# Prognosticate Diabetes Mellitus Using Machine Learning

##### **K.Thirunavukkarasu.,M.E**

Assistant Professor ,Thanthai Periyar Government Institute of Technology,Vellore-02.

**ABSTRACT**

Diabetes Mellitus often recognized as diabetes is a metabolic disease that leads to high blood sugar. It has reached an epidemic magnitude in many countries and more so in developing countries. In this project, our task is to classify whether a patient is diabetic or not supported by differing kinds of diabetes.

We have implemented different Machine Learning Algorithms for classification such as Decision Tree, Random Forest, Support Vector Machine, Kernel Support Vector Machine and obtained accuracy from each model using different performance metrics. The model with the best accuracy is being used for classification.

**INDEX TERMS**

Diabetes Mellitus, Exploratory Data Analysis, Principal Component Analysis, t-Distributed Stochastic Neighbor-hood Embedding, Decision Tree, Random Forest, SVM, Kernel SVM, Performance metrics.

**INTRODUCTION**

Diabetes mellitus is a union of metabolic diseases distinguished by hyperglycemia that results from the faults in insulin secretion, insulin action, or both. **D Atlas - … , 7th edn. Brussels, Belgium: International Diabetes …, 2015 - suckhoenoitiet.vn**The chronic hyperglycemia of diabetes is related along with the long-term injuries, dysfunction, and loss of different organs, particularly the eyes, kidneys, nerves, heart, and blood vessels. Various pathogenic procedures are included in the growth of diabetes. **D Mellitus - Diabetes care, 2006 - Am Diabetes Assoc**. Those range by autoimmune destruction of the β-cells of the pancreas with subsequent insulin scarcity to abnormalities resulting in confrontation to insulin action. The origin of the irregularities in carbohydrate, fat, and protein metabolic processes in diabetes is inadequate action of insulin on target tissues.**National Diabetes Data Group (US)… - 1995 - books.google.com**

Inadequate insulin effect results from deficient insulin production and decreased tissue react to insulin at one or more locations in the complex trackways of hormone activity.**G Williams, JC Pickup - 2004 - proformas.ljmu.ac.uk**. Disability of insulin production and flaws in insulin action often are associated with the same patient, and it is frequently uncertain which is a peculiarity, likewise, is the main source of the hyperglycemia.

[**AH Mokdad**](https://scholar.google.co.in/citations?user=L5xeTxkAAAAJ&hl=en&oi=sra)**, ES Ford, BA Bowman, DE Nelson… - Diabetes …, 2000 - Am Diabetes Assoc**. Symptoms of obvious hyperglycemia contain polyuria, polydipsia, weight loss, occasionally with polyphagia, and blurred vision. Disability of increase and susceptibility to definite infections may also accompany chronic hyperglycemia. Severe, life-threatening outcomes of non-controlled diabetes are hyperglycemia with ketoacidosis or the nonketotic hyperosmolar syndrome.

Long-lasting problems of diabetes involve retinopathy with the potential vision loss; nephropathy principal to renal loss; peripheral neuropathy with fear of foot ulcers, amputations, and Charcot's joints; and sovereign neuropathy producing, genitourinary, gastrointestinal and cardiovascular symptoms and sexual dysfunction**.World Health Organization - 2016 - apps.who.int**. Patients escorted by diabetes have growth in the incidence of atherosclerotic cardiovascular, peripheral arterial, and cerebrovascular disease.

Hypertension and irregularities of lipoprotein metabolism are frequently found in individuals with diabetes. The enormous majority of cases of diabetes come under two broad etiopathogenetic species (discussed in greater detail below). In the first category, type 1 diabetes, the source is a certain inadequacy of insulin production**.MI Harris - Diabetes Research and Clinical …, 1995 - diabetes research clinical practice …** Individuals at enlarged risk of growing this category of diabetes can frequently be recognized by the serological confirmation of an autoimmune pathologic process happening in the pancreatic islets and by genetic markers. **KD Crothall, B Butoi-Teodorescu… - US Patent 7,179,226, 2007 - Google Patents**

Whereas in the second category, much more prevalent kind, type 2 diabetes, the source is a combination of resistance to insulin action and an insufficient compensatory insulin secretory reaction. In the next category, a level of hyperglycemia enough to source path-logic and functional changes in numerous target tissues, but without clinical symptoms, conceivably present for a whole of time before diabetes is spotted.**HM Mather, JA Nisbet, GH Burton,** [**GJ Poston**](https://scholar.google.co.in/citations?user=oCBdzTkAAAAJ&hl=en&oi=sra)**… - Clinica Chimica …, 1979 - Elsevier.** Throughout this asymptomatic phase, it is feasible to illustrate an abnormality in carbohydrate metabolism by evaluation of plasma glucose in the fasting state or following a challenge with an oral glucose load.

The degree of hyperglycemia if present, may switch over time, depending on the scope of the essential disease procedure**.A Ramachandran,** [**RCW Ma**](https://scholar.google.co.in/citations?user=Y1W7OtoAAAAJ&hl=en&oi=sra)**, C Snehalatha - The Lancet, 2010 - Elsevier** A disease procedure may be existent but may not have developed far enough to produce hyperglycemia. The alike disease procedure can base impaired fasting glucose (IFG) and impaired glucose tolerance (IGT) without realizing the scale for the diagnosis of diabetes. In a few people along with diabetes, adequate glycemic management can be achieved with weight reduction, exercise, and oral glucose-lowering agents. These individuals accordingly do not need insulin**.**[**MG Kahn**](https://scholar.google.co.in/citations?user=UwLIDg4AAAAJ&hl=en&oi=sra)**, D Huang, SA Bussmann, SB Cousins… - US Patent …, 1993 - Google Patents**

Others who have some excess insulin secretion but need exogenous insulin for appropriate glycemic control can last without it**.**[**PZ Zimmet**](https://scholar.google.co.in/citations?user=d-NvY4cAAAAJ&hl=en&oi=sra)**,** [**DJ Magliano**](https://scholar.google.co.in/citations?user=Ylakn5IAAAAJ&hl=en&oi=sra)**, WH Herman… - The lancet Diabetes & …, 2014 - Elsevier** Individuals with extensive β-cell levelling and hence no residual insulin secretion need insulin for survival. The extremity of the metabolic irregularity can progress, regress, or continue the same. Accordingly, the degree of hyperglycemia gives back the predicamentof the first principles, metabolic process and its treatment more than the nature of the process by oneself.

**RESEARCH ELABORATION**

Machine Learning Algorithms are universally used for various Diagnosis. The Data has been gathered from the National Institute for Diabetes and Metabolism. The project was implemented on Two types of Diabetes that are Type 1 and Type 2 to predict whether the patient is Diabetic or NonDiabetic.

Some of the features have been collected from**Aishwarya R., Gayathri P., Jaisankar N.**

**A Method for Classification Using Machine Learning Technique for Diabetes**

**International Journal of Engineering and Technology (IJET), 5 (2013), pp. 2903-2908**

In the above mentioned Journal, there were relevant features that can be used to predict Whether the Patient is Diabetic or Nondiabetic.

The above-mentioned journal used several preprocessing and Data Cleaning techniques to feed the data to the model.

Various Machine Learning Algorithms were used to train the model for predicting the output.

**Aljumah A.A., Ahamad M.G., Siddiqui M.K.**

**Application of data mining: Diabetes health care in young and old patients**

**Journal of King Saud University-Computer and Information Sciences, 25 (2013), pp. 127-136.**

The above Journal has been found useful for Preprocessing and cleaning techniques for building the model.

**Bamnote M.P.G.R.**

**Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming**

**Advances in Intelligent Systems and Computing, 1 (2014), pp. 763-770**

The feature Extraction techniques EDA and Dimensionality Reduction were used from the above journal for reference.

Various EDA techniques Like Pair Plots, PDF, Box Plots, Violin Plots etc were used.

**Smith, L. I. (2002). A tutorial on principal components analysis. *Inform. Fusion* 51:52.**

For Dimensionality Reduction PCA ([Wang and Paliwal, 2003](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B50); [Polat and Günes, 2007](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B33); [You et al., 2018](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B55)) and t-SNE(**Visualizing Data using t-SNE *Laurens van der Maaten, Geoffrey Hinton*; 9(86):2579−2605, 2008.)** were used to extract the best features for building the model.

**Esposito F., Malerba D., Semeraro G., Kay J.**

**A comparative analysis of methods for pruning decision trees**

**IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (1997), pp. 476-491.**

Various Machine Learning Algorithms have been implemented such as Decision Tree( ([Salzberg, 1994](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B42); [Kohabi, 1996](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B20)), Random Forest([Friedl and Brodley, 1997](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B8); [Habibi et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B10); [Liao et al., 2018](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B26)), [Ozcivit And Gluten (2011)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B31)), SVM and Kernel-SVM([Kavakiotis et al., 2017](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B18)) for predicting whether the patient is Diabetic or Non-Diabetic. The accuracy has been computed for each model and best among the algorithms were picked.

Along with the Algorithms implemented in the above journal various other Algorithms were implemented and calculated the accuracy for each model using various production metrics and picked the best among them.

**METHODOLOGY**

The project has been performed on two types of diabetes Type 1 Diabetes and Type 2 Diabetes. The data for both type 1 diabetes and type 2 Diabetes has been abstracted from the National Institute for Diabetes and Metabolism.

Type 1 Diabetes Implementation

Generally Type 1 Diabetes will be seen in younger people whose age is less than or equal to 30 years. They will be having frequent urination, high blood glucose levels etc. ([Iancu et al., 2008](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B12))

The Data has been abstracted from the National Institute for Diabetes and Metabolism. The Raw data has been converted into a CSV file to feed the data into the model. During the implementation there were some duplicate values, Missing Values and data has not been standardized.

Various techniques are implemented for Data Cleaning and Data Preprocessing. Duplicates are removed using Data Deduplication, Missing Values are removed Using imputer method from scikit learn Library, and the data has been standardized using Standard Scalar method present in the scikit learn library.

[**Data preprocessing techniques** for classification without discrimination](https://link.springer.com/article/10.1007/s10115-011-0463-8)[F Kamiran](https://scholar.google.co.in/citations?user=yfyugf8AAAAJ&hl=en&oi=sra), [T Calders](https://scholar.google.co.in/citations?user=CcqxbMkAAAAJ&hl=en&oi=sra) - Knowledge and Information Systems, 2012 - Springer

There were more than 10 features present in the dataset in order to pick the best features from all the remaining features EDA (Exploratory Data Analysis) was implemented. The pair plots are being plotted to pick the best pair of features. But using the pair plots, we were inadequate to obtain the best pair of features.

# Exploratory Data Analysis by John Wilder Tukey. 16 June 1915 – 26 July 2000

# <https://doi.org/10.1098/rsbm.2003.0032>

But the features with the identical values on X-Axis and Y-Axis are far better than Features with different values X-Axis and Y-Axis. The features that we found relevant are Pregnancy, Glucose, Age.

Using these three features the PDF Graphs are being plotted, from the above PDF Graphs, the threshold value has been calculated for all three features.

PDF For Pregnancies



PDF For Glucose



PDF For Age



The Threshold values are being placed on the box plot and calculated the error using percentiles and quantiles. For all the above three features the error is more than 25%.

Box Plot For Pregnancies



Error for Pregnancies : 55 % for Class 1 and 25 % for Class 0.

Box Plot For Glucose



Error for Glucose : 26 % for Class 1 and 30 % for Class 0.

Box Plot For Age



Error for Age : 30 % for Class 1 and 40 % for Class 0.

After calculating the error the features are found to be irrelevant for building the model.

As we were unable to obtain the best features using EDA(Exploratory Data Analysis) Dimensionality Reduction using PCA(Principal Component Analysis)

[Polat and Günes (2007)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B33) distinguished diabetes from normal people by using principal component analysis (PCA) and neuro and t-SNE(t-Stochastic Neighbour-hood Embedding) was used for finding the best features.

**Using PCA Features on Different Machine Learning Algorithms**

Smith, L. I. (2002). A tutorial on principal components analysis. *Inform. Fusion* 51:52.

In PCA ([Wang and Paliwal, 2003](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B50); [Polat and Günes, 2007](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B33); [You et al., 2018](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B55)) the k vectors and unit eigenvectors are obtained by solving the correlation matrix of the observed features. Eigenvalues and EigenVectors are sorted in ascending order ([Smith, 2002](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B44)).

Using PCA (Principal Component Analysis) we were able to detect the eigenvalues and eigenvectors for each feature. After Calculating the eigenvalues and eigenvectors the top two values with the highest eigenvalues and eigenvectors were picked and using these features the model has been trained to predict the output.

Different Machine Learning Algorithms were used such as ([Kavakiotis et al., 2017](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B18)) support vector machine (SVM), decision tree (DT) Decision tree ([Friedl and Brodley, 1997](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B8); [Habibi et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B10); [Liao et al., 2018](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B26)), [Ozcivit And Gluten (2011)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B31)proposed a new method of ensemble learning known as Random Forest ([Breiman, 2001](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B4); [Lin et al., 2014](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B28); [Svetnik et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B46)) and Kernel-SVM are implemented using the above features and the output is predicted based on the train data for each model. After predicting the output the accuracy has been calculated using different production metrics such as Confusion Matrix, HeatMap, Accuracy, Precision, Recall.

After Calculating the Accuracy, Using PCA features Kernel-SVM has better accuracy of 94.33 %.

**Using t-SNE Features on Different Machine Learning Algorithms**

**Visualizing Data using t-SNE *Laurens van der Maaten, Geoffrey Hinton*; 9(86):2579−2605, 2008.**

t-SNE which is commonly known as t-Stochastic Neighbour-hood Embedding is an Unsupervised , non-Linear based approach for feature extraction in higher dimensions.t-SNE gives the overall view on how the data is arranged in the higher dimensions.

t-SNE is different from PCA , In t-SNE we preserve the small pairwise distance but when we take PCA we will be preserving the Large pairwise distance. Generally the t-SNE algorithm calculates the pairwise distance in both higher and lower dimensions and optimises the two similarity measures using cost function.

### [How to use **t**-**SNE** effectively](https://distill.pub/2016/misread-tsne/?_ga=2.135835192.888864733.1531353600-1779571267.1531353600)

[M Wattenberg](https://scholar.google.co.in/citations?user=pv54dqMAAAAJ&hl=en&oi=sra), [F Viégas](https://scholar.google.co.in/citations?user=GvXDNsYAAAAJ&hl=en&oi=sra), [I Johnson](https://scholar.google.co.in/citations?user=lml0nEgAAAAJ&hl=en&oi=sra) - Distill, 2016 - distill.pub

Different Machine Learning Algorithms were used such as ([Kavakiotis et al., 2017](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B18)) support vector machine (SVM), decision tree (DT) Decision tree ([Friedl and Brodley, 1997](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B8); [Habibi et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B10); [Liao et al., 2018](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B26)), [Ozcivit And Gluten (2011)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B31)proposed a new method of ensemble learning known as Random Forest ([Breiman, 2001](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B4); [Lin et al., 2014](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B28); [Svetnik et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B46)) and Kernel-SVM are implemented using the above features and the output is predicted based on the train data for each model. After Predicting the output the accuracy has been calculated using different Performance Metrics such as Confusion Matrix, HeatMap, Accuracy, Precision, Recall.

After Calculating the accuracy, Using t-SNE features Random Forest has better accuracy of 96.77 %.

Random Forest Generates Multiple Decision Trees ([Pal, 2005](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B32)). In the Random Forest approach each Decision Tree will be predicting the output. After Predicting the output from each tree Majority Voting technique will be applied. ([Liaw and Wiener, 2002](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B27); [Svetnik et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B46)).

Type 2 Diabetes Implementation

Generally Type 2 Diabetes is seen in elderly people and causes obesity, hypertension and other ([Robertson et al., 2011](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B40)).

The Data has been taken from the National Institute for Diabetes and Metabolism. The Raw data has been converted into a CSV file to feed the data into the model. During the implementation there were some Duplicate values, Missing Values and data has not been standardized. Various techniques have been implemented for Data Cleaning and Data Preprocessing.

Duplicates are removed using Data Deduplication, Missing Values are removed Using imputer method from scikit learn Library, and the data has been standardized using Standard Scalar method present in the scikit learn library, The categorical data were converted into numerical data using simple Python.

[**Data preprocessing techniques** for classification without discrimination](https://link.springer.com/article/10.1007/s10115-011-0463-8)[F Kamiran](https://scholar.google.co.in/citations?user=yfyugf8AAAAJ&hl=en&oi=sra), [T Calders](https://scholar.google.co.in/citations?user=CcqxbMkAAAAJ&hl=en&oi=sra) - Knowledge and Information Systems, 2012 - Springer

There were more than 10 features in the dataset in order to pick the best features from all the remaining features Dimensionality Reduction using PCA and t-SNE was used.

**Using PCA Features on Different Machine Learning Algorithms**

Smith, L. I. (2002). A tutorial on principal components analysis. *Inform. Fusion* 51:52.

In PCA ([Wang and Paliwal, 2003](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B50); [Polat and Günes, 2007](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B33); [You et al., 2018](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B55)) the k vectors and unit eigenvectors are obtained by solving the correlation matrix of the observed features. Eigenvalues and EigenVectors are sorted in ascending order ([Smith, 2002](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B44)).

Using PCA(Principal Component Analysis) we were able to find the eigenvalues and eigenvectors for each feature and top features with highest eigenvalue and eigenvectors were picked . [Polat and Günes (2007)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B33) distinguished diabetes from normal people by using principal component analysis (PCA) and neuro.

Different Machine Learning Algorithms were used such as ([Kavakiotis et al., 2017](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B18)) support vector machine (SVM), decision tree (DT) Decision tree ([Friedl and Brodley, 1997](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B8); [Habibi et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B10); [Liao et al., 2018](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B26)), [Ozcivit And Gluten (2011)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B31)proposed a new method of ensemble learning known as Random Forest ([Breiman, 2001](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B4); [Lin et al., 2014](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B28); [Svetnik et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B46)) and Kernel-SVM are implemented using the above features and the output is predicted based on the train data for each model. After Predicting the output the accuracy has been calculated using different Performance Metrics such as Confusion Matrix, HeatMap, Accuracy, Precision, Recall.

In type 2 Diabetes Using PCA features the Decision Tree has the highest accuracy of 92.44 %

**Using t-SNE Features on Different Machine Learning Algorithms**

**Visualizing Data using t-SNE *Laurens van der Maaten, Geoffrey Hinton*; 9(86):2579−2605, 2008.**

t-SNE which is commonly known as t-Stochastic Neighbour-hood Embedding is an Unsupervised , non-Linear based approach for feature extraction in higher dimensions.t-SNE gives the overall view on how the data is arranged in the higher dimensions.

t-SNE is different from PCA , In t-SNE we preserve the small pairwise distance but when we take PCA we will be preserving the Large pairwise distance. Generally the t-SNE algorithm calculates the pairwise distance in both higher and lower dimensions and optimises the two similarity measures using cost function.

### [How to use **t**-**SNE** effectively](https://distill.pub/2016/misread-tsne/?_ga=2.135835192.888864733.1531353600-1779571267.1531353600)

[M Wattenberg](https://scholar.google.co.in/citations?user=pv54dqMAAAAJ&hl=en&oi=sra), [F Viégas](https://scholar.google.co.in/citations?user=GvXDNsYAAAAJ&hl=en&oi=sra), [I Johnson](https://scholar.google.co.in/citations?user=lml0nEgAAAAJ&hl=en&oi=sra) - Distill, 2016 - distill.pub

Different Machine Learning Algorithms were used such as ([Kavakiotis et al., 2017](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B18)) support vector machine (SVM), decision tree (DT) Decision tree ([Friedl and Brodley, 1997](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B8); [Habibi et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B10); [Liao et al., 2018](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B26)), [Ozcivit And Gluten (2011)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B31)proposed a new method of ensemble learning known as Random Forest ([Breiman, 2001](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B4); [Lin et al., 2014](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B28); [Svetnik et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B46)) and Kernel-SVM are implemented using the above features and the output is predicted based on the train data for each model. After Predicting the output the accuracy has been calculated using different Performance Metrics such as Confusion Matrix, HeatMap, Accuracy, Precision, Recall.

In type 2 Diabetes Using PCA features the SVM has the highest accuracy of 90.44 %

**RESULT**

Type 1 Diabetes

Generally Type 1 Diabetes will be seen in younger people whose age is less than or equal to 30 years. They will be having frequent urination, high blood glucose levels etc ([Iancu et al., 2008](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B12)). In type 1 Diabetes Using t-SNE features the Random Forest has the highest accuracy of 96.77 % when compared with other models. RF generates many decision trees, which is very different from decision tree algorithms ([Pal, 2005](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B32)).

Random Forest Generates Multiple Decision Trees ([Pal, 2005](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B32)). In the Random Forest approach each Decision Tree will be predicting the output. After Predicting the output from each tree Majority Voting technique will be applied. ([Liaw and Wiener, 2002](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B27); [Svetnik et al., 2015](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B46)).

Therefore using Random Forest with t-SNE features was used to predict whether a patient is Diabetic or Non-Diabetic.

**HeatMap For Type 1 Diabetes**



**Type 2 Diabetes**

In type 2 Diabetes Using PCA features the Decision Tree has the highest accuracy of 92.44 % when compared with other models.

In this we have used Different Decision Tree models like C4.5, WEKA etc ([Salzberg, 1994](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B42); [Kohabi, 1996](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B20)).

Therefore using the Decision Tree with PCA features was used to predict whether a patient is Diabetic or Non-Diabetic.

**HeatMap For Type 2 Diabetes**



**CONCLUSION**

Diabetes mellitus could be a disease, which might cause many complications. a way to exactly predict and diagnose this disease by using machine learning is worth studying. In this project, we have implemented Machine Learning Algorithms to find out whether the Patient is Diabetic or Non-Diabetic based on the given features for both Type 1 and Type 2 Diabetes. Different Preprocessing and Data Cleaning Techniques were used. EDA and Dimensionality Reduction were wont to find the most effective features for building the model.

Exploratory Data Analysis has been used to find the best features using pair plots. But none of the pair plots found to be useful except three plots Pregnancies, Glucose, Age. Univariate Analysis was performed on this Using PDF and Box Plots. The error was high so the above mentioned features were also found to be irrelevant.

As we were unable to obtain the best features using EDA another method of feature Extraction was used that is Dimensionality Reduction Using PCA and t-SNE. Using these Features Different Machine Learning algorithms were implemented.

PCA was implemented on Both Type 1 and Type 2 Diabetes to find the best features using EigenVectors and EigenValues. To improve the accuracy another method of Dimensionality Reduction was used that is t-SNE on both Type 1 and Type 2 Diabetes.

In type 1 Diabetes Using t-SNE features the Random Forest has the highest accuracy of 96.77 % when compared with other models. Therefore using Random Forest with t-SNE features was used to predict whether a patient is Diabetic or Non-Diabetic.

In type 2 Diabetes Using PCA features the Decision Tree has the highest accuracy of 92.44 % when compared with other models. Therefore using the Decision Tree with PCA features was used to predict whether a patient is Diabetic or Non-Diabetic.

**REFERENCES**

1. Aishwarya R., Gayathri P., Jaisankar N.

A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET), 5 (2013), pp. 2903-2908

1. Bamnote M.P.G.R.

Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing, 1 (2014), pp. 763-770

doi:10.1007/978-3-319-11933-5.

1. Arora, R. Suman

Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications, 54 (2012), pp. 21-25

doi:10.5120/8626-2492.

1. Aljumah A.A., Ahamad M.G., Siddiqui M.K.

Application of data mining: Diabetes health care in young and old patients

1. Journal of King Saud University-Computer and Information Sciences, 25 (2013), pp. 127-136

doi:10.1016/j.jksuci.2012.10.003.

1. Kanchan Dhomse B.M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5-10.
2. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima Indian diabetes dataset using Naive Bayes with the genetic algorithm as an attribute selection, in Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451-455.
3. Esposito F., Malerba D., Semeraro G., Kay J.

A comparative analysis of methods for pruning decision trees

IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (1997), pp. 476-491.

1. Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed. Res. Int.* 2016:1654623. doi: 10.1155/2016/1654623
2. Cox, M. E., and Edelman, D. (2009). Tests for screening and diagnosis of type 2 diabetes. *Clin. Diabetes* 27, 132–138. doi: 10.2337/diaclin.27.4.132
3. Duygu,ç., and Esin, D. (2011). An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. *Expert Syst. Appl.* 38, 8311–8315.
4. Friedl, M. A., and Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sens. Environ.* 61, 399–409.
5. Habibi, S., Ahmadi, M., and Alizadeh, S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. *Glob. J. Health Sci.* 7, 304–310. doi: 10.5539/gjhs.v7n5p304
6. Han, L., Luo, S., Yu, J., Pan, L., and Chen, S. (2015). Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. *IEEE J. Biomed. Health Inform.* 19, 728–734. doi: 10.1109/JBHI.2014.2325615
7. Iancu, I., Mota, M., and Iancu, E. (2008). “Method for the analysing of blood glucose dynamics in diabetes mellitus patients,” in *Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics*, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883
8. Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74, 2204–2214. doi: 10.2307/1939574
9. Jiang, Y., and Zhou, Z. H. (2004). Editing training data for kNN classifiers with neural network ensemble. *Lect. Notes Comput. Sci.* 3173, 356–361. doi: 10.1007/978-3-540-28647-9\_60
10. Jolliffe, I. T. (1998). “Principal components analysis,” in *Proceedings of the International Conference on Document Analysis and Recognition* (Heidelberg: Springer).
11. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* 15, 104–116. doi: 10.1016/j.csbj.2016.12.005
12. Kohavi, R. (1995). “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal.
13. Kim, J. H. (2009). Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* 53, 3735–3745. doi: 10.1016/j.csda.2009.04.009
14. Kohabi, R. (1996). “Scaling up the accuracy of naive-bayes classifiers : a decision-tree hybrid,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR.
15. Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R. News* 2, 18–22.
16. Liao, Z., Ju, Y., and Zou, Q. (2016). Prediction of G protein-coupled receptors with SVM-Prot features and random forest. *Scientifica* 2016:8309253. doi: 10.1155/2016/8309253
17. Ozsoft, A., and Gluten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Comput. Methods Programs Biomed.* 104, 443–451. doi: 10.1016/j.cmpb.2011.03.018
18. Quinlan, J. R. (1986). Induction on decision tree. *Mach. Learn.* 1, 81–106. doi: 10.1007/BF00116251
19. Quinlan, J. R. (1996a). “Bagging, boosting, and C4.5,” in *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (Menlo Park, CA: AAAI Press), 725–730.
20. Refaeilzadeh, P., Tang, L., and Liu, H. (2016). “Cross-validation,” in *Encyclopedia of Database Systems*, eds L. Liu and M. T. Özsu (New York, NY: Springer), 532–538.
21. Sakar, C. O., Kursun, O., and Gurgen, F. (2012). A feature selection method based on kernel canonical correlation analysis and the minimum redundancy-maximum relevance filter method. *Expert Syst. Appl.* 39, 3432–3437. doi: 10.1016/j.eswa.2011.09.031
22. Salzberg, S. L. (1994). C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers, Inc., 1993. *Mach. Learn.* 16, 235–240.
23. Sharma, S., Agrawal, J., and Sharma, S. (2014). classification through machine learning technique: C4. 5 algorithm based on various entropies. *Int. J. Comput. Appl.* 82, 28–32.
24. Smith, L. I. (2002). A tutorial on principal components analysis. *Inform. Fusion* 51:52.
25. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2015). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inform. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

**AUTHOR**

K.Thirunavukkarassu.,Assistant Professor,CSE,TPGIT

thirunatech@gmail.com