CIND820-Big Data Analytics Project : Predictive Modeling for Type-2 Diabetes

Introduction

This repository includes a collection of different documents such as text files, pdf, CSV files and python codes, relevant to the preparation, processing and implementation of the CIND820: Big Data Analytics Project.This project,Specifically, involveds in predictive modeling of type 2 diabetes using three selected machine learning algorithms: Logistic regression, Decision Tree and Random Forest.

Data Description

For this project, a publicly available 2015 USA Behavioral Risk Factor Surveillance System phone call survey dataset, originally collected by the Center for Disease Control and Prevention (CDC), is used. The data set contains 70,692 observations and 21 features. The target variable has binary classes: diabetes and non-diabetes. While some of the features that are expected to be the main risk factors include cholesterol, smoking, alcohol, heart disease, blood pressure, body max index, other healthy habits expected to reduce the risk of diabetes complications include consumption of fruits, vegetable and physical activity. This study also checks the prevalence of diabetes by features such as age, income level and education.

The dataset used for this research project could be acced from the following Kagle website.

```
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?s
```

Data Processing steps

1. Data cleaning
2. Data preprocessing
3. Exploratory Data Analysis
4. Feature selection
5. Feature scaling
6. Encoding categorical variables
7. Train-test Splitting
8. Modeling
9. Evaluation

Visualization Techniques

1. scatter plots/Bar graphs
2. Histograms and density plots

3. Correlation heatmaps
4. Box plots

Predictive Modeling

1) Logistic Regression

2) Decision Tree

3) Random Forest

Results

Conclusions