M.Sc. (Five Year Integrated) in Computer Science (Artificial Intelligence & Data Science)

Fifth Semester

Project Proposal

21-805-0506: R Programming Lab

Submitted by

OMAL S.

(80521015)

DEPARTMENT OF COMPUTER SCIENCE
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY (CUSAT)
KOCHI-682022

AUGUST 2022

Dataset : 10-Year Diabetes Dataset

The 10-Year Diabetes Dataset is a collection of data that spans a period of ten years and contains information related to diabetes and its potential risk factors. This dataset aims to explore the relationships between various clinical and demographic variables and the presence or absence of diabetes in individuals.

Dataset Source : https://www.kaggle.com/datasets/jimschacko/10-years-diabetes-dataset

Exploratory Data Analysis Questions :

1. What are the basic statistics (mean, median, standard deviation) for numerical variables like glucose levels, BMI, age, etc.?

2. What is the distribution of the target variable (e.g., diabetes diagnosis) across the dataset?

3. How has the distribution of diabetes cases changed over the 10-year period? Are there any trends or patterns?

4. Is there a seasonal pattern in the data? For example, do certain months or seasons have higher incidences of diabetes?

5. What is the age distribution of the individuals in the dataset? Is there a correlation between age and diabetes?

6. What other clinical parameters (blood pressure, insulin levels, etc.) are associated with diabetes?

7. What are the pairwise correlations between numerical variables? Do any variables exhibit strong correlations?

8. Are there opportunities to create new features from existing ones that might better capture patterns related to diabetes risk?

9. Do outliers have any impact on the relationships between variables or the distribution of diabetes cases?

10. Are there any extreme values or outliers in the dataset that could affect the analysis?

Inferred Question :

Based on the insights gained from EDA, what variables might be strong predictors of diabetes risk?Are there variables that could potentially be dropped due to low variability or lack of relevance?