**DIABETES PREDICTION USING MACHINE LEARNING**

**Akshata Nitin Agine – RMCET, Ambav, Ratnagiri**

**Abstract** — Diabetes Mellitus is a chronic disease characterized by hyperglycemia. It is a common disease for human body caused by metabolic disorder when the sugar level is high. It can cause many complications. According to growing morbidity by the year 2050, the world’s diabetic patients will reach 740 millions, which means that one of ten adults or children may suffer diabetes. Early prediction of such disease can save human life. To achieve this goal researchers are mainly working on this risk factor related to diabetes using machine learning techniques. With rapid development of machine learning, machine learning has been applied in many aspects of medical health. In this study, we are using some popular machine learning algorithms namely K-Nearest Neighbour, Support Vector Machine (SVM) and Naïve Bayes to predict diabetes mellitus. In our experimental results it shows that KNN and SVM have achieved the highest accuracy compared to Naïve Bayes machine learning technique.

**Keywords**— Diabetes Mellitus, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, Machine Learning, Prediction.

1. **Introduction:** Diabetes is a common chronic disease which can pose great threat to human health. Diabetes can be identified when blood glucose is higher than normal level, which is caused by high secretion of insulin or biological effects. Diabetes can cause various damage to our body and can dis-function tissues, kidneys, eyes and blood vessels. Diabetes can be divided into two categories, type 1 diabetes and type 2 diabetes. Patients with type 1 diabetes are normally younger with an age less than 30 years old. The clinical symptoms are increase thirst and frequent urination this type of diabetes cannot be cleared by medications as it requires therapy. Type 2 diabetes occurs more commonly on middle-aged and old people, which can show hypertension, obesity and other diseases. With our living standards diabetes has increased commonly in people’s daily life. So how to analyse diabetes is worth studying. As we get the diagnosis earlier we can control it. Machine learning can make preliminary judgement on diabetes mellitus according to physical examination data, and by reference with doctors. Recently, many algorithms are used to predict diabetes, including machine learning methods like Support Vector Machine (SVM), (KNN) K-Nearest Neighbour, Naive Bayes and so on. With this machine learning techniques we are able to predict diabetes by constructing predicting models which are obtained by medical datasets. By extracting such knowledge we are able to predict diabetic patient. We use the best technique to predict based on our attributes of the given datasets in order to get the perfect accuracy to predict diabetes mellitus.
2. **Literature Survey:** This section shows our existing recent literature work and provide us the understanding the challenges of our given approaches. Various computing techniques were used in this healthcare domain. The focus on this literature survey is the use of different machine learning algorithms used for predicting diabetes mellitus. In order to get the perfect accuracy we extract the knowledge from the given medical data. The Asaduzzaman [1] used 10 fold cross validation on an evaluation method for three different algorithms which included decision tree, naive bias and SVM where naïve bias have shown the accuracy of 75% than other given algorithms. Chun li [2] used random forest, KNN, Naive Bayes, SVM, decision tree to predict diabetes mellitus early stage. Currently in the healthcare domain we are implementing machine learning algorithms and statistical data to understand the diseased data which was discovered. Since the machine learning domain consists of various techniques and researches to make a comparison based on which algorithm is faster in giving the results of prediction. The classification of algorithm was not evaluated by cross validation method. To predict and analyse diabetes mellitus different data mining techniques were used. As we use three data mining techniques we used real word data sets by collecting information from the given datasets. In this work we have analysed real diagnostic medical data based on various risk factors for the classification International Journal of Scientific & Engineering Research Volume 12, Issue 3, March-2021 ISSN 2229-5518 63 IJSER © 2021 http://www.ijser.org IJSER of machine learning techniques and for predicting diabetes mellitus.
3. **METHODOLOGY:** In order to achieve our goal, our methodology comprises if few steps from which we accumulate datasets of the given attributes for the patients and we will do the pre-processing of our given attribute to apply on the given machine learning techniques to find out the predictive analysis of the data.
4. **DATASET AND ATTRIBUTES**

In this work, we collect diabetes data from Medipath Diagnostic Center (MDC), from Mumbai, Maharashtra, India. The dataset consists of various attributes for diabetes mellitus for 700 patients. The attributes are given in the below table.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pregnancies | Glucose | Blood Pressure | Skin Thick-ness | Insulin | BMI | Diabetes Pedigree Function | Age | Outcome |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.2 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

1. **DATA PREPROCESSING**

To achieve the goal, some data pre-processing is done on the given diabetes dataset. As it converts raw data in numerical form from which we are able to get the values of the attribute to predict diabetes. Here for example we can say that the age of the patient can be divided into three categories, such as young (10-23 years), adult (24-49 years), old (50 and above). Similarly a patient’s weight can also be classified into three categories as less (below 40 kg), normal (40-60kg) and overweight (above 60kg). Blood pressure is classified as normal (120/80 mmdl), low (less than 80mmdl), high (more than 120 mmdl).

1. **APPLYING MACHINE LEARNING TECHNIQUES**

Once the data has been created for modelling we employ our four machine learning classification algorithm which we are going to implement to predict diabetes mellitus. Some overview of these techniques.

1. **K-NEAREST NEIGHBOUR:** The k-nearest neighbour is a non-parametric method used for classification and regression. The input consist of k-closest training example in the feature space. To determine the distance from point of interest to point of training data set it uses. In classification technique, the value of k is always a positive integer of the nearest neighbour. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

**2) SUPPORT VECTOR MACHINE:** The objective of the support vector machine algorithm is to find a hyper-plane in an N-dimensional space (N - the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyper-planes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

**3) NAÏVE BAYES:** Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

1. **COMPARISON RESUTS**

The performance of the given machine learning techniques shows us the prediction result where we recall the algorithms used as K-Nearest Neighbour, Support Vector Machine, Naive Bayes. Here we see that classifiers KNN and SVM show the better results than Naïve Bayes classifier to predict diabetes. According to the figure, both KNN and SVM show 81% accuracy on this dataset, which is greater than all other machine learning techniques. This it shows that KNN and SVM perform well on this given medical dataset for predicting diabetes mellitus. In this work, we have chosen the best machine learning technique to predict diabetes mellitus to that we can achieve high evaluation as shown in the box plot from which we can estimate the testing dataset to show high performance of precision for our given dataset As both KNN and SVM are considered the best accuracy choice for this dataset as it shows that it can achieve up to 81% of accuracy to predict diabetes mellitus for the medical dataset.

**4) CONCLUSION:**

In this experiment we have analysed the early prediction of diabetes by taking all the related factors in its tests and implementing using machine learning techniques by extracting knowledge from our real health care medical dataset to predict diabetic patients thus we have done our experiment using some various machine learning algorithms namely K-Nearest Neighbour, Support Vector Machine and Naïve Bayes on Indian datasets to predict diabetes.