CSC 635 Data Mining

## Project Report

### Submitted to:

### Dr. Jamil Saquer

### Author(s):

*Bandlamudi Srinivasa Rao,*

*Kollabathula Preetham,*

*Vaishali Tappeta Venkataramana*

**Diabetics Prediction**

**Abstract**

*Diabetes is a disease in which blood glucose (also known as blood sugar) levels are unusually high. Blood glucose, which originates from the food we eat, is the body's main source of energy. Diabetes can lead to wide range of issues which affect every organ of the body. The objective of our project is to perform exploratory data analysis on a diabetes data set and predict diabetes in a patient based on the diagnostic measures by using multiple Machine Learning Algorithms such as K- Nearest Neighbors, Support Vector Machine, Decision trees, Naïve Bayes and Random Forest classifier. We then compare the accuracy for each model so that we can know the best model which gives highest score for our data set.*

**Introduction**

Diabetes is becoming a more prevalent health problem as a result of our sedentary lifestyle. If caught early enough, negative consequences can be avoided with adequate medical care. Technology may be used very effectively and reliably to aid in early detection. We developed a predictive model based on machine learning that can predict if a patient is diabetic or not.

Diabetes is a disease in which your blood glucose level, commonly known as blood sugar, is too high. Your major source of energy is blood glucose, which comes from the foods you eat. Insulin, a pancreatic hormone, boosts glucose absorption into cells for energy production. The following include the symptoms that the person should watch out for diabetics:

* Hunger and exhaustion.
* Being thirstier and urinating more frequently
* Itchy skin and a dry mouth
* Vision is hazy

To predict if a patient has diabetes, machine learning algorithms may be developed using the Pima Indians Diabetes Database. Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age are all included in this data set. Several independent variables and one dependent outcome variable, are included in the datasets. independent variables include the patient's number of pregnancies, BMI, insulin level, age, and other characteristics. A total of 768 patients are included in the Pima Indian Diabetes Data set (268 of whom were tested positive and 500 of whom were tested negative). In 2015, an estimated 415 million individuals worldwide had diabetes, with type 2 diabetes accounting for about 90% of cases. This corresponds to 8.3% of the adult population. Diabetes can cause a variety of problems that impact every organ in the body.

The goal of our project is to use multiple Machine Learning Algorithms such as K-Nearest Neighbors, Support Vector Machine, Decision trees, Nave Bayes, and Random Forest classifier to perform exploratory data analysis on a diabetes data set and predict diabetes in a patient based on diagnostic measures. The accuracy of each model is then evaluated so that we can determine which model is the best for our data set.

**Data set Description:**

The data set is Pima Indians diabetic data set which consists of 769 samples, out of which 500 are non-diabetic while 269 are diabetic people. The data set has total 9 attributes of which 8 are independent variables and one is the dependent variable (i.e the Outcome variable which determines whether the patient is having diabetics or not.

Sample data is given below:

A black screen with white text

Description automatically generated with low confidence

**Data Understanding:**

The detailed attributes in the data set are listed as follows:

* Pregnancies  (number of times pregnant)
* Glucose (two hour plasma glucose concentration after 75g anhydrous glucose in mg/dl)
* Blood pressure  (Diastolic Blood Pressure in mm hg)
* Skin Thickness (Triceps skin fold thickness in mm)
* The Insulin level in the blood (2-hour serum insulin in mIU/ml)
* BMI (Body Mass Index in kg/m2)
* Diabetics pedigree function ('function that represents how likely they are to get the disease by extrapolating from their ancestor’s history')
* Age  (years),
* Outcome (class variable (0 or 1) which denotes whether the patient is having diabetics or not)

The below subplots will provide a overview of the proximity of various features which will contribute to the positive diabetics prediction.

Shape

Description automatically generated

**Approach**

We are using Exploratory data analysis approach which is the use of summary statistics and graphical representations, exploratory data analysis refers to the crucial process of doing first investigations on data to uncover patterns, spot anomalies, test hypotheses, and check assumptions.

**Data Cleaning:**

We're doing exploratory data analysis and using Pandas dataframe.corr to discover the pairwise correlation of all columns in a dataframe. Any NA values are eliminated from the equation immediately. For any non-numeric data type columns in the data frame, it is ignored.

The sns map of the correlation among the feature attributes for the labels with positive diabetics value is given below.



**Data Preprocessing**

The quality of the data has a significant impact on the prediction outcome. This indicates that data preparation is critical to the model's success. First, we looked at the medical implications of each attribute and how they relate to diabetes. We discovered that the number of pregnancies has little to do with diabetes. Due to mistakes or deregulation, the data set contains some missing and erroneous values. To decrease the impact of nonsensical values, we replaced any missing values with the training data's means.

On taking the mean of the class labels, it has been determined that average values of the feature variables for those with negative prediction is lower than the values of those with Diabetics positive prediction.

Chart, bar chart

Description automatically generated

Also, it has been determined that the number of records with positive prediction of Diabetics class is significantly lesser than the records with negative prediction of diabetics class.

**Data Classification**

In this, we have tested the data set with various algorithms to test the accuracy of each algorithm and thereby finalizing it for use in our analysis. Our main objectives are as follows:

* To use data mining techniques to preprocess and classify the data.
* Comparing the accuracies of various Machine Learning models.
* Determining the most appropriate model for predicting diabetes based on user input.

**Algorithms used:**

1. **KNN Algorithm**

The k-nearest neighbor's algorithm (k -NN) is a classification technique that categorizes variables based on the training data that is closest to them in the feature space. When given extra unlabeled data and labeled input data, a supervised machine learning algorithm (as opposed to an unsupervised machine learning algorithm) trains a function that produces the desired output. One of the most fundamental data mining methods used by K -NN is an instance-based learning approach. Before selecting whether or not to assign an object to classes, this function considers its nearest neighbors.

1. **SVM Algorithm:**

Support Vector Machine (SVM) is a supervised algorithm for splitting the feature space into hydroplanes based on target classes. To compute classification, the margin of the hyperplane that intercepts classes is maximized in SVM. This approach builds a multidimensional hyperplane that separates classes and increases the gap between them in order to enhance classification accuracy. We used multiple kernel functions included in the SVM class of the SVC library in Python framework, such as quadratic, polynomial, radial basis, and so on, to classify the cases and identify the greatest accuracy among them.

1. **Logistic Regression:**

To assign observations to a discrete set of classes, the logistic regression classification procedure is utilized. Some instances of categorization challenges are email spam vs. non-spam, online transaction fraud vs. non-fraud, and tumor malignant vs. benign.The logistic sigmoid function translates the output of logistic regression into a probability value.

A common approach for predictive modeling practices is logistic regression. For all values of X, the function p(X) returns a probability between 0 and 1, where X1–Xp are the predictors. Maximum likelihood estimation is used to estimate the coefficients 0–p.

1. **Decision Tree Classifier:**

Decision Trees (DTs) are a supervised learning approach that may be used for classification and regression. The objective is to construct a model that predicts the value of a target variable by learning basic decision rules from data features. The piece wise constant is approximated by a tree.

Decision trees are used in the example below to predict a sine curve using data using a sequence of if-then-else decision rules. As the tree gets deeper and the model gets more accurate, the choice criteria get more complicated.

1. **Random Forest Algorithm:**

It's a common tool for dealing with structured (tabular) data sets, such as data from a spreadsheet or database table, to solve classification and regression predictive modeling problems. Time series may also be forecasted using Random Forest, although this requires first transforming the time series data set into a supervised learning problem. It also needs the use of a model assessment approach called walk-forward validation, as k-fold cross validation would yield optimistically biased findings.

1. **Gaussian Naïve Bayes:**

The Gaussian Naive Bayes variant enables continuous data and follows the Gaussian normal distribution. The Bayes theorem is the foundation for the Naive Bayes collection of supervised machine learning classification methods. It's a basic classification method that packs a punch. They're handy when the dimensional of the inputs is large. Complex classification challenges can also be solved with the Naive Bayes Classifier.

1. **Gradient Boosting Classifier:**

Gradient boosting classifiers are a collection of machine learning algorithms that combine several weak learning models to generate a powerful predictive model. When doing gradient boosting, decision trees are commonly employed. Gradient boosting models are gaining popularity as a result of their ability to classify difficult information.

**Experimental Setup and Results**

**Data Splitting**

Firstly, the data is divided into training and testing in the ratio of 80% of the data set for training and the remaining 20% of the data for testing.

**Data Modeling**

Tools Used:

Python-Scikit Learn

Scikit Learn provides a wide range of supervised and unsupervised learning algorithms. The following libraries must be installed before using the above tool. It includes Numpy, Matplotlib, Pandas, seaborn and sklearn.

Here, we are determining the testing and training accuracy of the K-Nearest Neighbour algorithm. The k-nearest neighbors are selected in the range from 1 to 10 and and the K Neighbors Classifier fit model is built using training data comprising of features and the labels. The accuracy of both the training data and the testing data will be recorded for the k-nearest neighbors in the given range. The Accuracy and K-neighbors plot is plotted for both the training and testing data and it has been observed that accuracy of training set decreases as the number of neighbors increases in the algorithm and the accuracy of the testing set decreases as the number of neighbors i.e. k value increases and both will converge at the maximum value of k. The plot is shown in the figure below.

Chart, line chart

Description automatically generated

This phase consists of application of appropriate model to the data. A wide variety of Machine learning algorithms ranging from k-nearest neighbor's algorithm, Support Vector Machine, Logistic Regression, Decision Tree Classifier, Random Forest Algorithm, Gaussian Naïve Bayes and Gradient Boosting Classifier are applies to the training and testing data.

The accuracy of each classifier model is determined as below.

Text

Description automatically generated

The below figure shows the plot of accuracy with respect to different machine learning algorithms.

Chart, bar chart

Description automatically generated

Our working model comprises of a predictor tool which enables the person to predict whether he/she has diabetics or not. We built this tool using the logistic regression algorithm as it has been proven to give the highest accuracy among the given tested algorithms. On entering the available features of the diabetics, the tool will predict whether the person is diabetic or non-diabetic.

**Evaluation:**

We are using confusion matrix as a performance measurement for machine learning classification problem where the predicted output is plotted with actual class labels.

True Positive: Case where both the predicted output and actual output are the same(Correct Prediction as Diabetic)

True Negative: The prediction matches the actual false output (Correct Prediction as not diabetic)

False Positive: We predicted as diabetic, but they actually don’t have the disease. (incorrect Prediction as diabetic)

False Negative: We predicted nondiabetic, but they actually have the disease. (incorrect Prediction as nondiabetic)

In the matrix below 0 -> Diabetic and 1 -> NonDiabetic

**A picture containing text, screenshot, electronics

Description automatically generated**

**Result:**

Nondiabetic patient prediction:

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Diabetic Patient prediction:

Text

Description automatically generated



So, based on all the attributes, the model will determine the probability of the person developing diabetics.

**Conclusion & Future Work:**

The diabetics prediction tool is helpful for doctors to identify the patients suffering from diabetics. It is also very helpful for normal people to check whether they are prone to diabetics or not by just entering simple values of pregnancy, blood pressure, Glucose, age, etc. and determine if they are developing any early signs of diabetics using this. Further, predicting this disease early leads to treating the patient before it becomes critical. Some future improvements to improve our predictive accuracy could be like using an error metric other than accuracy score, like f1 score or precision. Acquiring domain knowledge (acquired from a doctor or more qualified person) to make better choices about feature engineering. Implementing the model prediction with algorithms like Neural Networks instead. Having a large data set to train the model.

**References**:

[1] S. S. Patil and K. Malpe, "Implementation of Diabetic Retinopathy Prediction System using Data Mining," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1206-1210, doi: 10.1109/ICCMC.2019.8819865.

[2]Nahar N, Ara F. Liver disease prediction by using different decision tree techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol. 2018;8.

[3] The Pima Indians Diabetics Dataset, Version 1. Retrieved December 09, 2021 from https://www.kaggle.com/lucky1/pima-indians-diabetes-dataset-part-1/notebook#Pima-Indians-Diabetes-Data-Set.

[4] Juntao Wang and Xiaolong Su, An improved K-Means clustering algorithm, 2011 IEEE 3rd International Conference on Communication Software and Networks.