

## **Diabetes Prediction(statistical modelling)**

### Introduction

Artificial Intelligence is intelligence demonstrated by machines, in contrast to the natural intelligence possessed by humans. It is used to describe machines that mimic cognitive functions that humans associate with the human mind such as learning and problem solving. It aims to mimic human cognitive functions. Artificial Intelligence has various applications in today's society. It is transforming the way we interact, consume information, and obtain goods and services across industries. In health care, AI is bringing paradigm changes. It is already changing the patient experience, how clinicians practice medicine, and how the pharmaceutical industry operates. Recently AI techniques have sent vast waves across healthcare, even fuelling an active discussion of whether AI doctors will eventually replace human physicians in the future. They will definitely help physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare. The increasing availability of healthcare data and rapid development of big data analytic methods has made possible the recent successful applications of AI in healthcare. Guided by relevant clinical questions, powerful AI techniques can unlock clinically relevant information hidden in the massive amount of data, which in turn can assist clinical decision making.

### Aim and Scope

Classification is a supervised Machine Learning technique which assigns labels or classes to not the same objects or groups. Classification strategies are broadly used in the medical field for classifying data into different classes according to certain criteria. Classification is a two-step process: First step is model construction which is defined as the analysis of the training records of a database. Second step is model usage to the constructed model is used for classification. The classification accuracy is estimated by the percentage of test samples or records that are correctly classified. In the Classification has been successfully applied to a wide range of application areas, such as scientific experiments, medical diagnosis, weather prediction, credit approval, customer segmentation, target marketing and fraud detection. Classifiers are used extensively for diagnosis of breast tumor in ultrasonic images, ovarian cancer, heart sound diagnosis, diabetes prediction and so on.

Diabetes affects over 415M people around the world. It is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. Many

complications occur if diabetes remains untreated and the only way to stay away from complications is early detection. This is when Machine Learning Algorithms come into use. This project demonstrates Diabetes Prediction using Classification algorithms. The aim here is proactive analysis and detection of diabetes in order to delay, prevent or control its onset to prevent detrimental health effects. In this model, predictions have been made based on pregnancies, blood pressure, skin thickness, bmi, age, insulin level, etc.

### Random Forest Algorithm

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Step 1 – First, start with the selection of random samples from a given dataset.

Step 2 – Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3 – In this step, voting will be performed for every predicted result.

Step 4 – At last, select the most voted prediction result as the final prediction result.

### Advantages

It overcomes the problem of overfitting by averaging or combining the results of different decision trees.

Random forests work well for a large range of data items than a single decision tree does.

Random forest has less variance than single decision tree.

Random forests are very flexible and possess very high accuracy.

Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling.

Random Forest algorithms maintain good accuracy even a large proportion of the data is missing.

### Disadvantages:

Complexity is the main disadvantage of Random forest algorithms.

Construction of Random forests are much harder and time-consuming than decision trees.

More computational resources are required to implement Random Forest algorithm.

It is less intuitive in case when we have a large collection of decision trees.

The prediction process using random forests is very time-consuming in comparison with other algorithms.

### Steps:

1.import libraries

2.loading dataset(data used: <https://www.kaggle.com/johndasilva/diabetes>)

3.exploring dataset

4.data cleaning

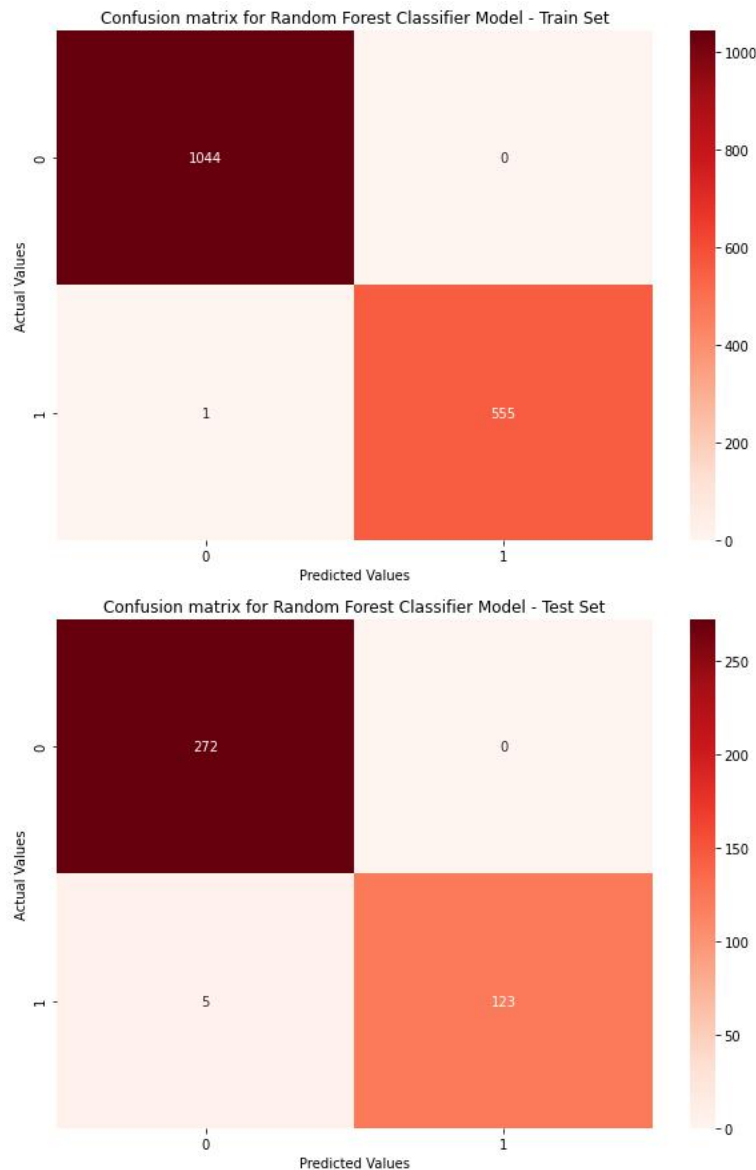
5.model building(setting X and y,splitting dataset into train and test set,finding out best model and parameters and fitting them)

6.model evaluation

7.data visualisation

8.predictions

Output: The model will predict if an individual has diabetes or not.



Accuracy on test set:98.75%  
Accuracy on training set:99.94%

## Related work

AI aims to make healthcare more proactive and consumer centric. Diabetic Retinopathy is the fastest growing cause of blindness. It is diagnosed with the help of Retinal Fundus Images by ophthalmologists. The primary issue is the grading of retinal images by ophthalmologists or trained persons, whose numbers are very scarce compared to the load of patients requiring screening. This can be solved with provision of an automated imaging system within easy reach of the patient. Deep NN are used to read these images and rate them on a scale of 1-5 i.e. from No DR to Proliferative DR. According to JAMA, Deep Learning Algorithm for Detection of Diabetic Retinopathy has an F-Score of 0.95.

In 2016, its accuracy was on par with that of general ophthalmologists and in 2017, it was on par with retinal specialist ophthalmologists.

## Conclusion

AI application is increasing sharply, not only in the general sectors but also in healthcare. Many different interdisciplinary pieces of research with different AIs and its functions suggest the fundamental and necessary gaps of AI support patients. AI-powered solutions have made small steps towards addressing key issues, but still have yet to achieve a meaningful overall impact on the global healthcare industry, despite the substantial media attention surrounding it. If several key challenges can be addressed in the coming years, it could play a leading role in how healthcare systems of the future operate, augmenting clinical resources and ensuring optimal patient outcomes.