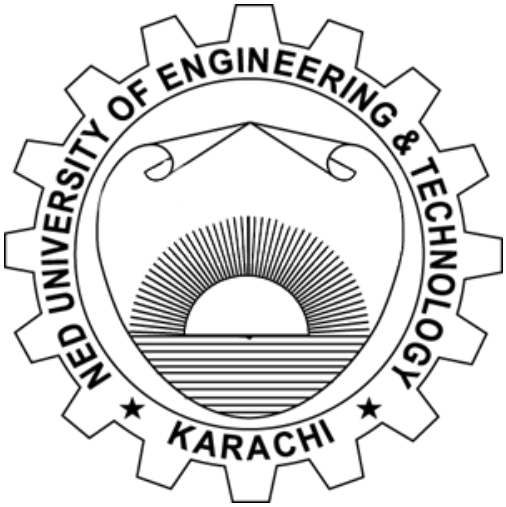

Machine Learning

CS - 324

Project Report

Submitted to	**Miss Mahnoor Malik**
Submitted by	Barira Khan Ghauri CS 21001 Fatima Ali CS 21012 Hasnain Ali Naqvi CS 21035 Syeda Shiza Rizvi CS 21040

Introduction

The objective of this Open Ended Lab was to collect a dataset and create predictive models for it after relevant data preprocessing and EDA. We trained the following classifiers for our dataset, both with Python's Scikit Learn package and using custom implementations.

1. KNN Classifier
2. Decision Tree Classifier
3. Random Forest Classifier
4. Support Vector Machine
5. Logistic Regression
6. Naive Bayesian Classifier

Dataset

Our dataset is one for obesity data, taken from Kaggle. It uses multiple features and one target variable to classify patients into different levels of body weight/obesity.

Following is a brief description about the features and target variable of the data.

id: A unique identifier for each individual in the dataset.

Gender: The individual's gender, indicating whether they are male or female.

Age: The age of the individual, representing their age in years.

Height: The height of the individual, typically measured in meters.

Weight: The weight of the individual, typically measured in kilograms.

family_history_with_overweight: Indicates whether there is a family history of overweight for the individual (yes/no).

FAVC: Stands for "Frequency of consuming high caloric food," representing how often the individual consumes high-calorie foods (yes/no).

FCVC: Stands for "Frequency of consuming vegetables," representing how often the individual consumes vegetables.

NCP: Stands for "Number of main meals," indicating the number of main meals the individual consumes daily.

CAEC: Stands for "Consumption of food between meals," representing the frequency of consuming food between meals.

SMOKE: Indicates whether the individual smokes or not (yes/no).

CH2O: Represents the amount of water consumption for the individual.

SCC: Stands for "Calories consumption monitoring," indicating whether the individual monitors their calorie consumption (yes/no).

FAF: Stands for "Physical activity frequency," representing the frequency of the individual's physical activities.

TUE: Stands for "Time using technology devices," indicating the amount of time the individual spends using technology devices.

CALC: Stands for "Consumption of alcohol," representing the frequency of alcohol consumption.

MTRANS: Stands for "Mode of transportation," indicating the mode of transportation the individual uses.

NObeyesdad: The target variable, representing the obesity risk category of the individual. It has 7 classes: 'Overweight_Level_II', 'Normal_Weight', 'Insufficient_Weight', 'Obesity_Type_III', 'Obesity_Type_II', 'Overweight_Level_I', and 'Obesity_Type_I'.

Model Evaluation

The metrics used for evaluation of the models are provided below.

1. **Precision, Recall, F1-Score, and Support**
    - **Macro Average:** This calculates the metrics independently for each class and then takes the average. It gives equal weight to each class, regardless of its size.
    - **Weighted Average:** This computes the weighted average of metrics, weighted by the number of instances of each class. It's useful when classes are imbalanced.
2. **Accuracy**
3. **Comparison Across Models:**
    - Models like the Support Vector Machine (sklearn) show very high precision, recall, and F1-score across all classes, indicating excellent performance.
    - Decision Tree Classifiers (both sklearn and custom) also perform well, achieving high accuracy and balanced F1-scores across classes.
    - Custom implementations of classifiers (like KNN, Decision Tree, and Logistic Regression) often show comparable or slightly better performance than their sklearn counterparts, due to custom tuning or implementation specifics.

The evaluation suggests that different classifiers have varying strengths. Models like SVMs and Decision Trees generally perform robustly across different metrics. KNN models show good precision but may vary in recall depending on the implementation.

- Logistic Regression, while generally reliable, might vary in performance based on whether it's a custom or sklearn implementation.

- Bayesian Networks, shown in both custom and general evaluations, indicate lower performance metrics compared to other models, suggesting they might not be as suitable for this dataset.s

Conclusion

The following table shows an accuracy comparison between all the models we trained. The Random Forest Classifier fits our dataset the best, giving the highest accuracy score. This is understandable, because RFC works well on large datasets that have features of mixed data types and non-linear relationships. It is often used for disease diagnosis.

MODELS	SKLEARN MODEL	CUSTOM MODELS
KNN CLASSIFIER	0.89	0.94
Decision Tree Classifier	0.95	0.96
Random Forest Classifier	0.97	0.97
Support Vector Machine	0.96	0.75
Logistic Regression	0.86	0.88
Bayesian Network	0.88	0.71