**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |  |
| --- | --- | --- |
| **Module Title:** | *Data Preparation*   |  | | --- | | *Machine Learning* | |
| **Assessment Title:** | *CA2\_DP\_ML\_HDip\_Lvl8* |
| **Lecturer Name:** | *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Carla Dubehiza Ribera Arizaga |
| **Student Number:** | 2022368 |
| **Assessment Due Date:** | Tuesday, 2 January 2024, 11:59 PM |
| **Date of Submission:** | Tuesday, 2 January 2024, 11:59 PM |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Predictive Analytics in Public Health: A Machine Learning Approach to NHANES Data**

**Introduction**

The ever-increasing volume of data in the modern world presents both a challenge and an opportunity for organizations. With a particular focus on healthcare, the analysis of extensive datasets like the National Health and Nutrition Examination Survey (NHANES) can unearth insights with significant implications for public health policies and individual healthcare decisions. The dataset from NHANES, encompassing a range of health and nutritional parameters from a diverse cross-section of the U.S. population, offers a fertile ground for applying machine learning techniques to predict and classify health outcomes.

**Rationale for Data Preparation Techniques**

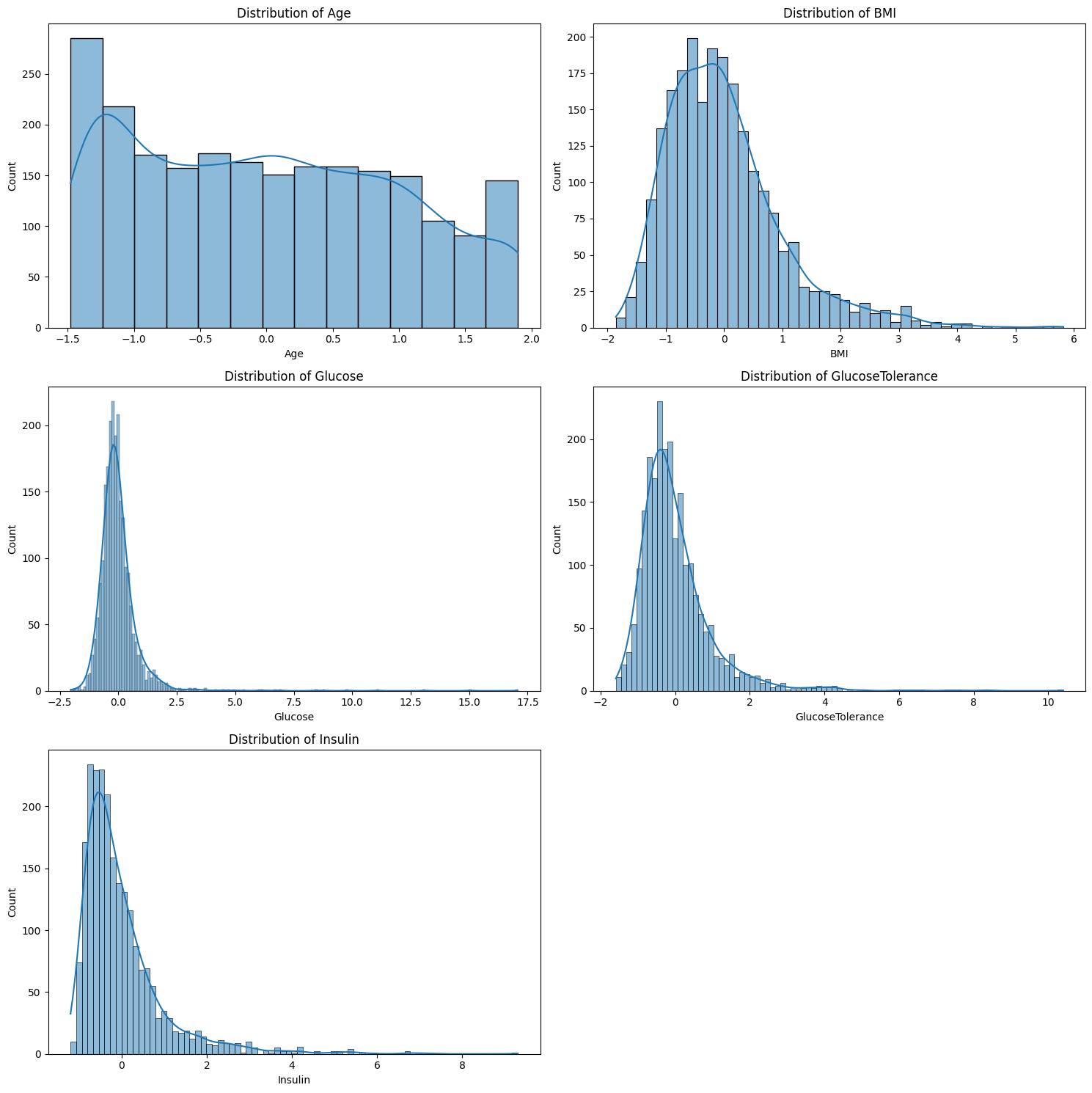
The initial step in the analytical journey involves rigorous data preparation, an indispensable phase that significantly impacts the performance of subsequent machine learning models. The dataset under scrutiny was subjected to a methodical process starting with cleaning, where descriptive column names were introduced for clarity. The renaming of attributes to more understandable terms was an essential step to demystify the data, making subsequent analysis more intuitive.

**Data Cleaning and Exploratory Data Analysis (EDA)**

The primary dataset contained 2278 observations across ten attributes, including age, gender, physical activity levels, BMI, and various blood markers. Initial exploration revealed a dearth of missing values, suggesting a well-maintained and robust data collection methodology by the CDC.

Subsequent histograms (see Image 1: Histograms for Numerical Features) unveiled the distributions of critical numerical features post-scaling. These features showed varying degrees of skewness and kurtosis, which were anticipated given the biological variability inherent in a cross-sectional population sample.

**Image 1: Histograms for Numerical Features**



The gender-based boxplots (see Image 2: Boxplot for Numerical Features by Gender) highlighted differences in distributions of variables like BMI, glucose levels, and insulin between the two genders, which could potentially inform gender-specific health interventions.

**Image 2: Boxplot for Numerical Features by Gender**

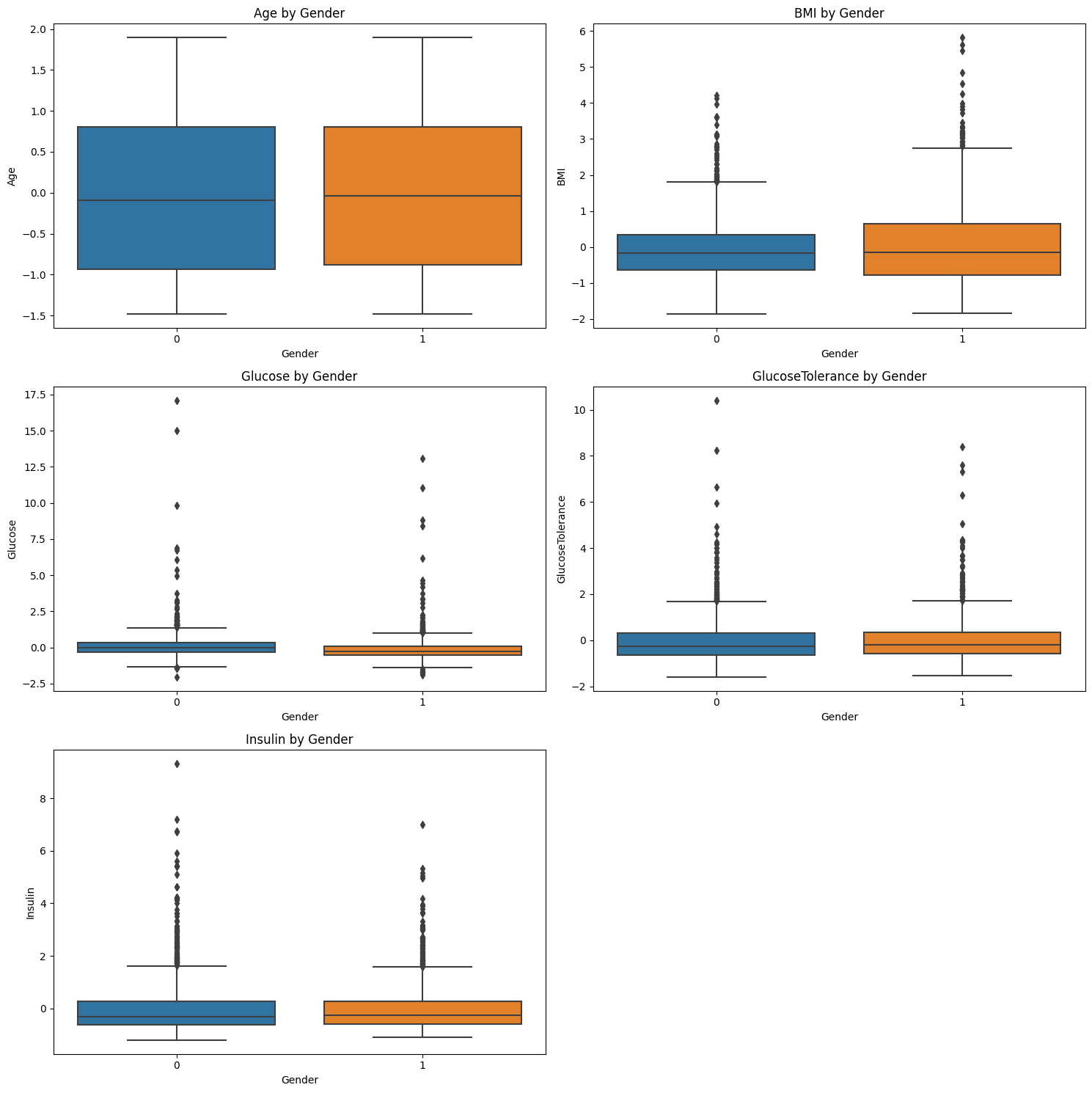
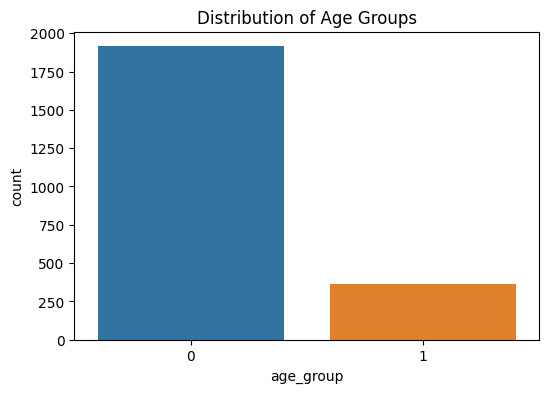


Image 3 (see Image 3: Count Plot for Distribution of Age Groups) illustrates the distribution of age groups, pointing to an imbalance with a higher proportion of adults compared to seniors. This disparity may reflect demographic patterns within the population or could be indicative of the dataset's sampling strategy.

**Image 3: Count Plot for Distribution of Age Groups**



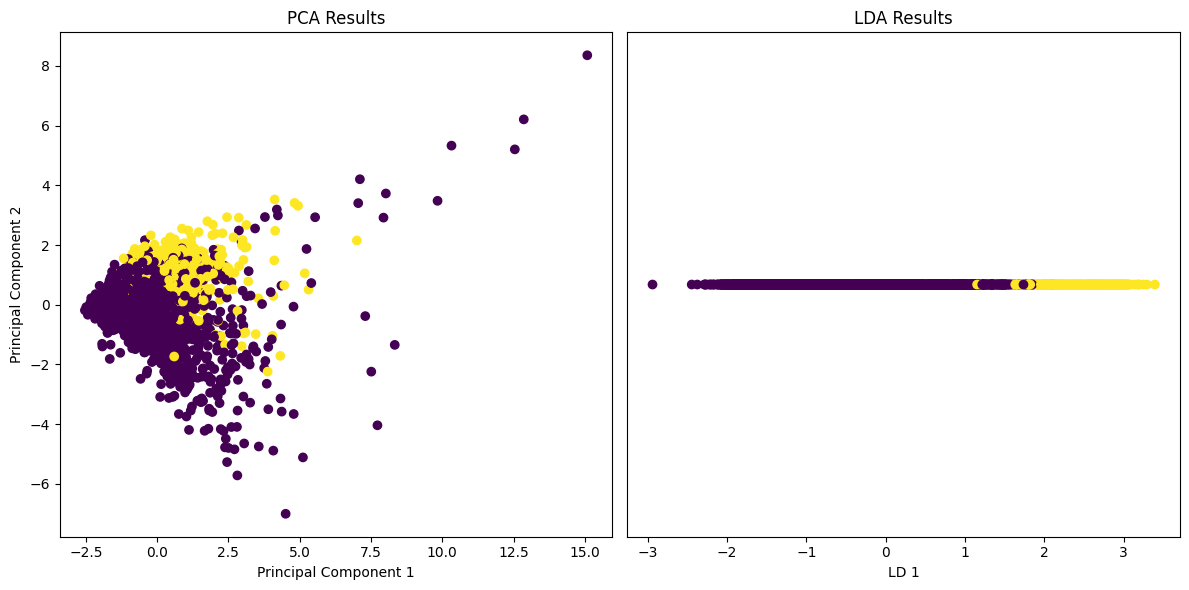
**Feature Engineering**

The encoding of categorical variables, like gender and age group, and the scaling of numerical features, were imperative to normalize the data, rendering it suitable for the application of machine learning algorithms. An innovative step was the creation of a new 'BMI\_Category' feature, classifying individuals into various weight categories based on their BMI - a feature that could potentially enhance the models' predictive capabilities.

**Dimensionality Reduction**

Dimensionality reduction was executed using PCA and LDA to distil the data into its most informative components. The PCA scatter plot (see Image 4: PCA and LDA Results) exhibited a somewhat dispersed cloud of points, indicating variance within the data but less class separability. Conversely, the LDA graph showed a stark demarcation between the age groups, affirming LDA's efficacy in maximizing class distinction.

**Image 4: PCA and LDA Results**



**Machine Learning Models**

Within the realm of machine learning, the NHANES dataset presents a unique opportunity to apply various algorithms that can classify and predict health outcomes. Supervised learning models are particularly suited for this dataset due to its structured format and the clearly defined target variable of age groups. Two models that stand out in their performance and application are the Random Forest Classifier and Logistic Regression.

**Random Forest Classifier**

The Random Forest Classifier is a robust and versatile machine learning algorithm capable of performing both regression and classification tasks. The model operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. This ensemble method is favored for its high accuracy, ability to run in parallel, and immunity to overfitting due to the averaging of multiple trees.

In the analysis of the NHANES dataset, the Random Forest Classifier was optimized through hyperparameter tuning using GridSearchCV. This methodical approach involved iterating over a predefined grid of parameters, which included the number of trees in the forest (n\_estimators), the maximum depth of the tree (max\_depth), and the minimum number of samples required to split a node (min\_samples\_split). The best parameters yielded from this process were a maximum depth of 10, a minimum sample split of 2, and 100 trees.

The model's performance was evaluated using metrics such as precision, recall, and f1-score, with a focus on the accuracy metric, which reached an exceptional score of 1.0. This level of accuracy suggests that the model is highly reliable in classifying individuals into the correct age groups based on their health and nutritional parameters.

However, it's important to consider the context of the dataset when evaluating the model's performance. With a perfect classification score, there is a possibility of data leakage or an overfitting scenario where the model may not generalize well to unseen data. In practice, a perfect score is often scrutinized and subjected to further validation to ensure the model's robustness.

**Logistic Regression**

In contrast to the Random Forest Classifier, Logistic Regression is a more simplistic approach that models the probability of a binary outcome based on one or more predictor variables. The strength of Logistic Regression lies in its interpretability; the coefficients of the model can be directly related to the odds ratios for the respective predictors, offering clear insight into the relevance of each feature.

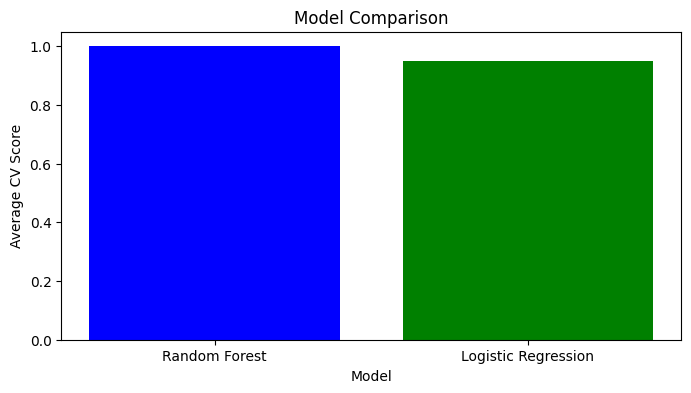
For the NHANES dataset, Logistic Regression was applied, and its performance was evaluated, demonstrating a commendable average cross-validation score of approximately 0.952. This score, while not as high as the Random Forest Classifier, is still indicative of a strong model, especially considering the interpretability and simplicity of Logistic Regression.

The classification report for Logistic Regression showed a slightly lower precision and recall for predicting the senior age group compared to the Random Forest. Specifically, the model achieved a precision of 0.99 and a recall of 0.99 for the non-senior group, and a precision and recall of 0.97 for the senior group. Despite the slight decrease in metrics, these results are still notable and reinforce the model's utility as a predictive tool in public health settings.

The comparison between the two models (as seen in Image 5: Model Comparison) highlights the trade-off between complexity and interpretability that often arises in machine learning. While the Random Forest Classifier may offer higher accuracy, Logistic Regression provides greater simplicity and ease of understanding, which can be crucial in certain applications, such as clinical settings where explanations are necessary for decision-making.

In conclusion, both models demonstrated significant potential in leveraging the NHANES dataset to inform health outcomes. The Random Forest Classifier showcased exceptional accuracy, and Logistic Regression offered valuable insights with its interpretability. These findings contribute to the broader discourse on the application of machine learning in healthcare and underscore the importance of model selection based on the specific needs of the analysis.

**Image 5: Model Comparison**



It is essential for further research to explore the implications of these findings and to validate the models against new and unseen data. The continuous evolution of machine learning algorithms and their applications in health data analysis will undoubtedly reveal new insights and strategies for public health advancements.

**Conclusions**

The analytical foray into the NHANES dataset culminated in the development of two predictive models, each with its unique strengths. The Random Forest Classifier emerged as the superior model in terms of raw performance metrics. However, the Logistic Regression model's high interpretability and the substantial cross-validation score cannot be overlooked, presenting a compelling case for its use in scenarios where model transparency is paramount.

These models, armed with the insights drawn from the NHANES dataset, could potentially guide healthcare professionals in crafting targeted interventions and preemptive measures. Moreover, the findings underscore the importance of considering demographic discrepancies, such as the uneven representation of age groups, in the planning and execution of public health strategies

**References**

1. Lopez-Jimenez, F., Attia, Z., Arruda-Olson, A. M., Carter, R., Chareonthaitawee, P., Jouni, H., ... & Kapa, S. (2020). Artificial intelligence in cardiology: Present and future. Mayo Clinic Proceedings, 95(5), 1015-1039. This study discusses the use of artificial neural networks in predicting hypertension using NHANES data, highlighting the importance of selecting relevant features for developing classification models.
2. Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28(3/4), 321-377. Hotelling's foundational work on Canonical Correlation Analysis (CCA) provides a statistical method to understand and explore the relationship between two sets of variables, relevant to the NHANES dataset's multiview analysis.
3. Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10(1), 1-11. This paper implements a Support Vector Machine model for diabetes classification using the NHANES dataset, which could inform the selection of features and the development of prediction models.
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). The XGBoost algorithm, renowned for its performance in classification tasks, can be applied to the NHANES dataset for predictive analysis, as suggested by the successful application of similar methodologies in healthcare data.
5. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317-1318. This reference provides a broader perspective on the implications and applications of machine learning in healthcare, which is pertinent to the analysis of the NHANES dataset for predictive health modeling.

Word Count

The word count for each section is as follows:

* Report Carla Ribera: 1322 words

Github Repository

https://github.com/CarlaDubehizaRibera/CARLA-RIBERA\_CA2ML-DATA