

**Prediction of Diabetes using Machine Learning**

A Report Submitted for Partial Fulfilment of

Course-

**Project Based Learning**

First Year Engineering Programme of SPPU, Semester-II

(2020-21)

Submitted By:

|  |  |
| --- | --- |
| Roll No. | Name of Student |
| FE2021ENTC101  FE2021ENTC107 | Vaishnavi Kachare  Pooja Katre |
| FE2021ENTC112 | Swapnil Jadhav |
| FE2021ENTC113 | Siddhesh Munfan |
| FE2021ENTC115 | Sharayu Munde |
|  |  |
|  |  |

Under the Guidance of

**Prof. Jayant Shimpi**

June-July 2021

**CERTIFICATE**

This is to certify that, report **“Prediction of Diabetes using Machine Learning**

submitted by **Vaishnavi Kachare ( 101 )**

**Pooja Katre ( 107 )**

**Swapnil Jadhav ( 112 )**

**Siddhesh Munfan ( 113 )**

**Sharayu Munde ( 115 )**

of First Year Engineering Programme, is bonafide work completed in partial fulfilment of Course **Project Based Learning**.

**Prof. Jayant Shimpi**

**Project Guide**

**Prof. Hnnie Williams Dr. Sanjay K. Babar**

**Academic Coocrdinator Dean- FE Section**

**Dr. Vijay M. Wadhai**

**Principal**

**SYNOPSIS**

Diabetes is an illness caused because of high glucose level in a human body. Over 40 million people in India are suffering from diabetes and many others are under the risk. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. . The risk of Type 2 diabetes was predicted using machine learning algorithm as these algorithm is highly accurate which is very much required in the health profession. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by constructing models from datasets collected from patients. In order to conduct the experiment, 768 instances have been collected through an online Pima Indian database. In this work we will use Machine Learning Classification on a dataset to predict diabetes, which is K-Nearest Neighbor (KNN). The Project work gives the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively. Our Result shows that we have achieved good accuracy.

INDEX

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **TITLE** | **PAGE NO** |
| 1 | PREAMBLE | 4 |
| 2 | LITERATURE REVIEW/ THEORETICAL | 5 |
| 3 | DESIGN AND DEVELOPMENT | 7 |
| 4  5 | CONCLUSION  REFERANCES | 11  12 |

Chapter: 1

Introduction

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it’s tough to choose best technique. Thus for this purpose we apply popular classification and ensemble method on dataset for prediction.

Chapter: 2

Literature Review

In India, diabetes is a widespread problem as more than 70% of the adult population is suffering from this disease. Various researchers have worked to predict symptoms of diabetes by applying different approaches such as machine learning and data mining [11]. Few of them have also applied neural network and genetic algorithm. Since the problem of prediction of diabetes is supervised in nature, the supervised methods of machine learning, data mining and ANN have been applied by many.

Some closely related works are discussed in this section. Many of the research studies have used Pima Indians Diabetes Dataset (PIDD) for diabetes prediction. Machine learning methods and Weka tool were applied by [13, 14, 16, 17, 20, 21, 23]. The different approaches applied by researchers can be broadly classified as machine learning methods, data mining techniques, hybrid methods and neural network or genetic algorithms.

Swapna et. al. in [12] used deep learning methods on electrocardiogram (ECG) signals for detection of diabetes. Specifically, convolution neural network and long short-term memory has been used by them and then features were extracted by support vector machine. As a result, they found a very high accuracy of 95.7%.

Sisodia et. al. in [13] applied three machine learning methods i.e. decision tree (DT), naïve based (NB) and support vector machine (SVM) on PIDD in order to predict the diabetes. Naïve bayes classifier was found to be 76.30% accurate.

Han Wu et. al. in [14] applied data mining techniques (i.e. improved kNN and logistic regression) to accurately predict up to 95.42% the risk to an individual of developing type 2 diabetes. The modification was done by selecting value of initial seed point experimentally. The initial seed point was selected by conducting 100 experiments in which they selected smallest value of ‘within cluster sum of squared errors.’

Meng et. al. in[15] compared logistic regression, artificial neural network (ANN) and decision tree (DT) for identifying the risk of diabetes and prediabetes based on 12 risk factors which included education level, work stress, BMI, age, sleep duration, gender, marital status, family history of diabetes, coffee drinking, preference to salty foods, physical activity, and consumption of fish. DT was found to provide best results among the three methods.

Choubey et. al. in [16] applied a hybrid algorithm using genetic algorithm (GA) and radial basis function neural network (RBFNN), wherein first GA is applied for feature selection then RBFNN is applied for classification. Their findings were that hybrid method performed better than RBFNN alone.

Tigga et. al. in [17] applied logistic regression on PIDD for diabetic prediction and found number of pregnancies, BMI and glucose level are most significant variables for diabetes prediction among all features in PIDD.

Huang et. al. in [18] did feature selection and classification of diabetes by applying naïve bayes, IB1 and C4.5 algorithms. The study concluded that patient age, diagnosis duration, need of insulin and diet control are most important factors for blood sugar control. Some other factors are also affecting results that are type of care, home monitoring and importance of smoking.

Saravana et. al. in [19] collected raw data from various places in form of Electronic Reports (EHR) that may be clinical reports, prescriptions given by doctors, diagnostic centre reports, pharmacy related information, and data asked by insurance personals. All this information collectively put in a map reduce to exact features which are directly related to diabetes.

Nongyao et. al. in [20] compared four classification techniques i.e. decision tree, ANN, logistic regression and naïve bayes. Further bagging and boosting were applied on all and random forest was also included. The maximum accuracy achieved by all was in between 84% and 86%.

Zou et. al. in [21] applied Random Forest, J48, ANN for classification after the feature reduction is done by unsupervised methods: Principle Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) methods. Accuracy for mRMR is found to be better than PCA with all features.

Perveen et. al. in [22] were concerned in finding risk of metabolic syndrome and diabetes. For prediction Naïve Bayes and J48 (C4.5) decision tree model were applied and the balancing of training set was done by k-medoids sampling. In their study, NB outperformed the others.

Rahman et. al. in [23] summarises the effect of various data mining techniques for diabetes diagnosis. For the prediction purpose, Multilayer Perceptron (MLP), Bayes Classification, J48graft, JRip (RIP- PER), Fuzzy Lattice Reasoning (FLR) classification methods were applied. J48graft was found most accurate.

Choi et. al. in [24] applied machine learning algorithms on patients having history of non-diabetes having cardiovascular risk. Five years data has been collected in form of EMR from Korea University Guro Hospital. Then, machine learning methods were applied with 10-fold cross validation. Highest accuracy was obtained in logistic regression model.

Chapter : 3

**Design And Development/ Methodology Used To Collect Data And Its Analysis**

Goal of the project is to investigate for model to predict diabetes with better accuracy. We experimented with classification and algorithms as kNN to predict diabetes. In the following, we briefly discuss the phase.

1. **Dataset Description -** the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. 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.

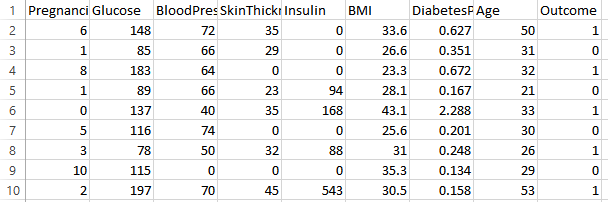


Fig : Pima Indians Dataset

**Distribution of Diabetic patient-** We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.

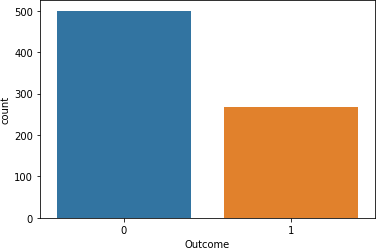
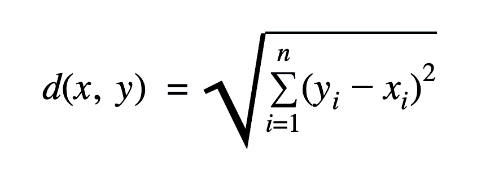
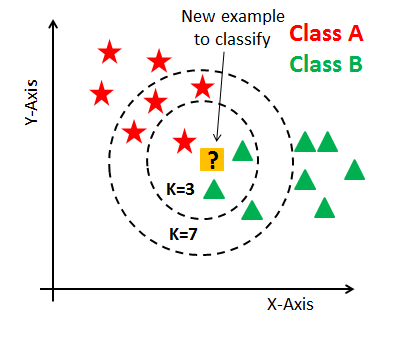


Figure 1: Ratio of Diabetic and Non Diabetic Patient

1. **Data Pre-processing -** Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effective- ness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform preprocessing in two steps.
2. **Missing Values removal-** Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces diamentionality of data and help to work faster.
3. **Splitting of data-** After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.
4. **Apply Machine Learning-** When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyse the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction. The Techniques are follows :-

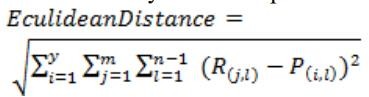
**K-Nearest Neighbor -** KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique. KNN assumes that similar things are near to each other. Many times data points which are similar are very near to each other. KNN helps to group new work based on similarity measure. KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — it’s nearest neighbors. Here **K**= Number of nearby neighbors, it’s always a positive integer. Neighbor’s value is chosen from set of class. Closeness is mainly defined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e. P (p1,p2, …. Pn) and Q (q1, q2,..qn) is defined by the following equation:-





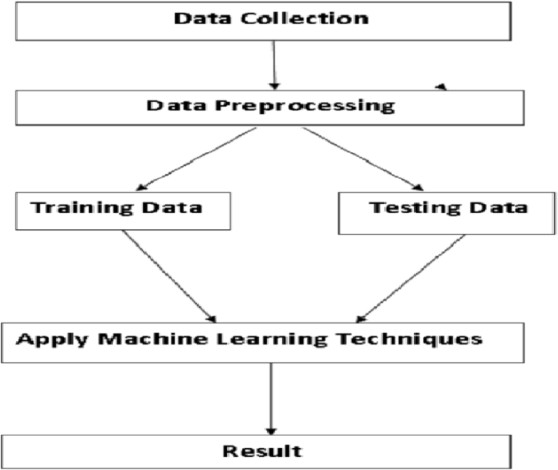
## Algorithm-

* + 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 formula :-



* + 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.

1. If the values are same, then the patient is diabetic, otherwise not.

Figure 2: Overview of the Process

**MODEL BUILDING: -**

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology-

**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

**Step5:** Build the classifier model for the mentioned machine learning algorithm based on training set.

**Step6:** Test the Classifier model for the mentioned machine learning algorithm based on test set.

**Step7:** Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

**Step8:** After analyzing based on various measures conclude the best performing algorithm.

Chapter: 4

**Conclusion**

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Method and Performance Analysis of that method and it has been achieved successfully. The proposed approach uses classification and ensemble learning method which is kNN method is used. And 78% classification accuracy has been achieved. The Experimental results can be asst. health care to take early prediction and make early decision to cure diabetes and save humans life.

**References**

1. Strategies to make online teaching more effective Project by “Pooja Rathi” from Researchgate.in publication.
2. Sci-kit learn Python Module
3. Pima Indains Dataset from Kaggle.
4. Some of the Youtube Channels.
5. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.
6. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.
7. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.