BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Work Integrated Learning Programmes Division

Machine Learning

S2-23_DSECLZG565

Second Semester, 2023 -24 Assignment 1 – Part 1

Diabetes Risk Prediction Analysis

Instructions for Assignment Evaluation

1. Please follow the naming convention as <Group no>_<Problem Statement No.>.ipynb. Eg – for group 1 with problem statement 1, your notebooks should be named as – Group1_Problem_Statement_1.ipynb.
2. Inside each jupyter notebook, you are required to mention your name, Group detailsand the Assignment dataset you will be working on.
3. Organize your code in separate sections for each task. Add comments to make the code readable.
4. Deep Learning Models are strictly not allowed. You are encouraged to learn classical Machine learning techniques and experience their behavior.
5. Notebooks without output shall not be considered for evaluation.
6. Prepare a jupyter notebook to build, train and evaluate a Machine Learning model on the given dataset. Please read the instructions carefully.
7. Each group consists of up to 5 members. All members of the group will work on thesame problem statement.
8. Each group should upload assignment on LMS in respective locations under ASSIGNMENT Tab. Assignment submitted via means other than through LMS will not be graded.
9. The executed ipynb file with clear subdivision of the codes and brief description of the purpose of respective code needs to be uploaded on LMS. All the executed tablesor graphs and results should be present in the ipynb file.
10. Only two files should be uploaded in canvas without zipping them. One is ipynb fileand other one html output of the ipynb file. No other files should be uploaded.

Problem Statement

Diabetes mellitus (DM) is one of the most devastating metabolic diseases affecting people today. This study aimed to examine DM and the ways in which machine learning algorithms are utilized to detect DM in its early stages. Several things can put you at risk for developing diabetes, including being older, not moving around much, having a history of the disease in your family, having high blood pressure, being depressed or stressed out, eating poorly, and so on.

Dataset: diabetes_risk_prediction_dataset.csv format can be downloaded from drive https://drive.google.com/file/d/1fQhVx2on3WoMCu-R8PPCNvCAWdknTOAR/view?usp=drive_link

Metadata file

- Age (1-20 to 65): Age range of the individuals.
- Sex (1. Male, 2. Female): Gender information.
- Polyuria (1. Yes, 2. No): Presence of excessive urination.
- Polydipsia (1. Yes, 2. No): Excessive thirst.
- Sudden Weight Loss (1. Yes, 2. No): Abrupt weight loss.
- Weakness (1. Yes, 2. No): Generalized weakness.
- Polyphagia (1. Yes, 2. No): Excessive hunger.
- Genital Thrush (1. Yes, 2. No): Presence of genital thrush.
- Visual Blurring (1. Yes, 2. No): Blurring of vision.
- Itching (1. Yes, 2. No): Presence of itching.
- Irritability (1. Yes, 2. No): Display of irritability.
- Delayed Healing (1. Yes, 2. No): Delayed wound healing.
- Partial Paresis (1. Yes, 2. No): Partial loss of voluntary movement.
- Muscle Stiffness (1. Yes, 2. No): Presence of muscle stiffness.
- Alopecia (1. Yes, 2. No): Hair loss.
- Obesity (1. Yes, 2. No): Presence of obesity.
- Class (1. Positive, 2. Negative): Diabetes classification.

1. **Import Libraries/Dataset**

   a. Download the dataset.
   b. Import the required libraries.

2. **Data Visualization and Exploration [1M]**

   a. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   b. Provide appropriate data visualizations to get an insight about the dataset.
   c. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. **Answer without justification will not be awarded marks.**

3. **Data Pre-processing and cleaning [2M]**

    a. Do the appropriate pre-processing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Mention the pre-processing steps performed in the markdown cell.

    b. Apply appropriate feature engineering techniques. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset. Provide proper justification. **Techniques used without justification will not be awarded marks**. Explore a few techniques for identifying feature importance for your feature engineering task.

4. **Model Building [5M]**

    a. Split the dataset into training and test sets. **Answer without justification will not be awarded marks.** [1M]

        i. Train = 80 % Test = 20%

        ii. Also, try to split the dataset with different ratios of your choice.

    b. Build model using Logistic model and decision tree [4 M]

        i. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. Justify your answer.

5. **Performance Evaluation [2M]**

    a. Compare the performance of the Logistic Regression and Decision Tree models using appropriate evaluation metrics.

    b. Provide insights into which model performs better and why. **Answer without justification will not be awarded marks.**