# Diabetes Prediction using Machine Learning

Amit Kumar Ashutosh(11906959)
Computer Science, Lovely Professional University

## Abstract

Diabetes is an illness induced due to excessive glucose degree in a human frame. Diabetes ought to now no longer be unnoticed if it's miles untreated then Diabetes may also purpose a few foremost troubles in someone like: coronary heart associated problems, kidney problem, blood pressure, eye harm and it may additionally impacts different organs of human frame. Diabetes may be managed if it's miles expected earlier. To reap this intention this undertaking paintings we can do early prediction of Diabetes in a human frame or a affected person for a better accuracy thru applying, Various Machine Learning Techniques. Machine studying strategies Provide higher end result for prediction via way of means of constructing fashions from datasets accumulated from patients. In this paintings we can use Machine Learning Classification and ensemble strategies on a dataset to expect diabetes. Which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). The accuracy is unique for each version whilst as compared to different fashions. The Project paintings offers the correct or better accuracy version indicates that the version is capable of predicting diabetes effectively. Our Result indicates that Random Forest carried out better accuracy as compared to different system studying strategies.

# Introduction

Diabetes is noxious sicknesses with inside the world. Diabetes precipitated due to weight problems or excessive blood glucose stage, and so forth. It influences the hormone insulin, ensuing in unusual metabolism of crabs and improves stage of sugar with inside the blood. Diabetes takes place whilst frame does now no longer make sufficient insulin. According to (WHO) World Health Organization approximately 422 million humans laid low with diabetes particularly from low or idle earnings countries. And this can be accelerated to 490 billion as much as the 12 months of 2030. However incidence of diabetes is discovered amongst numerous Countries like Canada, China, and India etc. Population of India is now extra than a hundred million so the real wide variety of diabetics in India is forty million. Diabetes is primary motive of demise with inside the world. Early prediction of sickness like diabetes may be managed and shop the human life. To accomplish this, this paintings explores prediction of diabetes with the aid of using taking numerous attributes associated with diabetes sickness. For this motive we use the Pima Indian Diabetes Dataset, we observe numerous Machine Learning class and ensemble Techniques to expect diabetes. Machine Learning Is a way this is used to educate computer systems or machines explicitly. Various Machine Learning Techniques offer green end result to acquire Knowledge with the aid of using constructing numerous class and ensemble fashions from accumulated dataset. Such accumulated information may be beneficial to expect diabetes. Various strategies of Machine Learning can successful to do prediction, but its hard to select pleasant technique. Thus for this motive we observe famous class and ensemble techniques on dataset for prediction.

# PROPOSED METHODOLOGY

Goal of the paper is to investigate for model to predict dia- betes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

1. Dataset Description- the data is gathered from dropbox . The dataset have many attributes of 768 patients.

   Table 1: Dataset Description

| S No. | Attributes |
|---|---|
| 1 | Pregnancy |
| 2 | Glucose |
| 3 | Blood Pressure |
| 4 | Skin Thickness |
| 5 | Insulin |
| 6 | BMI(Body Mass Index) |
| 7 | Diabetes Pedigree Function |
| 8 | Age |

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

Distribution of Diabetic patient- We made a version to expect diabetes but the dataset changed into barely imbalanced having round 500 instructions categorized as 0 method negative method no diabetes and 268 categorized as 1 method advantageous method diabetic.

# Data Preprocessing

Data preprocessing is most important technique. Mostly healthcare associated statistics consists of lacking vale and different impurities that could purpose effectiveness of statistics. To enhance nice and effectiveness received after mining technique, Data preprocessing is done. To use Machine Learning Techniques at the dataset efficiently this technique is crucial for correct end result and a hit prediction. For Pima Indian diabetes dataset we want to carry out pre processing in steps.

1. Missing Values removal- Remove all of the times which have zero (0) as really well worth. Having 0 as really well worth isn't always possible. Therefore this example is eliminated. Through eliminating beside the point capabilities/times we make characteristic subset and this technique is referred to as capabilities subset selection, which reduces diamentonality of statistics and assist to paintings faster.

2. Splitting of data- After cleansing the records, records is normalized in schooling and trying out the version. When records is spitted then we teach set of rules at the schooling records set and preserve take a look at records set aside. This schooling procedure will produce the schooling version primarily based totally on common sense and algorithms and values of the characteristic in schooling records. Basically purpose of normalization is to deliver all of the attributes beneathneath equal scale.

# Apply Machine Learning

When records has been prepared we follow Machine Learning Technique. We use specific class and ensemble techniques, to expect diabetes. The techniques implemented on Pima Indians diabetes dataset. Main goal to use Machine Learning Techniques to research the overall performance of those techniques and locate accuracy of them, and additionally been capable of discern out the responsible/critical function which play a chief position in prediction. The Techniques are follows-

1. **Support Vector Machine-** Support Vector Machine additionally called svm is a supervised gadget mastering algo- rithm. Svm is maximum famous category technique. Svm creates a hyperplane that separate instructions. It can create a hyperplane or set of hyperplane in excessive dimensional space. This hyper aircraft may be used for category or regression additionally. Svm differentiates times in unique instructions and also can classify the entities which aren't sup- ported via way of

means of data. Separation is carried out via way of means of via hyperplane plays the separation to the nearest education factor of any class.

**Algorithm-**

- Select the hyper plane which divides the class better.
- To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to
- Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

2.**K-Nearest Neighbor-** KNN is likewise a supervised ma- chine studying set of rules. KNN allows to remedy each the class and regression problems. KNN is lazy predic- tion technique.KNN assumes that comparable matters are close to to every other. Many instances information factors that are comparable are very close to to every other.KNN allows to organization new paintings primarily based totally on similarity measure.KNN set of rules file all of the data and classify them in line with their similarity measure. For locating the gap among the factors makes use of tree like structure. To make a prediction for a brand new information point, the set of rules unearths the nearest information factors withinside the train- ing information set its nearest neighbors. Here K= Number of close by neighbors, its constantly a high quality integer. Neighbors cost is selected from set of class. Closeness is mainly defined in terms of Euclidean distance. The Euclidean dis- tance between two points P and Q i.e. P (p1,p2, . Pn) and Q (q1, q2,..qn) is defined by the following equation.

**Algorithm**-

- Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.
- Take a test dataset of attributes and rows.
- Find the Euclidean distance by the help of formu- la-
- Then, Decide a random value of K. is the no. of nearest neighbors
- Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
- Find out the same output values.

If the values are same, then the patient is diabetic, other- wise not.

**3.Decision Tree-** Decision tree is a simple type method. It is supervised gaining knowledge of method. Decision tree used while reaction variable is categorical. Decision tree has tree like shape primarily based totally version which describes classi- fication technique primarily based totally on enter feature. Input variables are any kinds like graph, text, discrete, non-stop etc.

**Algorithm-**

- Construct tree with nodes as input feature.
- Select feature to predict the output from input fea- ture whose information gain is highest.
- The highest information gain is calculated for each attribute in each node of tree.
- Repeat step 2 to form a subtree using the feature which is not used in above node.

**4.Logistic Regression-** Logistic regression is likewise a supervised mastering category algorithm. It is used to estimate the possibility of a binary reaction primarily based totally on one or extra predictors. They may be non-stop or discrete. Logistic regression used while we need to categorize or distinguish a few facts objects into categories. It classify the facts in binary shape approach handiest in 0 and 1 which refer case to categorize affected person this is high-quality or negative for diabetes. Main goal of logistic regression is to satisfactory healthy that's answerable for describing the connection among goal and predictor variable. Logistic regression is a primarily based totally on Linear regression version. Logistic regression version makes use of sigmoid characteristic to are expecting possibility of high-quality and negative class.

Sigmoid function P = 1/1+e – (a+bx) Here P = probability, a and b = parameter of Model.

**5.Random Forest-** It is form of ensemble studying method and extensively utilized for type and regression tasks. The accuracy it offers is grater then in comparison to different models. This technique can without difficulty deal with huge datasets. Random Forest is evolved via way of means of Leo Bremen. It is famous ensemble Learning Method. Random Forest Improve Performance of Decision Tree via way of means of lowering variance. It operates via way of means of building a mess of selection bushes at education time and outputs the elegance this is the mode of the instructions or type or imply prediction (regression) of the individual bushes.

**Algorithm-**

- he first step is to select the R features from the total features m where R<<M.
- Among the R features, the node using the best split point.
- Split the node into sub nodes using the best split.

- Repeat a to c steps until l number of nodes has been reached.

- Built forest by repeating steps a to d for a num- ber of times to create n number of trees.

  The random forest finds the best split using the Gin-Index Cost Function which is given by:

The first step is to want the take a look at selections and use the rules of every indiscriminately created choice tree to expect the end result and shops the expected final results at durations the goal place. Secondly, calculate the votes for every expected goal and ultimately, admit the excessive voted expected goal due to the final prediction from the random wooded area formula. Some of the alternatives of Random Forest does accurate predictions end result for a range of programs are offered.

6.Gradient Boosting- Gradient Boosting is maximum power- ful ensemble approach used for predictio and it's miles a clas- sification approach. It integrate week learner collectively to make sturdy learner fashions for prediction. It makes use of Decision Tree version. it classify complicated records units and it's miles very ef- fective and famous method. In gradient boosting version overall performance enhance over iterations.

**Algorithm-**

- Consider a sample of target values as P

- Estimate the error in target values.

- Update and adjust the weights to reduce error M.
- P[x] =p[x] +alpha M[x]
- Model Learners are analyzed and calculated by loss function F
- Repeat steps till desired & target result P.

# Model Building

This is maximum critical segment which incorporates version build- ing for prediction of diabetes. In this we've got applied numerous gadget gaining knowledge of algorithms which might be mentioned above for diabetes prediction.

**Procedure-**

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K- Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned ma- chine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned ma- chine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experi- mental performance results obtained for each classifier.

Step8: After analyzing based on various measures con- clude the best performing algorithm.

# Result

In this paintings distinct steps had been taken. The proposed approach makes use of distinct type and ensemble techniques and applied the usage of python. These techniques are standard Machine Learning techniques used to reap the exceptional ac- curacy from data. In this paintings we see that random wooded area classifier achieves higher as compared to others. Overall we've used exceptional Machine Learning strategies for prediction and to gain excessive overall performance accuracy. Figure indicates the end result of those Machine Learning techniques.

Here characteristic performed essential function in prediction is present- ed for random woodland algorithm. The sum of the significance of every characteristic gambling important function for diabetes were plotted, in which X-axis represents the significance of every characteristic and Y-Axis the names of the features.

# Conclusion

The major intention of this undertaking turned into to layout and put in force Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that techniques and it's been completed successfully. The proposed technique makes use of diverse class and ensemble gaining knowledge of technique wherein SVM, KNN, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. And 77% class accuracy has been completed. The Experimental consequences may be asst fitness care to take early prediction and make early selection to treatment diabetes and keep humans life.

# References

1. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importances for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

2. Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

3. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

4. https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques

5. https://www.hindawi.com/journals/jhe/2021/9930985/

6. https://www.researchgate.net/publication/339543101_Diabetes_Prediction_using_Machine_Learning_Algorithms