**Project Description**

According to the National Institutes of Health, Diabetes is a disease that develops when the

blood glucose, commonly known as blood sugar, is too high [1]. Blood glucose, which comes

from the food we eat, is the primary source of energy. To turn glucose into energy, we need

insulin, which is a hormone produced by the pancreas. However, sometimes the body doesn't

produce enough insulin, in which case the glucose remains in the circulation and does not

reach the cells as a result. Having too much glucose in the blood without management might

lead to health problems over time.

The main objective of this project is to build a classifier using machine learning techniques

that can accurately tell us if the person has diabetes or not?

The dataset used for this project is originally from the National Institute of Diabetes and

Digestive and Kidney Diseases.

Task 1: Discovery

1. Read/Load the dataset into R. ( <https://www.kaggle.com/code/jamaltariqcheema/model-performance-and-comparison/data> )

2. Discard 200 random healthy instances.

3. Drop the following attribute:

a. Use all the attributes except: remove Age

4. Display the first 10 instances. After step 2 and 3, how many instances and attributes do you

have?

5. Check the structure of the data and answer the following: is there missing values? How

many? In which columns (if any)? What is the datatype for each attribute?

6. Provide summary statistical description.

7.Visualize the data in three different ways. Describe the plots in detail.

**Task 2: Hypothesis Testing**

1. Write a hypothesis that you would like to verify. For example, there is no statistically

significant difference in the mean of age between the diabetics and healthy people.

2. Test your hypothesis using an appropriate test. Why did you choose this test?

3. What conclusion can you infer from the sample?

**Task 3: Model Building & Evaluation**

1.Build a classifier that can accurately tell you if the person has diabetes or not using:

**A. Regression method**

a. What type of regression will you use? why?

b. Does the data need any preparation for this algorithm? What did you do? Why?

c. Discuss the result.

i. Which variables are important/non-important?

ii. Which variables have a positive/negative impact on the outcome?

iii. Pick two variables and write a complete interpretation/explanation for the

coefficient effects on the outcome.

d. Plot ROC curve. Report AUC.

e. Evaluate the developed model using at least three different metrics.

i. Explain each metric. What is it and how is it calculated?

ii. What value did you get?

iii. What does that mean?

**B. Naïve Bayes**

a. Does the data need any preparation for this algorithm? What did you do? Why?

b. Evaluate the developed model using at least three different metrics.

i. What value did you get?

ii. What does that mean?

**C. Decision Trees**

a. Does the data need any preparation for this algorithm? What did you do? Why?

b. Draw the tree and write at least two rules.

c. What are the variables that best split the data in the first and second level of the tree?

d. Evaluate the developed model using at least three different metrics.

i. What values did you get?

ii. What does that mean?

2.Which method performed the best? Based on what?

**Deliverables:**

**• A report that shows the R code and output for each task along with an explanation for each**

**task.**

**• R source code file**