Categorising The Symptoms of Diabetes Type

By

Mr. Gopalakrishnan Kumar-IIT-B Alumnus

Consulant Data Scientist

Introduction

- Diabetes mellitus is a chronic metabolic disease characterized by hyperglycemia and causes long term complications in multiple systems of the human body.

- The purpose of this project is to predict the ability of machine learning for classifying the types of Diabetes Mellitus (DM) using factors that determine diabetic types in susceptible individual, thus helps health care providers in different settings to find out the disease in their earliest phase which in turn enhance good prognosis.

Aim of Project Work

- To identify the factors and symptoms of diabetes in patients and classify the patients as per "Type 1 diabetes" and "Type 2 diabetes". (Predictive analytics)

- To group the patients as per the types of diabetes they suffer and admit them in the proper ward. (Descriptive analytics)

Objectives

i. Applying proper supervised ML techniques to identify "Type 1 diabetes" or "Type 2 diabetes" diabetic patients.

ii. Harnessing suitable unsupervised learning to group "Type 1 diabetes" and "Type 2 diabetes" diabetic patients for admitting them in their respective wards.

iii. Conducting the performance evaluation for achieving the best results.

Results in Anticipation

- Variables description used in the analysis.
- Exploratory Analysis
- Data Pre-processing
- Models
- Model Comparative analysis
- Compare accuracy, precision and execution time

Methods and Techniques Adopted

Project Dataset, Code:

The project work is done on "Clean_Dia.xlsx" dataset.

The Code used in this project is from Jupyter/Python Consoles.

Data Quality:

In this project, wrangling is done by library loading and attaching various program packages. Data linkage is conducted by read.csv(file.choose()). File management and storage is done by R-Code aggression of R-Scripts and saving appropriate documents.

Analytics:

Exploratory Data Analysis:

In this project, EDA is done by conducting by Visual Inspection of data, Description of data, Histogram of Integer distributions.

Data Processing and Preprocessing:

In this project, 4 blocks of python-codes- Apriori, Binning, Handling missing values, PCA have been used.

Methods and Techniques Adopted

Model Feature Selection:

In this present project, the target variable is used by the following the steps below:

Feature selection is done by using Python Programming. It involves Python flask, shell to load dataset and subset only numerical features.

Model Comparative Analysis:

Model Evaluation:

It involves developing algorithms, namely LR, LDA, KNN, CART, NB, and SVM, to effective verify Model accuracy of the models.

Model evaluation is done by Gradient Regression Algorithm in which accuracy is calculated.

Conclusion

- Based on the Model Scores, Decision Tree Classifier (CART), Gaussian NB (NB) are the effective Models, in this project.

- The patients suffering from DM have been classified into Type1DM, Type2DM, and Type0DM using Gradient Boosting Algorithm.

- iii. The high risk factors and symptoms predominant in Type1DM are identified as

- a. Predominant in the age-group 0-19 years and 57-90 years.

- b. Alcoholism

Conclusion

- c. Dinner Habits
- d. Wound Healing
- e. Family History
- f. Weight Loss
- iv. The high risk factors and symptoms predominant in Type2DM are identified as
- a. Allergy
- b. Age Group 19-55 years
- c. PCos
- d. Gender
- e. Hand and Feet Numbness [100% Predominant in Type2]
- f. Gradual Weight Gain [Extremely High]
- v. The Risk Factors and Symptoms equally predominant in both Type1 and Type2
- are identified
- a. Micturition
- b. Visual Impairment
- c. Frequent Thirst

Conclusion

- The Probability Density and Bsl_ae Curve is narrowly distributed among Type1DM in the range of 180-350 bsl_ae and widely distributed among Type2DM in the range of 100-400 bsl_ae.

- Hence, the bsl_ae of majority of the patients of Type1DM falls in the range of 180-350 and the bsl_ae of the patients of Type2DM are in the range of 100-400.

References

- References- From *Kaggle*

- Detecting early Alzheimer's using MRI data and machine learning.

- Diabetes prediction with demographic features and body measurements.

- How to install Nvidia-Apex in notebook.

- Random Forest Benchmark ®.

- Back to (predict) the future - Interactive M5 EDA.