Journal of Control System and Control Instrumentation
Volume 4 Issue 3

An Analysis of Predicting Diabetes using Machine Learning

[1]Mr. Ujjwal Anand, [2]Dr. Amit Sehgal, [3]Mr. Shashank Tripathi, [4]Mr. Gagandeep Singh, [5]Mr. Rinku Sharma, [6]Ms. Manisha

[1,4,6] Innovatores, [3] Project Engineer, [5] Research Associate,
Department of Electronics & Communication Engineering, G L Bajaj Innovation Center, Greater Noida, UP, India.

[2] Professor, Department of Electronics & Communication Engineering, GLBITM, Greater Noida, UP, India.

Email: *[1]ujjwalanand1997@gmail.com, [2]amit.sehgal@glbitm.org, [3]shashanktripathi030@gmail.com*

DOI:

Abstract

Diabetes comes under chronic disease, in which cells are not able to use blood sugar (glucose) efficient enough for energy. This condition arrives when the cells become non-responsive to insulin and the blood sugar increases gradually. The known types of diabetes are, Type1, Type 2 and Type 3. Type 1 and type 2 diabetes are in hyperglycemia category (caused by increase in blood sugar), while Type 3 diabetes (Alzheimer's disease) which is caused by resistance to insulin in the brain. Prediction of preliminary stage diabetes is very important as it becomes worse in next stages. This Prediction can be done using machine learning classification models, which are more widely being used for other medical purposes. To predict, if a person is diabetic we need data about insulin, blood pressure, skin thickness and glucose. This data will be fitted in classification models of machine learning with a target vector of conclusion. This will prepare a model that can predict if a certain patient is a diabetic or not. We have implemented many classification models on the data. We have also used neural networks to serve the same purpose.

Keywords: *Diabetes diagnosis, Medical ontology.*

INTRODUCTION

Machine learning is one of the technique to withstand computational capacity of a computer, where a computer is programmed with the ability to learn and improve its performance on specific tasks. In Short, machine learning is all about analyzing data, extracting information which is used to make predictions, check whether the prediction was correct, and if in correct, learning from that to make a more correct prediction in the future.

Machine learning is being used in many parts of our life. Nowadays, it is being much implemented in the Medical diagnostics.

Medical diagnostics are a category of medical exams completed to become aware of infections, situations and sicknesses. These days, system mastering is gambling a key issue in reaching automation in clinical diagnostics.

Diagnosis via machine learning works while the condition may be reduced to a classification venture on physiological facts, in areas where we presently rely upon the clinician in an effort to visually perceive patterns that indicate the presence or form of the condition.

Classification is a technique to solve a problem where the output that we want is a category, such as “red” or “blue” or

Page 16-25 © MAT Journals 2018. All Rights Reserved

"disease" and "no disease"[1]. A classification model makes computation to come on conclusions using observed values or labelled data. Given one or more inputs to a classification model, it can predict the value of one or more outcomes. In case of Medical Diagnostics, the output we want will be either a "disease"or"nodisease",orinmoredeepclassification,whichtypeofacertaindisease is it, for example to predict which type of diabetes is it, the output we desire should be, "type 1" or "type2"[2].

The data is the most important part in machine learning, and no prediction model can be implemented without data[3].
The data can be a table of multiple records of patients, or the images of different patient's diseased parts. For example, if we want to detect tuberculosis, then we need to train our model with chest x-rays of tuberculosis patients, but if we want to detect if there is a presence of diabetes in a patient, then we will need a feature matrix containing the features of patients, and another list of conclusion if the patient is having the Diabetes or not.

Table: 1. *Feature Matrix for patients came for Diabetes diagnosis*

Blood Pressure	Insulin	Glucose	Skin Thickness	....
72	85	148	35	....
66	94	89	23	....
84	230	118	47	....
88	235	126	41	....

Then there will be another list containing the conclusion of diabetes in patients.

The machine learning model will be trained according to the feature matrix and conclusion list.

Overfitting in Machine learning models
In this condition, the model learns most of the detail and noise available in the training data to that extent, that it negatively influences the presentation of the model on new data. Actually, the noise or random variations in the exercise data is picked up and learned as concepts by the model[4].This comes on a problem that we need to tackle in every machine learning model is that when the new data will come for prediction, it will show wrong result.

This problem of overfitting, will become worse in medical diagnostics. As in other predictionsitwouldmaypredictwrongandincuralossofmoneyorelse,butincaseof medical diagnostics it would be a matter of life and death of a patient, i.e. if a patient is being diagnosed for diabetes and model made a prediction "negative", then the patient's health may lose in further days and this might make his/her health worse. So, we need to be very careful in avoid overfitting.

Data Pre-processing
Data pre-processing means altering raw information into a comprehensible format. Real data is mostly imperfect, in consistent and complicated. During the capturing of data there are certain features that are not required. During the pre-processing step all such features are removed. Pre-processing steps:
Data Reducing: Data reduction means selecting only desired features. We have

Page 16-25 © MAT Journals 2018. All Rights Reserved

used Back Prorogation algorithm to determine the features which affect the data most. We found out that Glucose level, Insulin level, Skin thickness and Blood Pressure affect the data most.
Data Cleaning: Data cleaning has lot of steps such as filling missing values, normalize the data and resolving in consistency.

Cases in Data Cleaning:
CASE 1 (Using mean to handle missing values):
Missing data is one of the common problems in data which can actually have a significant effect on the conclusions that can be drawn from the **data**. So, it is important to fill all missing values. Here we are using mean to fill all missing values and mean is calculated for each column individually. The results after replacing the missing data were astonishing, with only slight changes in data we were able to get a high accuracy.
CASE 2 (by replacing missing values with custom values):

In our case, we have taken avg. Values of all features of a healthy person and a diabetic person.Thevaluewegetisthenusedtofillallthemissingvaluesinthedataset. In our case Skin thickness was missing for many rows. We replaced it with the avg. values in accordance with whether he is healthy or diabetic.

Table: 2. *custom values for missing data*

	Skin Thickness	BP	…
Diabetic	30	130	
Non-diabetic	22	120	

Classification Models implemented
Decision Tree: Decision tree algorithm is a supervised earning algorithm that can be used for both regression and classification problems[5].Decision tree is a rule based algorithm that creates its own rules at the time of training. Decision tree tries to solve a problem by tree representation. When we used decision tree with case 1(i.e. replacing missing values by mean) the accuracy was not very great. It was low due to fact that it was replacing every missing value with mean which is not a good idea.

In second case where missing values were replaced by threshold of health and diabetic, the accuracy was pretty high than first time. It was because of handing missing values in a better way.
Random Forest Classifier: Like decision tree it is also supervised learning algorithm. Random forest builds multiple decision tree and uses them to give more accurate results. Higher number of trees in random forest can lead to more accurate results.

Random forest produces lightly better result than decision tree for case1.However, it was also not in acceptable state due to low accuracy.
There was a slight improvement in accuracy when the random forest was used in case

Page 16-25 © MAT Journals 2018. All Rights Reserved

There was slight (2%) increase in accuracy. It is because random forest uses number of decision trees to predict are sult.

K-Nearest Neighbor (KNN): KNN algorithm uses feature similarity to find its neighbors. During the training period it tries to distribute the data into different classes. When a new data arrives, the algorithm finds features in the data and then finds its closest neighbor or class. In this model we had only 2 classes- diabetic or non-diabetic. The closest neighbor at the end is calculated using distance vector.

KNN didn't do a good job in case 1 when the data was replaced with mean values.
Its accuracy was lower than random and decision.

KNN did better in case 2. It was able to find features very accurately. However, its accuracy was little less than random forest classifier.

Support Vector Machine: SVM is mostly used for classification but it can used for both regression and classification problems. In this classification model, we plot each data point in n-dimensional space consisting of each feature at a particular data point. We then find a plane which can separate classes.

The main important aspect in svm is kernel, which decides the type of classification. There are 3 main types of kernel in svm, namely, Linear, poly and rbf.
Linear kernel is for linear classification of data, which is not much helpful in case of sparse data.

Polykernel is for polynomial classification, whereas, rbf (radial basis function) is the most popular svm kernel.

Weused"rbf"kernelandSVMdidn'tdogoodi nbothcases.Thekernelfunctionisa measure of similarity between two sets of features. Hence, SVM is not an appropriate choice.

Naïve Bayes: Naïve Bayes uses conditional probability to predict the class of unknown data. Conditional probability[**8] is** the probability that something will happen, taken as something has occurred already[6].
□(|) = □(|) ∗ □()/□()
Where

P (A) is the probability of hypothesis H being true. This is known as the prior probability.
P (B) is the probability of the evidence (regardless of the hypothesis). P (B|A) is the probability of the evidence given that hypothesis is true.

P (A|B) is the probability of the hypothesis given that the evidence is there.
In both cases Naïve Bayes didn't do a good job.

Artificial Neural Networks (ANN): [9] this works as an information processing system that works in the same way as the like biological nervous systems work, such as the brain works in processing facts[5]. It comprises of a large number of highly intersected processing elements (neurons) working in unison to solve a particular problem. An ANN is arranged in single for a precise application, such as pattern acknowledgement or information sorting, using knowledge procedure[7].Knowledge in biological systems involves adjustments to the synaptic connections that exist between the neurons[8].

However, we experimented our data on many models, but the best result that we got is on a 5-layer neural networks, Having 3 hidden layers consisting of 15, 8 and 15 neurons. In both cases we get similar results [9].

Page 16-25 © MAT Journals 2018. All Rights Reserved

But, we do not recommend to use neural networks for medical diagnosis classification, as there is a wide range of fluctuation in results.

RESULTS

The Accuracy for each model can be obtained from confusion matrix of each model that we get from 768 people data, where we took 192 patients data as the test set, on which the model is tested. The results from the model and the original data are matched to see if the output is correct or not.

Table: 3. *Confusion Matrix for models*

	0	1
0	111	19
1	6	56

Decision Tree

	0	1
0	117	13
1	30	32

Gaussian

	0	1
0	117	13
1	11	51

KNN

	0	1
0	115	15
1	12	50

Random Forest

	0	1
0	130	0
1	62	0

SVM

	0	1
0	117	13
1	30	32

Artificial Neural Network

This Confusion Matrix Actually shows, how many test sets out of 192 patients where correct. Through this we can calculate the accuracy of each model. In the confusion matrix, the values at [0,0] and [1,1] is correct, while values at [0,1] and [1,0] are wrong

Table: 4. *Accuracy of Model*

Model Name	Accuracy
DecisionTree	86.9%
Gaussian	77.6%
KNN	87.5%
Random Forest	88.7%
SVM	67.6%
ANN	73.6%

Page 16-25 © MAT Journals 2018. All Rights Reserved

However, the accuracy is quite speaking about the model to be selected as best on the data that we used to build the model.

Simulation Setup

The model that is implemented is coded on software **anaconda (python 3.6.5),** which runs on a system having a configuration of 8GB RAM, 1TB Hard Drive and core i3 generation 5processor.

In Anaconda, An IDE named **Spyder** is used to run the code.

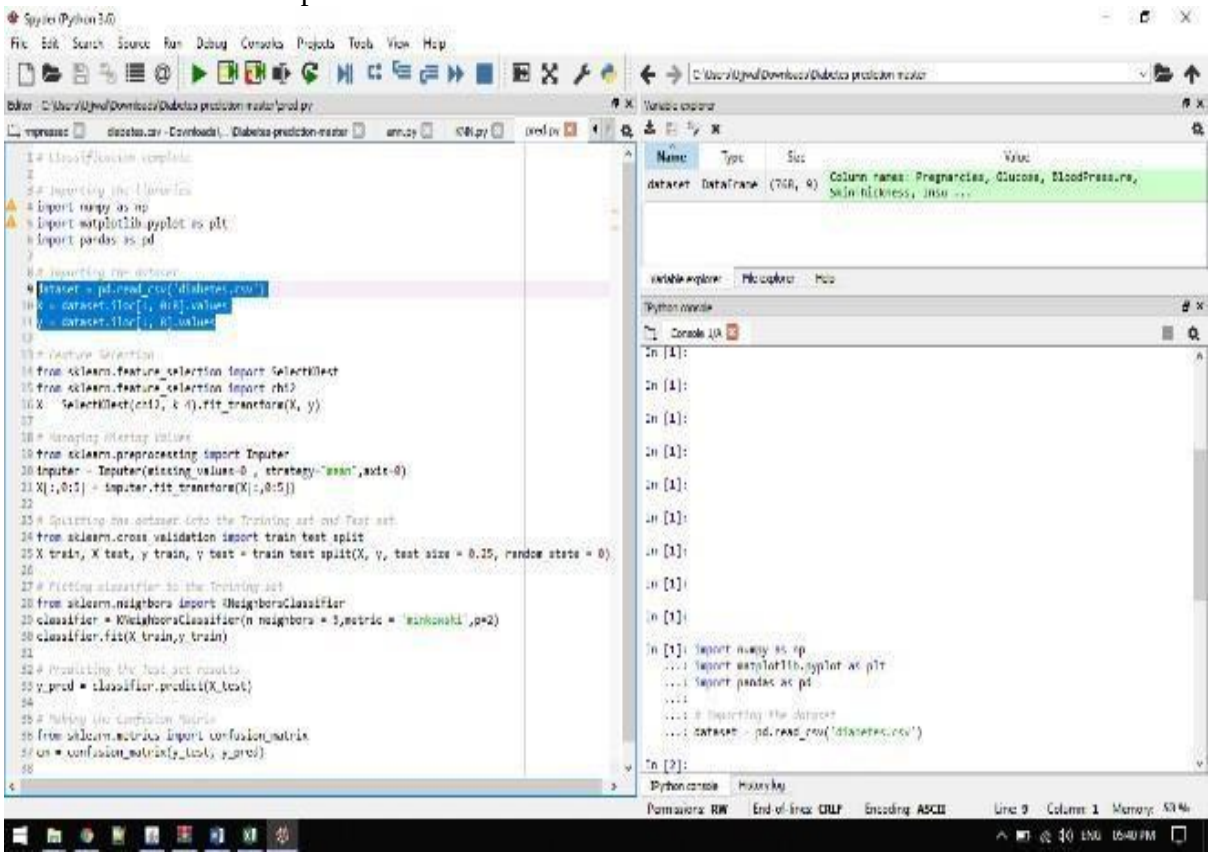


Fig: 1. *Importing Libraries and reading the input data which is in CSV format.*

Page 16-25 © MAT Journals 2018. All Rights Reserved

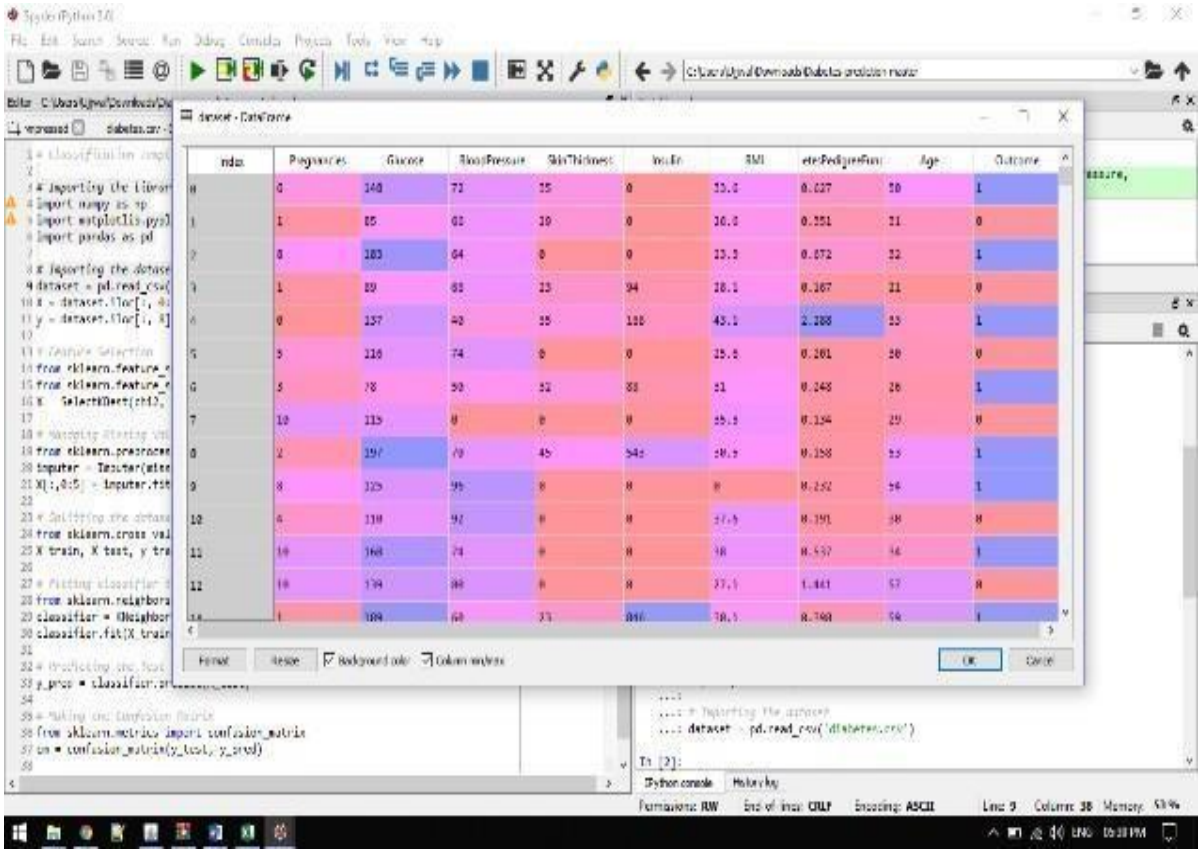


Fig: 2. *The dataset imported is like this.*

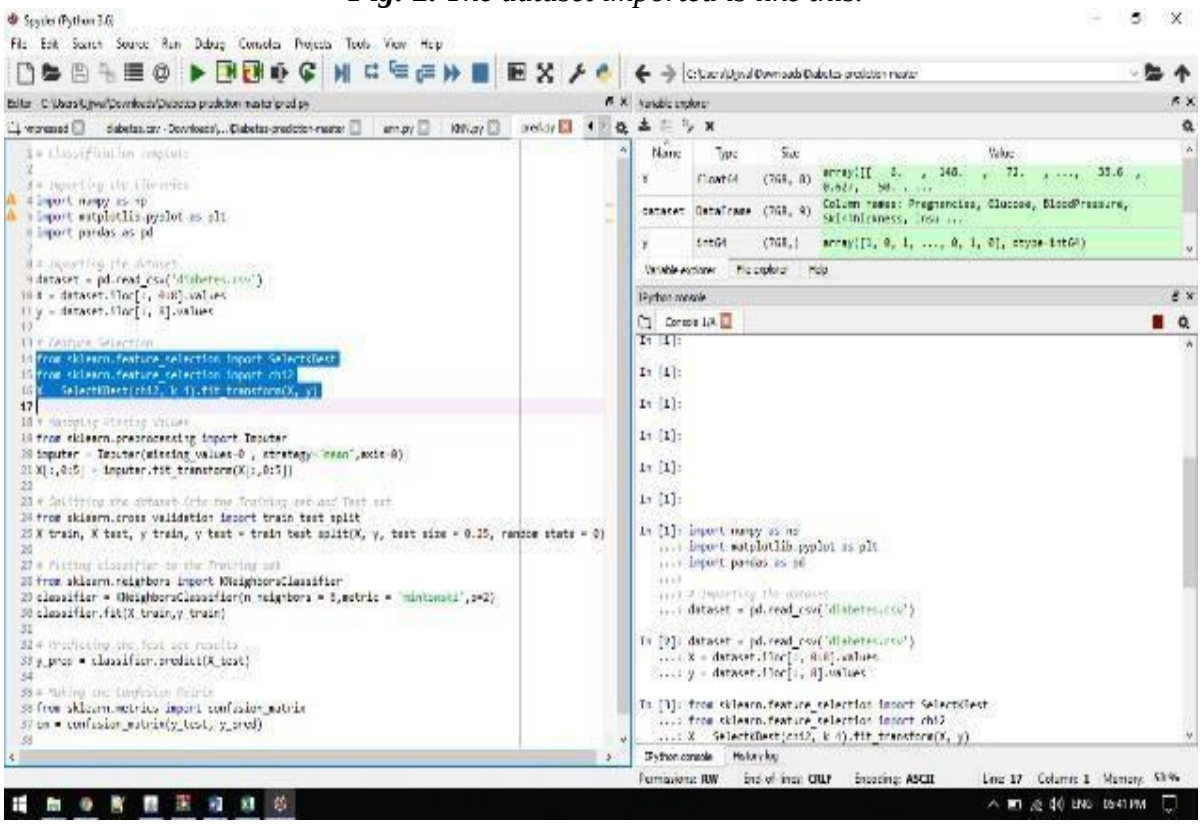


Fig: 3. *Feature Selection After feature selection we got just 4 features, i.e. Glucose level, Insulin level, Skin thickness and Blood Pressure.*

Page 16-25 © MAT Journals 2018. All Rights Reserved

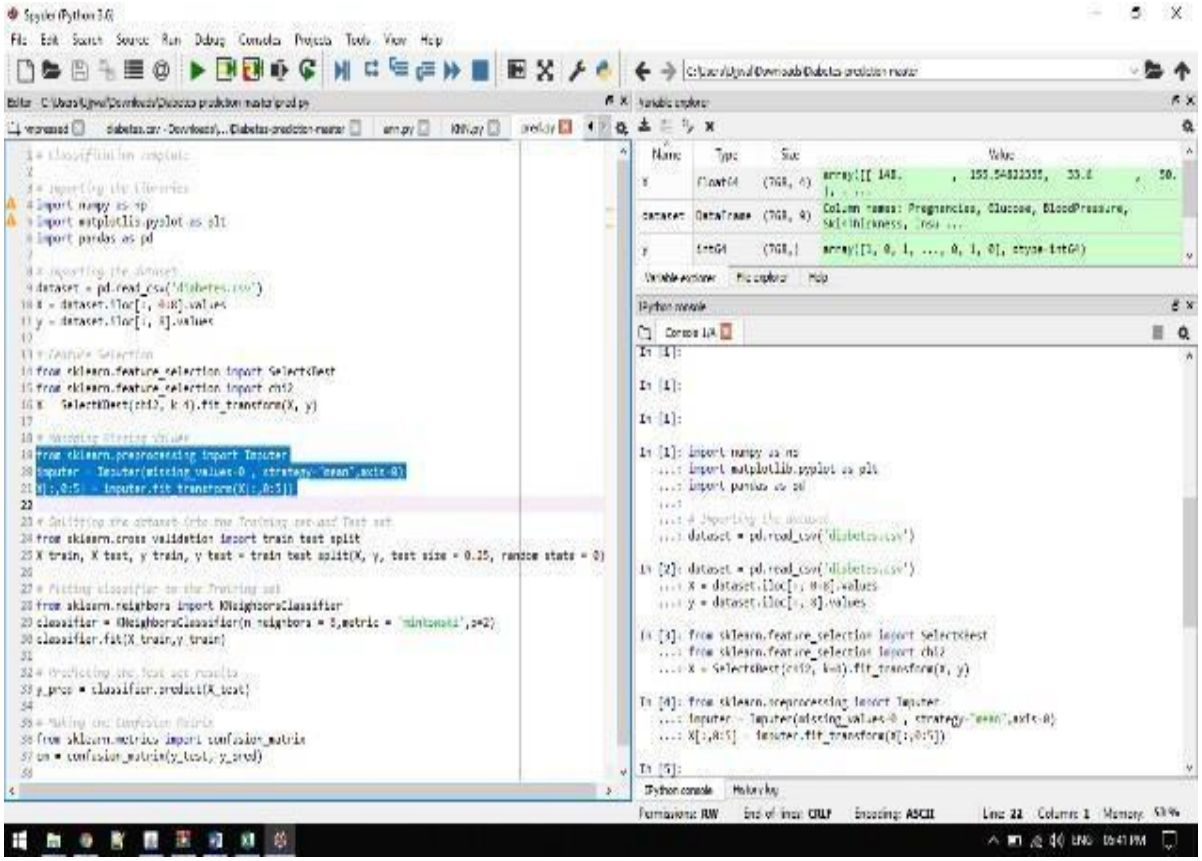
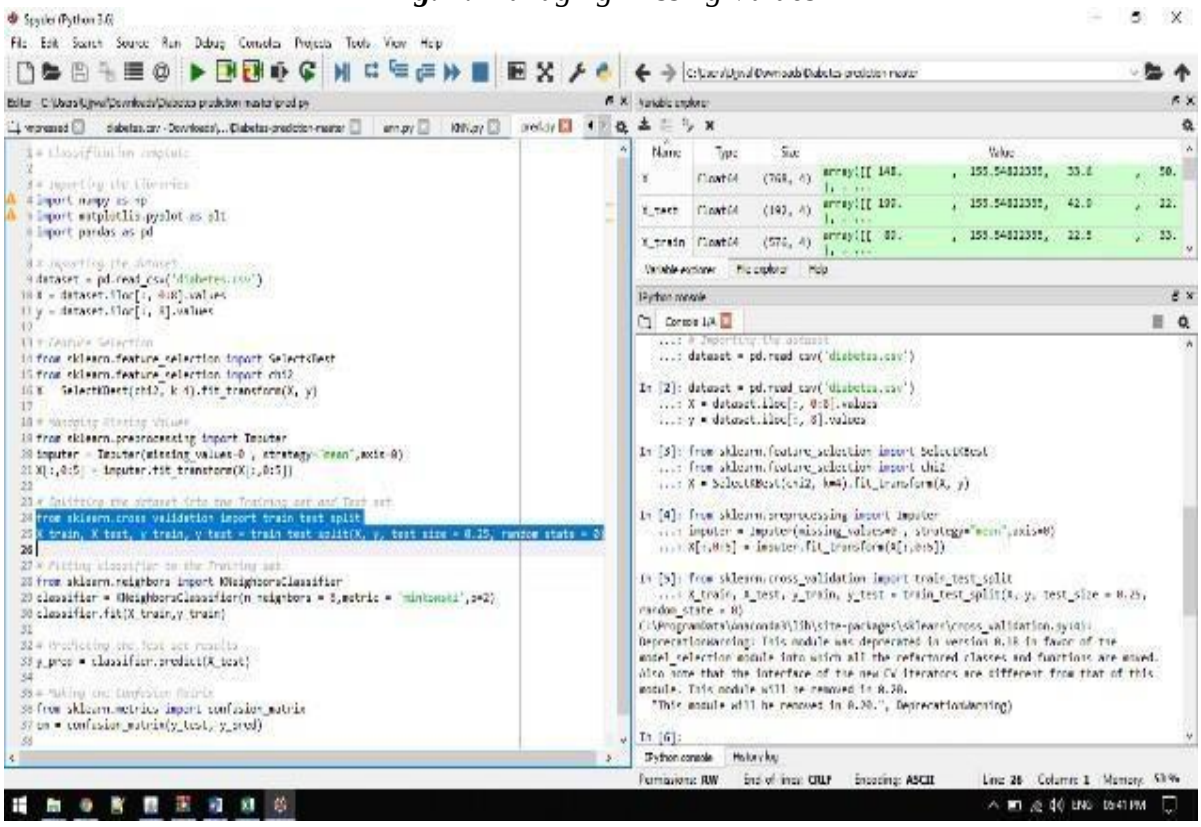


Fig: 4. *Managing Missing Values*



Page 16-25 © MAT Journals 2018. All Rights Reserved

Fig: 5. *Splitting the dataset in training and test set*

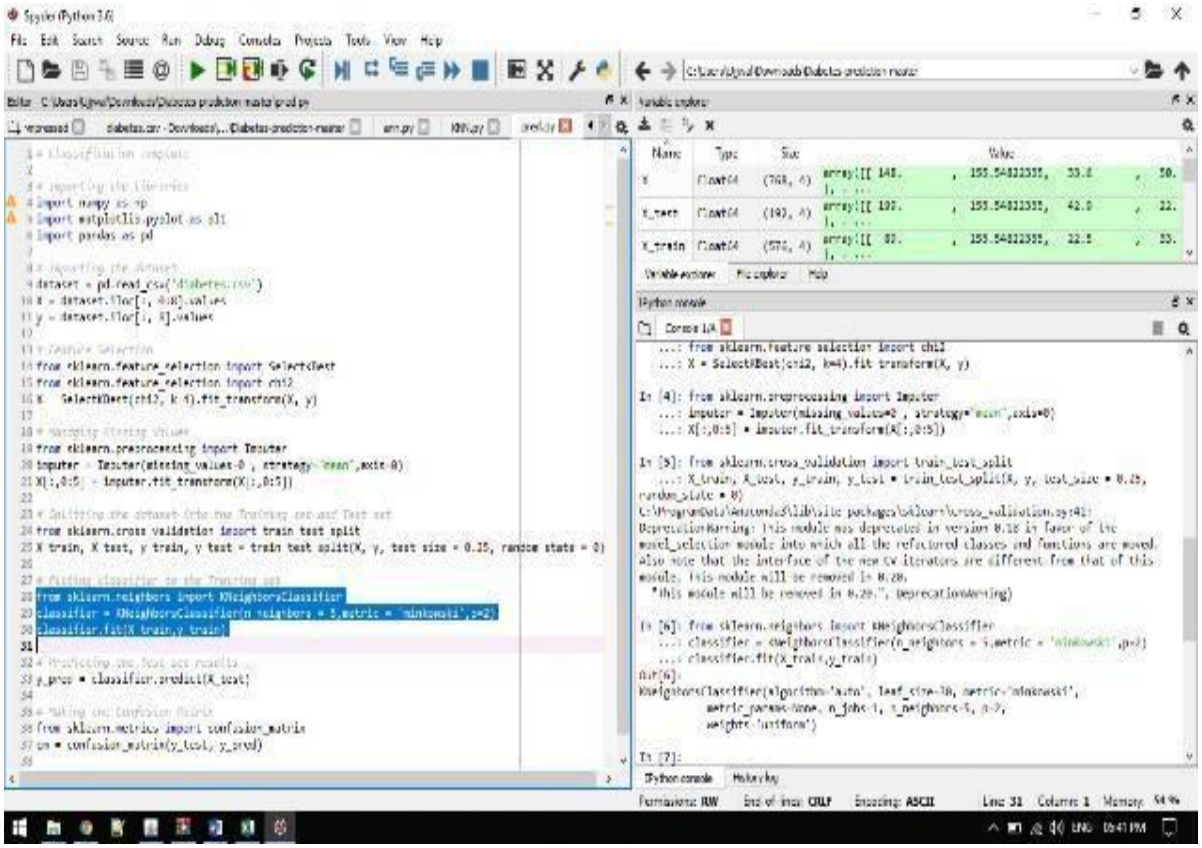


Fig: 6. *Fitting classifier to the dataset*

Page 16-25 © MAT Journals 2018. All Rights Reserved

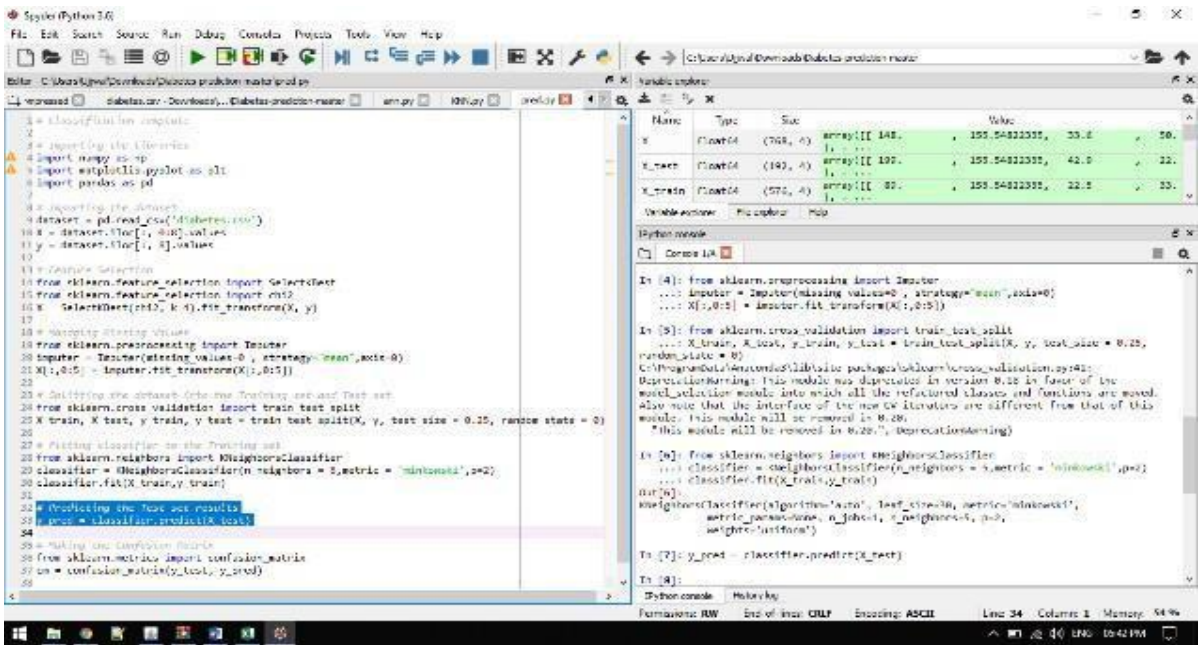


Fig: 7. *Predicting the test set*

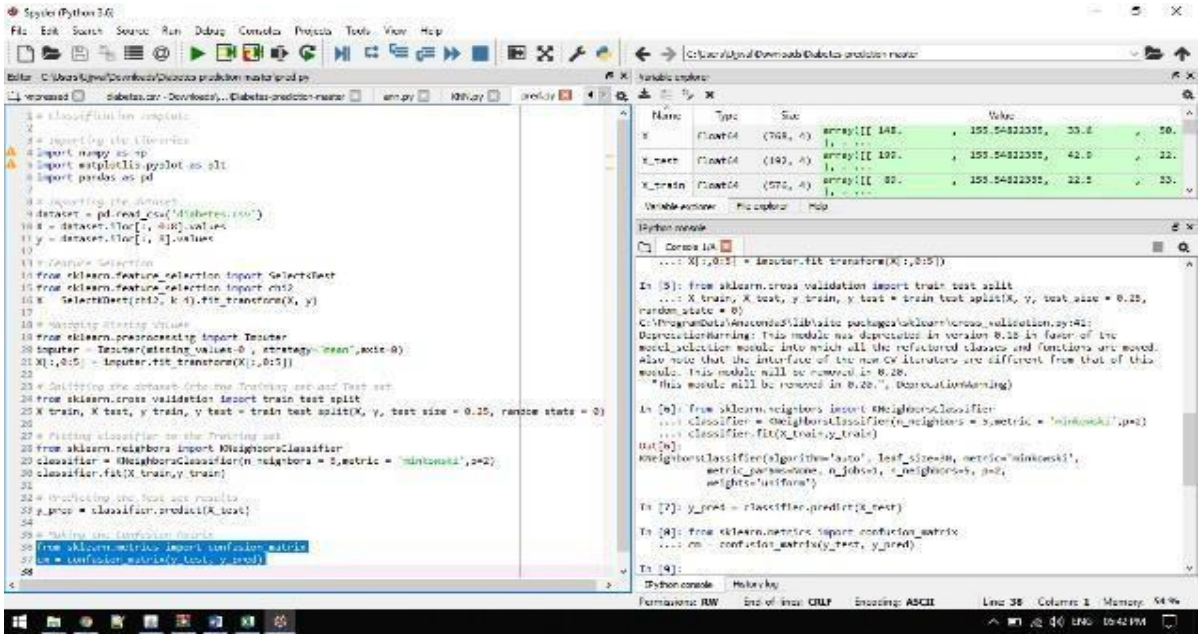


Fig: 8. *Making the confusion matrix*

CONCLUSION

The conventional method for detecting diabetes is time consuming, a clinical doctor must take many factors into consideration. The system we have developed uses a ML model to correctly determine health status of a person. So far Random Forest model proved to be the best amongst all in terms of accuracy in our type 2 diabetes data. Random forest model uses rule based system to tell whether a person is diabetic or not. The model can not only serve as a guide for doctors specializing in diabetes but also help family practitioners and interns in prescribing medications. It should be noted that diabetes cannot be predicted in a single event. Readings are taken for one

Page 16-25 © MAT Journals 2018. All Rights Reserved

week before any conclusion can be made from it. Our model serves the preliminary step of detecting diabetes. After that a check-up is recommended from doctor.

Future Scope

Predicting disease using AI is soon going to reduce the workload of clinical doctors. It is still under research that how accurate the results will be and continuous efforts are made to make intelligent systems. So far neural networks have outperformed very model in various domain in terms of accuracy but when it comes to predicting disease, neural networks do not perform better. Research is still going on how to make neural networks more accurate for predicting disease. Neural Networks provide a wide range of possibilities which in future may be efficient enough to predict diseases too.

REFERENCES

1. HLA Nomenclature in WMDA file format–details of current HLA alleles and where known their unambiguous, possible or assumed serologically equivalent antigens. http://hla.alleles.org/wmda/rel_dna_ser .txt. Accessed April 13, 20164
2. International Diabetes Federation. Diabetes Atlas. 5th ed. Brussels, Belgium: IDF Publications. (2011) The Global Burden of Diabetes; pp. 7–13. Available from http://www.idf.org/diabetesatlas/news/ fifth-edition-release. Accessed 25 May2015
3. Georgiev, D., Houdová, L., Fetter, M., Jindra, P.: A Scalable Method for Efficient Stem Cell Donor HLA Genotype Match Determination. In: Energy, Environment, Biology and Biomedicine, Proceedings of the 2014 International Conference on Biology and Biomedicine II, pp. 28-32. INASE, Prague (2014)3
4. Hayuhardhika W, Putra N, Sugiyanto, Sarno R, Sidiq M (2013) Weighted Ontology and Weighted Tree Similarity Algorithm for Diagnosing Diabetes Mellitus. IEEE International Conference on Computer, Control, Informatics and Its Applications pp.267-272.
5. Chen R, Huang Y, Bau C, Chen S (2012) A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Systems with Applications 39: 3995–4006
6. NMDP allele codes. https://bioinformatics.bethematchclinical.org/hla-resources/ allele- codes/. Accessed April 13,2016
7. A. S. Abdelmoneim, D. T. Eurich, J.-M. Gamble, and S. H. Simpson, "Use patterns of antidiabetic regimens by patients with type 2 diabetes," Canadian Journal of Diabetes, vol. 37, no. 6, pp. 394–400, 2013.
8. World Marrow Donor Association International Standards for Unrelated Hematopoietic Stem Cell Donor registries (2014), https://www.wmda.info/images/ pdf/20140101-STDC-WMDA_Standards_New_Housestyle.pdf. Accessed April 7,2016
9. Y. Handelsman, Z. T. Bloomgarden, G. Grunberger et al., "American Association of Clinical Endocrinologists and American College of Endocrinology - clinical practice guidelines for developing a diabetes mellitus comprehensive care plan - 2015," Endocrine Practice, vol. 21, Supplement 1, pp. 1–87,2015

practice guidelines for developing a diabetes mellitus comprehensive care plan - 2015," Endocrine Practice, vol. 21, Supplement 1, pp. 1–87,2015

Page 16-25 © MAT Journals 2018. All Rights Reserved

Page 16-25 © MAT Journals 2018. All Rights Reserved