comparatively minor role.

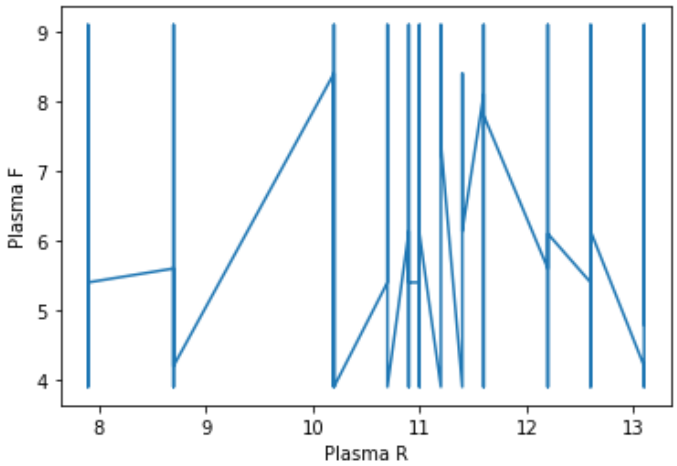
On data visualisation, we found that none of the given factors were dependent on each other. They gave us pretty inaccurate graphs and values. As a result, using mean value to fill missing values in Plasma F and BS Fast columns gave us more accurate results than using Multivariate Linear Regression based on Age and Plasma R, and Age and BS pp columns respectively. This step has been illustrated in the data visualisation and data cleaning process thoroughly.

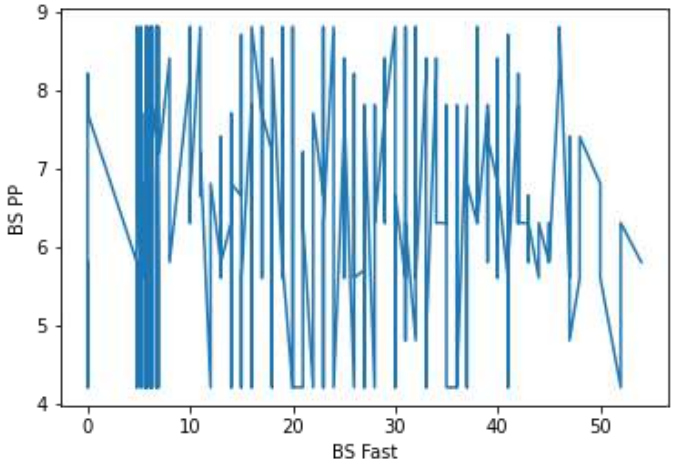
The following table summarises the average accuracy obtained by each and every model:

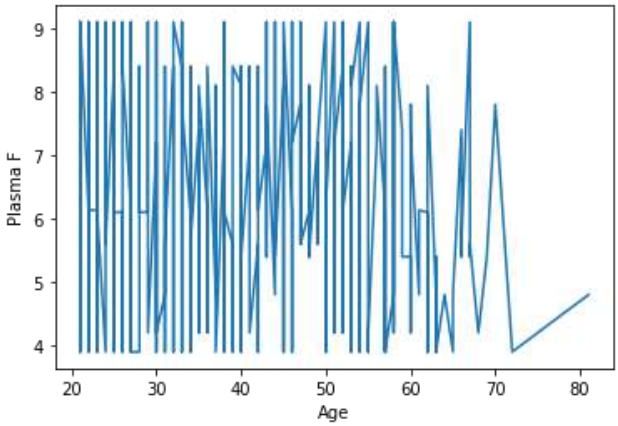
Data Mining Model	Average Accuracy
Multivariable Logistic Regression	82%-89%
Decision Tree (Gini Index)	96%-100%
Decision Tree (Entropy)	98%-100%
Random Forest Classifier	99.5%-100%
Naïve Bayes Classification	88%-94%
K Nearest Neighbors (k=5)	89%-93%
Support Vector Machines	99.5%-100%

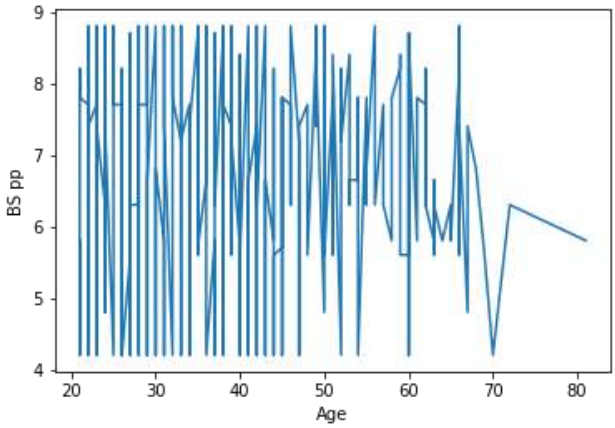
6.2 Data Visualisation Graphs











6.3 Why certain models Succeeded

Models such as Decision Trees, Random Forest and Support Vector Machines gave us pretty accurate results. One of the biggest reasons in their success is because they are supervised models. With the presence of Training Output Data, these models could highly capitalise on the output and form their respective mathematical equations to provide us with the perfect results. The Entropy Decision Tree gave a slightly better accuracy than Gini Index Decision Tree. This might have happened because the former uses Logarithmic Production to provide us with the perfect split while the latter uses Square Function. The logarithmic function provides with much better splits when used on probability values (i.e 0-1) while the square goes a bit inaccurate and might provide us with the wrong split due to confusions araised by squaring.

Random Forest Classifier works more accurately because it takes in consideration a mixture of best splits of various decision trees (in our case 200). It takes in consideration every factor and every data point in the dataframe and hence explains its accuracy.

Due to presence of classified Training output, Support Vector Machine works accurately in linear kernel mode for this classification. Visualising the SVM output will be tough due to presence of various factors, but in a 6-dimensional plane, SVM finds the perfect splits between the three classes.

6.4 Why certain models Failed

Models such as Multivariate Logistic Regression, Naïve Bayes Classification and K Nearest Neighbors failed due to two of them being unsupervised models. Logistic Regression works accurately in binary classification, but when it comes to multi-class classification, the logit function, which is used in the model, fails in providing accuracy. This happens because logit function has a range of 0-1 and here in multi-class classification we break down our set of classes into two classes (0 and 1) and do simple

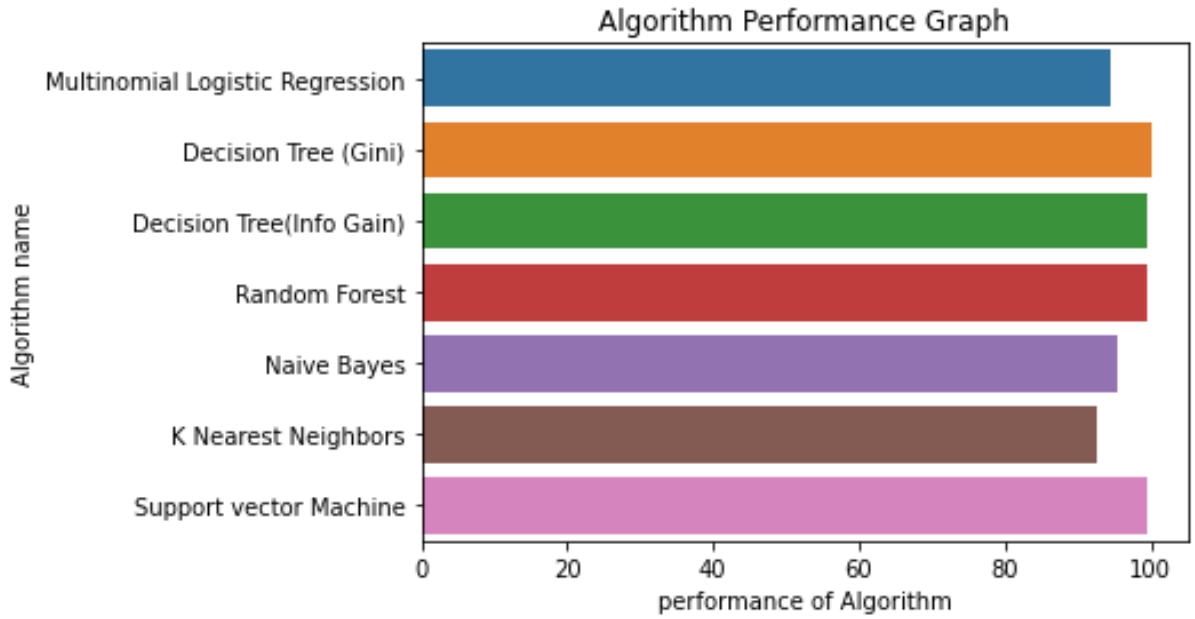
Logistic Regression on them. The second mixed class will be again given as an input to the model. This results in high cases of mixed classification.

Unlike the rest of models, KNN and Naïve Bayes classification don't work on Output Training Data. They cluster the data points together and forms classes. But here, due to limited presence of classes (i.e. 0,1 and 2), these models will fail as they will not be able to accurately cluster datapoints and would also have to adhere to the given training output data. In case our dataset would have been synced with unsupervised models, these models would have provided much better accuracy.

7. Conclusion

In conclusion, we can say Support Vector Machines to be the most accurate model of the 7 to provide with results. Random Forest Classification comes to a close second, but we preferred SVM over Random Forest, because SVM doesn't takes in consideration, the factors that don't play a major role in diabetes prediction. Random Forest also takes in consideration, factors such as Age and , that don't play a major role in diabetes prediction. SVM is also better than Decision Trees due to its consistency. Decision Trees under both the indices, do provide us with 100% accuracy but that completely depends on the training data given to it as input. SVM is independent of the training data used and hence provides us with high accuracy under any given Training Data Conditions.

The following Bar Chart summarises the performances of each and every model:



Coming to second conclusion, the Young Adult population suffering from Diabetes is 12.45% in our dataset. Young Adult population, as described by us, is the population of people in age range 20-25 years of age. These, also referred as Youth, are also the building blocks of any country's development. Having such a high ratio of diabetic youth population remarks as a threat to future. Government should carry out diabetes prevention and awareness campaigns amongst youths to stop this disease at such a young age from further spreading. At an individual level, there is a need of improving the lifestyle of youths. We should make it a practice to daily workout, eat healthy food and avoid all kinds of addictions such as smoking and alcohol, to prevent further spread of diabetes.

On an ending note, we would like to say that diabetes is a very dangerous and fast spreading disease amongst countries, especially in Young Adults. We need to be careful about our lifestyles to ensure

we don't fall into its trap. The proposed model helps early-diabetic or pre-diabetic patients to predict their diabetic state and the type of diabetes they suffer from. This model helps patients at an individual level and speeds up the process of recovery.

8. References

[1] Srinivas, P., Devi, K. P., & Shailaja, B. (2014). Diabetes mellitus (madhumeha)-an ayurvedic review. *Int J Pharm Pharm Sci*, *6*(Suppl 1), 107-110.

[2] Tattersall, R. B. (2017). The history of diabetes mellitus. *Textbook of diabetes*, 1-22.

[3] Kerner, W., & Brückel, J. (2014). Definition, classification and diagnosis of diabetes mellitus. *Experimental and clinical endocrinology & diabetes*, *122*(07), 384-386.

[4] Inzucchi, S. E. (2012). Diagnosis of diabetes. *New England Journal of Medicine*, *367*(6), 542-550.

[5] Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. *The Lancet*, *383*(9911), 69-82.

[6] Chatterjee, S., Khunti, K., & Davies, M. J. (2017). Type 2 diabetes. *The lancet*, *389*(10085), 2239-2251.

[7] World Health Organization. (2016). *Global report on diabetes*. World Health Organization.

[8] Expert Committee on the Diagnosis and Classification of Diabetes Mellitus*. (2003). Follow-up report on the diagnosis of diabetes mellitus. *Diabetes care*, *26*(11), 3160-3167.

[9] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.

[10] Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 367-371). IEEE.

[11] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, 1578-1585.

[12] Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, *44*(1), 197-200.

[13] Rokach, L., & Maimon, O. (2005). Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165-192). Springer, Boston, MA.

[14] Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.

[15] Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*, 105361.

[16] Laaksonen, J., & Oja, E. (1996, June). Classification with learning k-nearest neighbors. In *Proceedings of international conference on neural networks (ICNN'96)* (Vol. 3, pp. 1480-1483). IEEE.

[17] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.