**MACHINE LEARNING LAB 4**

*Name: Vishal Sreekumar*

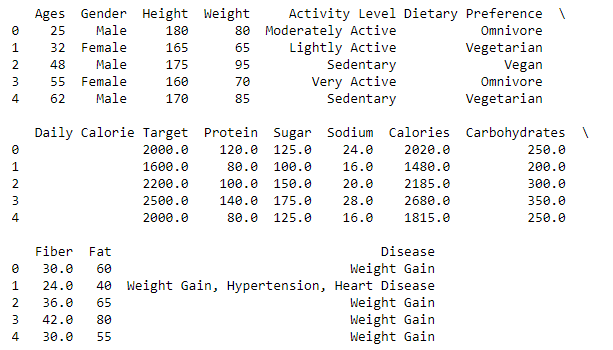
*Regnum: 2241163*

*Class: 5 BCA B*

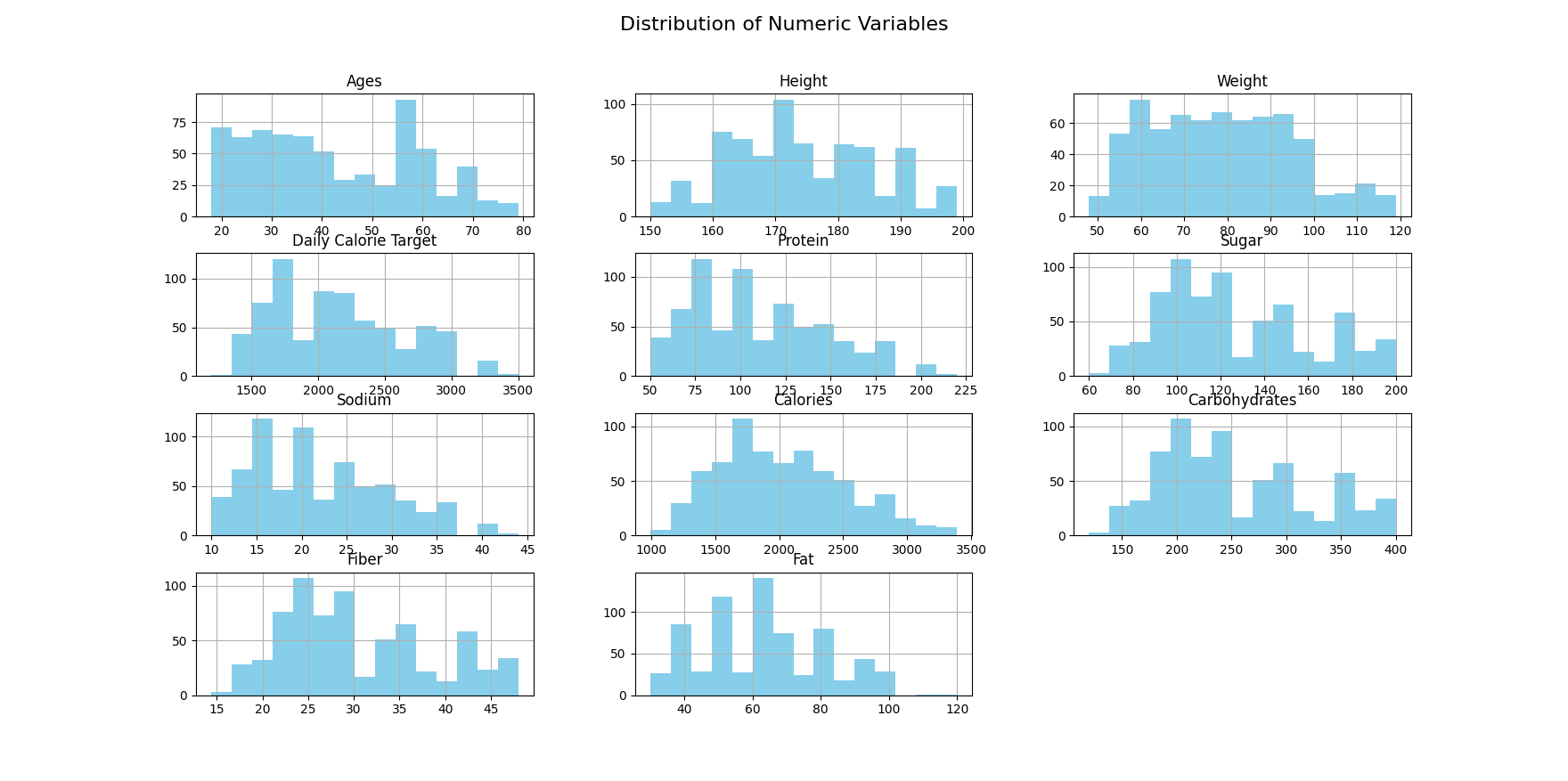
**Introduction**

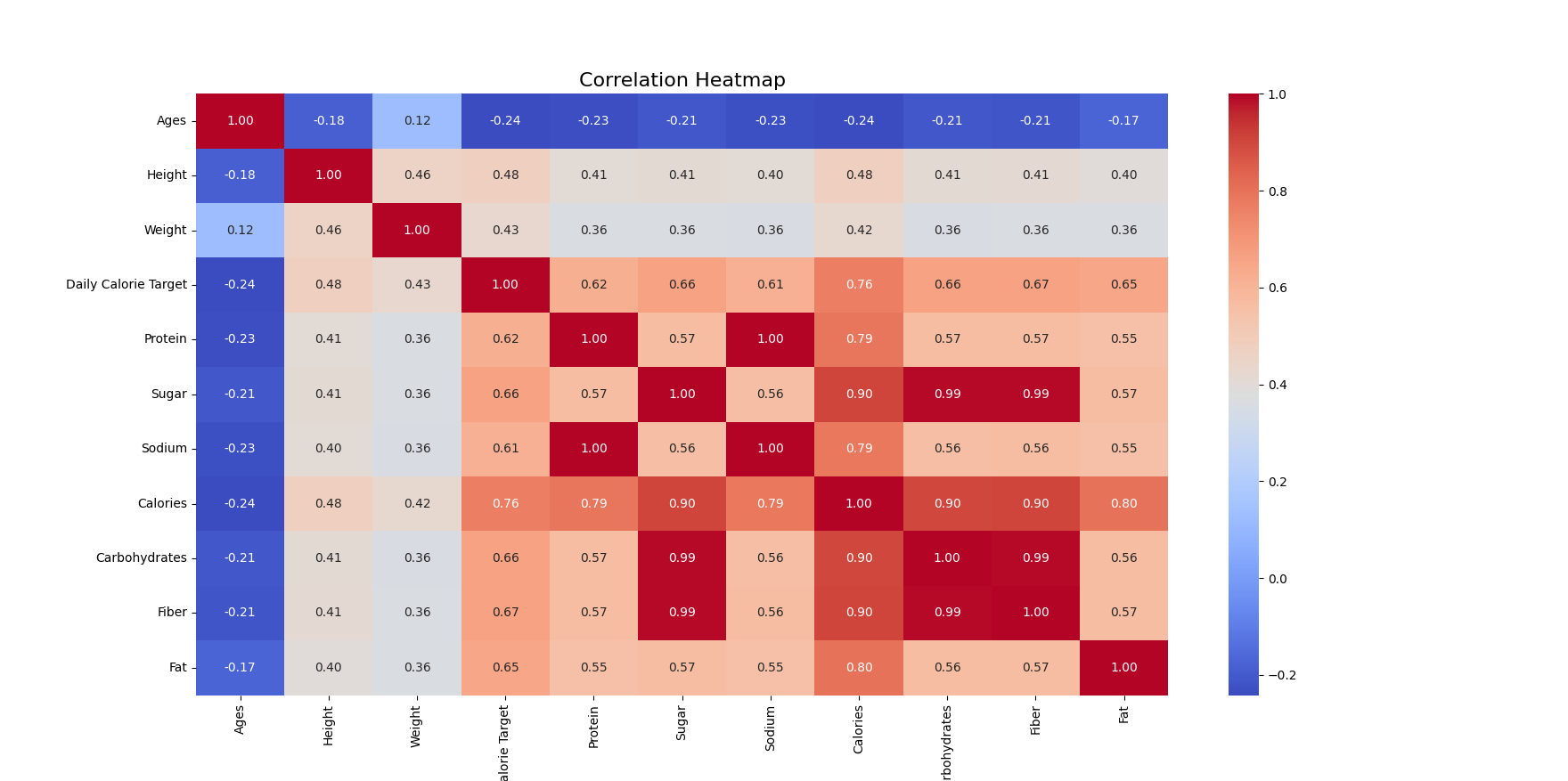
Data analysis and classification models are fundamental components of machine learning pipelines. This project focuses on a systematic approach to data exploration and classification, utilizing two popular Naïve Bayes algorithms: Gaussian and Multinomial Naïve Bayes. The analysis includes exploratory data analysis (EDA) to uncover hidden patterns, training classifiers for predictive modeling, and evaluating their performance using accuracy scores, confusion matrices, and ROC curves. By combining these techniques, the study aims to derive meaningful insights and assess the models' predictive capabilities effectively.

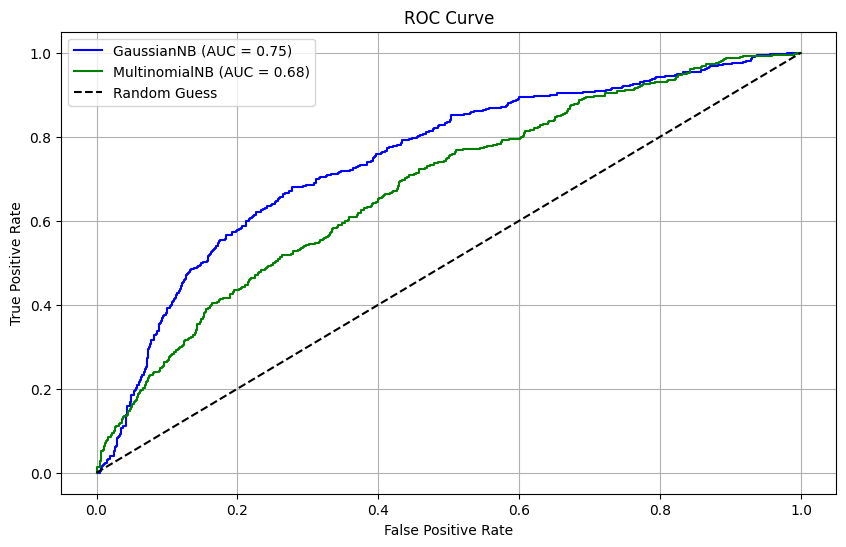
1. **Detailed EDA**:
   * EDA provides a comprehensive understanding of the dataset by exploring its numerical and categorical features. This step highlights data distributions, relationships, missing values, and anomalies, offering a solid foundation for preprocessing and feature engineering.
2. **Train Gaussian Naïve Bayes Classifier**:
   * Gaussian Naïve Bayes is particularly suited for continuous data. It assumes a normal distribution of features and is effective in scenarios where this assumption holds. The training process involves using these properties to compute probabilities for classification.
3. **Train Multinomial Naïve Bayes Classifier**:
   * Multinomial Naïve Bayes is ideal for discrete data, often used in text classification and count data scenarios. It leverages frequency counts for each feature to determine class probabilities.
4. **Check Accuracy Score**:
   * The accuracy score provides a straightforward metric to measure the proportion of correct predictions out of all predictions made by the model, helping to evaluate its overall performance.
5. **Confusion Matrix with Cross Validation**:
   * The confusion matrix summarizes the model's classification results into true positives, false positives, true negatives, and false negatives, while cross-validation ensures that performance metrics are robust and generalizable.
6. **ROC (Receiver Operating Characteristic) Curve**:
   * The ROC curve visualizes the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across different threshold values, providing an overall assessment of the classifier's discriminative power.
7. **Interpretation**:
   * Insights derived from the EDA and classification metrics are used to understand the model's behavior, including strengths, weaknesses, and areas for improvement. Comparing Gaussian and Multinomial classifiers sheds light on which is better suited for the dataset.

****

**Detailed EDA:**



****



**Interpretations:**

1. Accuracy of Gaussian Naive Bayes: 0.79
2. Accuracy of Multinomial Naive Bayes: 0.80
3. Confusion Matrix for Gaussian Naive Bayes:

[[1549 50]

[ 367 35]]

1. Cross-Validation Confusion Matrix for Gaussian Naive Bayes:

[[7680 284]

[1880 158]]

1. Confusion Matrix for Multinomial Naive Bayes:

[[1599 0]

[ 402 0]]

**Conclusion**

The project demonstrates the importance of detailed exploratory analysis and rigorous model evaluation in machine learning. Through EDA, we identified key patterns and prepared the data effectively for modeling. Gaussian and Multinomial Naïve Bayes classifiers were trained, and their performances were evaluated using accuracy scores, confusion matrices, and ROC curves. The results highlight the suitability of different classifiers for various data types, providing actionable insights for future applications. The study emphasizes the importance of combining domain knowledge with analytical tools to achieve accurate and interpretable results.