Assignment 7 Report

Mohamed Morshedy| 202100327

Under supervision of:

Dr. Khaled Mostafa

Eng. Gamal Zayed

Eng. Marwa Monier

1. Problem Definition and Motivation:

The task involves conducting an in-depth data analysis of a provided dataset, with the objective of addressing key questions through the exploration of visualizations and relationships among variables. The motivation behind this effort is to extract meaningful insights from the data, which can then contribute to the informed design and evaluation of diverse machine learning models.

2. Importance of the Task:

Comprehending the dataset holds paramount importance in the development of successful machine learning models. Through the formulation of questions, visualization of data relationships, and addressing aspects like cleaning and scaling, we can improve the overall performance of models. The utilization of various algorithms, including KNN, Logistic Regression, SVM, Naive Bayes Classifier, and K-means, enables the exploration of diverse approaches, facilitating a comparison of their effectiveness. This analysis is critical for making informed decisions, recognizing model limitations, and selecting the most appropriate algorithm for the given task.

3. Dataset

Detailed Description of the Dataset:

The dataset titled "ObesityDataSet.csv" consists of various attributes related to individual health and lifestyle factors. Here's an overview based on the initial inspection:

Gender: Categorical (e.g., Male, Female)
Age: Numerical
Height: Numerical (presumably in meters)
Weight: Numerical (presumably in kilograms)
family_history_with_overweight: Categorical (yes or no)
FAVC: Categorical (presumably relates to frequent consumption of high caloric food, yes or no)
FCVC: Numerical (frequency of consumption of vegetables)
NCP: Numerical (Number of main meals)
CAEC: Categorical (Consumption of food between meals)
SMOKE: Categorical (yes or no)

CH2O: Numerical (Consumption of water daily)
SCC: Categorical (Caloric consumption monitoring, yes or no)
FAF: Numerical (Physical activity frequency)
TUE: Numerical (Time using technology devices)
CALC: Categorical (Alcohol consumption)
MTRANS: Categorical (Mode of Transportation)
NObeyesdad: Categorical (Levels of Obesity)
Target Variable: The *NObeyesdad* column seems to be the target variable, indicating different levels of obesity such as "*Normal_Weight*", "*Overweight_Level_I*", and "*Overweight_Level_II*".

Possible Data Types and Range of Values: The numerical columns appear to have values that could represent age in years, height in meters, weight in kilograms, and other lifestyle factors on various scales (e.g., frequency of activities or consumption).

Missing Values and Data Quality: A more detailed analysis is needed to identify missing values or potential issues with data quality such as outliers or inconsistent entries.

Data Pre-processing Methods:

Given the nature of the dataset, several pre-processing steps are recommended:

Handling Missing Data: Check for any missing values and decide on a strategy for handling them (e.g., imputation, removal).

Data Normalization/Standardization: For numerical columns, especially those like height and weight, normalization or standardization might be necessary to bring all variables to a comparable scale.

Encoding Categorical Variables: Convert categorical variables into a format that can be provided to machine learning models. This could involve one-hot encoding or label encoding.

Outlier Detection and Handling: Identify and handle outliers in the dataset to prevent them from skewing the results.

Feature Engineering: Depending on the analysis or model-building objectives, new features could be created from the existing data to better capture the underlying patterns.

Data Splitting: If the dataset is to be used for building predictive models, it should be split into training and testing sets to evaluate the model's performance.

Correlation Analysis: Analyze the correlation between different variables, especially with respect to the target variable, to understand their relationships and importance.

4. Implementation

Tools and Libraries Used

In this project, a variety of tools and libraries were utilized to facilitate the processes of data handling, analysis, visualization, and machine learning model development. Keys among these are:

Pandas and NumPy: Essential for data manipulation and numerical calculations. They are typically used for data cleaning, transformation, and preparation tasks.

Matplotlib and Seaborn: These libraries are used for data visualization, providing insights into the distribution of data, trends, and patterns.

Scikit-Learn: A pivotal tool for machine learning. It offers a range of algorithms for classification, regression, clustering, and model validation techniques.

Additional libraries may include data preprocessing tools, model evaluation metrics, and possibly deep learning frameworks like TensorFlow or Keras, depending on the complexity of the models trained.

Model Training and Validation

This section likely covers the comprehensive process of selecting, training, and validating various machine learning models. Key aspects include:

Model Selection: Discussion on choosing appropriate models based on the problem type (e.g., classification, regression). This could range from simple models like Logistic Regression to more complex ones like Neural Networks.

Hyperparameter Tuning: Detailing the process of optimizing model parameters to improve performance.

Cross-Validation: Employing techniques like k-fold cross-validation to ensure that the model generalizes well to unseen data.

Performance Metrics: Utilizing metrics such as accuracy, precision, recall, F1-score, ROC-AUC for classification tasks, or MSE, RMSE for regression tasks to evaluate model performance.

5. Conclusion

In summary, among the models under consideration:

The Neural Network achieved an impressive accuracy of 98.7%, showcasing its capability to capture intricate data relationships with its potentially more complex structure. KNN, Logistic Regression, SVM, and Naive Bayes Classifier demonstrated accuracies of 81%, 86%, 88%, and 64%, respectively.

When evaluating the optimal model, it's essential to consider various factors. Neural Networks, with their heightened complexity, excel at capturing nuanced patterns in the data, while simpler models like Logistic Regression and Naive Bayes offer interpretability.

In terms of robustness and generalization, Neural Networks, especially with deep architectures, may perform well across diverse datasets featuring complex patterns. Conversely, simpler models may exhibit robustness if their assumptions align closely with the data distribution.

Furthermore, the nature of the data, including the presence of linear or non-linear relationships, significantly influences model performance. Careful consideration of these factors is crucial when selecting the most suitable model for the given task.

Recommendations: experimenting with different models, additional feature engineering, or acquiring more diverse data.