Prognostication of Diabetes using Random Forest

Harsh Harwani
Undergraduate Student
Thadomal Shahani Engineering College, Mumbai, India

Mohammed Omar Khan
Undergraduate Student
B.M.S. College of Engineering, Bangalore, India

Ananya Arora
Undergraduate Student
Thadomal Shahani Engineering College, Mumbai, India

ABSTRACT
Diabetes is a serious malady where one has abnormally high blood sugar levels. Despite being so deadly, it is quite common as anyone is susceptible to it. If untreated, it can damage a person's kidneys, eyes, nerves, and other organs. Genes, environment, and preexisting medical conditions can all affect a person's odds of developing diabetes. The bottom line is that it can be extremely deadly if discovered late. Thus, it is imperative that researchers devise an accurate diabetes predictor in order to enable early treatment of diabetic people. This paper demonstrates the prediction of diabetes using the Random Forest algorithm on the PIMA Indians Diabetes dataset. Using important data points and features of several healthy and diabetic PIMA Indians, the model predicts the onset of diabetes. The performance of this algorithm is evaluated using metrics like Accuracy, Precision, and Recall. Furthermore, several suggestions to improve the effectiveness of this model are discussed.

Keywords
Machine learning, Diabetes prediction, Random Forest, PIMA Indians diabetes dataset

1. INTRODUCTION
Diabetes, commonly known as the 'Silent Killer' [1], can be caused by two factors mainly. First, diabetes can be caused by unhealthy eating habits and eating foods that are rich in sugar. It can further be exacerbated by poor lifestyle choices like ignoring physical activity. Second, genetics can too lead to the onset of diabetes. It is predicted that by 2045, the number of diabetic patients would reach around 700 million, based on statistics calculated by the International Diabetes Federation [2]. There are two types of diabetes, and their differences and severity require them to have slightly different approaches to their treatment. Glucose is what gives energy to the human body, but for this to enter the cells, insulin is required. Without insulin, the body is vulnerable to several fatal side effects. Thus, insulin is fundamental for the healthy functioning of the body. Type 1 diabetes is where the pancreas gets damaged by antibodies and is unable to produce any insulin. Type 1 diabetes usually begins at an early age, and genetics are the major cause of this. People with this type of diabetes experience symptoms that include mood swings, weight loss, and restlessness. The symptoms of this type develop faster, over the span of a few weeks. Even though this type usually begins during childhood, its possibility of coming up during adulthood still exists. Type 2 diabetes, on the other hand, is where the pancreas produces insulin but in insufficient quantities for the body. It is also possible that the body doesn't use the insulin or respond to it like it normally should. This type of diabetes is more common. It is comparatively milder than Type 1, but it mustn't be taken lightly either as it raises the risk of heart diseases. The early recognition of diabetes before its onset is very important for helping people who are developing it. The U.S. Department of Health and Human Services [3] explains that the early detection of diabetes and pre-diabetes is important so that patients can begin to control the disease early and avert or delay the pernicious complications that can decrease quality of life. Thus, this necessitates the importance of a good model that can predict the onset of this deadly disease through several parameters. Using the concepts of machine learning, one can train a model to accurately predict if a person has diabetes or not. For training the model, Random Forest classifier is used. It is an algorithm that requires very little pre-processing and the data need not be rescaled or transformed. It works great with high dimensional data and allows for quick predictions due to fast training speeds. It is robust and is able to handle unbalanced data with ease. With these advantages in mind, this paper attempts to accurately predict the onset of diabetes and contribute to the research being done in the domain of disease prediction. This paper is organized as follows. Section 2 discusses related research in the diabetes prediction domain. In order to approach a standard context, Section 3 describes the dataset used and the Random Forest classifier. Section 4 presents an overview of the implementation of the technique. Section 5 mentions the results of the experiment. The results are further analysed in section 6. Section 7 discusses the conclusions and proposes possible avenues that might be undertaken to improve the models implemented in this paper.

2. LITERATURE REVIEW
Various studies have been carried out to accurately predict diabetes. This disease has been thoroughly researched by scholars. Various machine learning algorithms have been used to forecast diabetes in patients. For instance, Quan et al. [4] selected 68994 healthy people and diabetic patients' data as a training set. They leveraged techniques like Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) to lower the dimensionality. Through this study, Quan et al. [4] realized that prediction with the Random Forest resulted in the highest accuracy of 0.8084. Furthermore, Sneha and Gangil's [5] research had the objective of finding an optimal classifier that gave a predictive result similar to clinical outcomes. The method they proposed, used significant features that aided in the early detection of diabetes mellitus. Their research showed that the Decision Tree algorithm and Random Forest algorithm had the highest specificity of 98.2% and 98%, while Naive Bayesian outcomes had the best accuracy of 82.3%. They also generalized the selection of optimal features

from the dataset to improve classification accuracy. A literature review has shown that there have been many studies for the comparison of machine learning classifiers on diabetes datasets. For example, Alam et al. employed ANN, Random Forest, and K-means clustering to predict diabetes on the Pima Indian dataset [6]. Additionally, in the same context, Ratna and Tamane [7] compared the performance of eight classifiers including Logistic Regression, KNN, SVM,

Gradient Boost, Decision tree, MLP, Random Forest, and Gaussian Naive while predicting the population who are most likely to develop diabetes. They used Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Receiver Operating Characteristic (ROC), Accuracy, Precision, and Recall as performance measurement parameters. Moreover, an exhaustive exploration of machine learning techniques for diabetes prediction was proposed by Wei et al. [8], who predicted diabetes using different algorithms.

3. DATASET DESCRIPTION

This paper uses the popular Pima Indians diabetes dataset. This dataset contains a total of 768 entries and eight medical predictor attributes. The multiple features included are the number of pregnancies, glucose levels, insulin levels, etc. All the attributes along with their descriptions are stated in Table 1 below.

Table 1. Features of the PIMA Indians Diabetes dataset

FEATURES	DESCRIPTION	DATATYPE
AGE	Age (in years)	numeric
PREGNANCIES	Number of times pregnant	numeric
GLUCOSE	Plasma Glucose concentration (Glucose Tolerance Test)	numeric
BLOOD PRESSURE	Diastolic Blood Pressure (mm Hg)	numeric
SKIN THICKNESS	Triceps Skin Fold Thickness (mm)	numeric
INSULIN	2-Hour serum insulin (mu U/ml)	numeric
BMI	Body Mass Index (weight[kg]/height[m])	numeric
DIABETES PEDIGREE	Diabetes Pedigree Function	numeric
OUTCOME	Class variable (0 or 1)	binary

This dataset can be used to predict whether a patient has diabetes. There is one target variable called 'Outcome' that is of a binary data type i.e. there are two outcomes only. '1' indicates that the patient is diabetic while '0' indicates the opposite. The dataset has an uneven distribution of the outcomes. Out of the 768 entries, 500 entries are of healthy patients and 268 entries are of diabetic patients. Additionally, there are some missing values for some of the attributes. Thus, this dataset is not perfect and needs to be cleaned and pre-processed.

4. METHODOLOGY

Random Forest is an accumulation of decision trees where the outcomes of each tree are aggregated into the concluding result. Random Forest is conventionally trained by the 'Bootstrap algorithm' where a subset of features is selected randomly from the main set and the 'Bagging' method where the normal paradigm of the bagging method is that the combination of different learning modes will increase the overall accuracy and result of the model being trained. As the number of trees in the Random Forest algorithm increases, so does the accuracy of the model. Random Forest is a supervised algorithm that is popularly used for classification problems due to the very high accuracy depicted by it. This makes Random Forest a perfect choice for a high accuracy model to predict diabetes. Random Forest works efficiently on large databases making it optimal for use in the field of disease prediction where ample data is available. It provides an estimate of the variables that are comparatively more important than the others. Due to this, it can also maintain high accuracy even when large amounts of data are missing. Its fast training rate and parallelizable training allow us to implement the model faster. In this experiment, 70% of the data was used as training data and 30% as testing data.

4.1. Algorithm of Random Forest

Step 1: The first step in the Random Forest algorithm is to select random samples (K) from the dataset. These samples are chosen from the training dataset.

Step 2: The second step comprises constructing a decision tree for all the selected samples. Then a corresponding prediction based on the decision tree made for each sample is given.

Step 3: In the third step of Random Forest, voting is done for each of the predicted results.

Step 4: In the final step of Random Forest, the most voted predicted result is selected as the final predicted result of the algorithm.

5. RESULTS

A confusion matrix aids the evaluation of the performance of a classification algorithm or a classifier. The calculations of metrics like Accuracy, Precision, and F1-score can be easily done using the confusion matrix and the results help in identifying the errors in the classifier. Hence, the confusion matrix is also known as the error matrix. Table 2 shows the structure of the confusion matrix.

Table 2. Structure of confusion matrix

		Predicted Class	
		P	**N**
Actual Class	**P**	TP	FN
	N	FP	TN

* P = the number of real positive cases in data
* N = the number of real negative cases in data
* TP = True Positive; TN = True Negative
* FP = False Positive; FN = False Negative

Table 3 is the confusion matrix obtained by using Random Forest on the PIMA Indians Diabetes dataset.

Table 3: Confusion matrix obtained by using Random Forest

		Predicted Class	
		P	**N**
Actual Class	**P**	50	30
	N	22	128

The confusion matrix of this model shows that out of the 230 testing data, the model accurately predicts 50 cases as positive (diabetic) and 128 cases as negative (non-diabetic). On the contrary, 22 cases are incorrectly predicted as positive (diabetic) and 30 cases are erroneously predicted as negative (non-diabetic). Using the confusion matrix, the following metrics were calculated. Table 4 and Figure 1 show the values of the metrics of the model. Looking at these parameters, one can judge how efficient the model is.

Table 4. Calculated metrics of the model

PARAMETERS	VALUE
ACCURACY	77.39%
SENSITIVITY	85.33%
SPECIFICITY	62.50%
PRECISION	81.01%
NEGATIVE PREDICTIVE VALUE	69.44%
F1-SCORE	65.79%

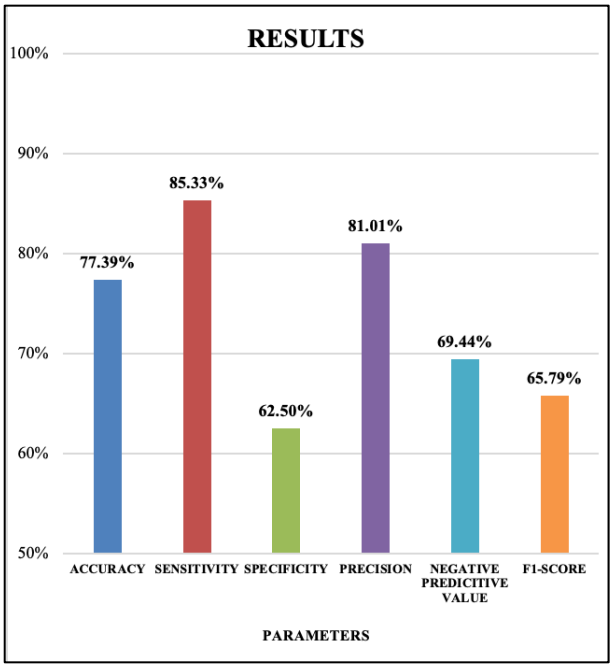

Fig 1: Calculated metrics of the model

6. DISCUSSION OF RESULTS

Accuracy is the fraction of predictions that the model gets right. It is the measure of the proportion of right predictions to the total outcomes. The Accuracy of this model is 77.39%. This signifies that the model accurately predicts approximately 77% of the outcomes. Specificity is defined as the ratio of the actual negatives which got predicted as negative. The Specificity is 62.5%. This reveals that this model predicts 65% of the actual negative outcomes as negative. Precision is also called Positive Predictive Value. It is the ratio of true positive data to the total positive data. The higher the precision, the more relevant instances are being extracted by the model. The Precision of the model is 81.01%, which means that approximately 81% of the actual positive outcome was accurately predicted. Negative Predictive Value (NPV) is the number of true negatives out of the sum of true negatives and false negatives. This model's NPV is 69.44%. F1-score signifies the balance between the precision and the recall of the model. It is the harmonic mean of both the metrics. The model's F1-score is 65.79%. The higher the F1-score, the better the model is.

7. CONCLUSION

Diabetes is a prevalent disease that affects millions of people throughout the world. Since anyone is susceptible to diabetes, early recognition and prevention of it plays a pivotal role. Extensive research has been done on diabetes datasets among which the PIMA Indian diabetes dataset consisting of 768 records is the most pervasive dataset used. This study has made an effort to implement a high accuracy data mining algorithm i.e. Random Forest, which was applied to the dataset chosen. The dataset used was pre-processed and cleaned before training the model which allowed for faster training time and enhanced results. The model trained was able to achieve an Accuracy of 77.39%. Additionally, the model achieved a Precision of 81.01%.

8. SCOPE

To further improve the Accuracy and other performance metrics of this model, the Random Forest algorithm can be complemented with several other algorithms. For instance, Panambar et al. [9], have constructed a hybrid model of Decision Tree and Random Forest algorithms. This combination enhances the performance of each individual algorithm and can be further refined to give a highly accurate predictor. Furthermore, this model can be improved by having a larger dataset where the data is of high quality and has fewer missing values. More parameters can be considered in the future that play a cardinal role in diabetes prediction. Alongside predicting diabetes, predicting diabetes-related problems is also a prime aspect of dealing with diabetes. Hence, better quality datasets and studies of diabetes-related problems can better guide doctors to treat patients in an effective manner.

9. REFERENCES

[1] "Diabetes-A Silent Killer", *Radixhealthcare.org*, 2020. [Online]. Available: https://radixhealthcare.org/blog/5e1ece326e6197396a35cd46/diabetes. [Accessed: 02- Oct- 2020].

[2] A. diabetes, W. diabetes and F. figures, "International Diabetes Federation - Facts & figures", *Idf.org*, 2020. [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html. [Accessed: 02- Oct- 2020]. "The Importance of Early Diabetes Detection", *ASPE*, 2020.

[Online]. Available: https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection. [Accessed: 02- Oct- 2020].

[3] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", *Frontiers in Genetics*, vol. 9, 2018. Available: 10.3389/fgene.2018.00515 [Accessed 2 October 2020].

[4] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection", *Journal of Big Data*, vol. 6, no. 1, 2019. Available: 10.1186/s40537-019-0175-6 [Accessed 2 October 2020].

[5] T. Mahboob Alam et al., "A model for early prediction of diabetes", *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019. Available: 10.1016/j.imu.2019.100204.

[6] R. Patil and S. Tamane, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, p. 3966, 2018. Available: 10.11591/ijece.v8i5.pp3966-3975.

[7] S. Wei, X. Zhao and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification", *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, 2018. Available: 10.1109/wf-iot.2018.8355130 [Accessed 2 October 2020].

[8] A. P, M. M V and S. H A, "DRAP: Decision Tree and Random Forest Based Classification Model to Predict Diabetes", *2019 1st International Conference on Advances in Information Technology (ICAIT)*, 2019. Available: 10.1109/icait47043.2019.8987277 [Accessed 2 October 2020].