Assessing Heart Disease Risk with Everyday Body and Behavioral Data

Brown Data Science Institute
Yicheng Lu
https://github.com/luananc/HeartDiseasePrediction

Introduction

Purpose

Heart disease remains a leading cause of death worldwide [1], with its development closely related to body characteristics and daily behaviors [2]. Early detection and intervention of the disease is crucial in reducing disease fatalities. This project empowers individuals to assess their risk of heart disease using readily available metrics, such as body measurements and daily consumption behaviors. Such self-assessments encourage the identification of the necessity of further formal medical evaluations, potentially leading to timely and life-saving interventions.

Dataset, target variables, and features

The CDC's Behavioral Risk Factor Surveillance System (BRFSS) is an ongoing telephone survey monitoring health-related behaviors and conditions in the United States. The dataset used for this project is a subset of the BRFSS sourced from Kaggle, containing 308,854 entries and 19 columns [3].

This project focuses on 'Heart Disease', the binary ('No': 0, 'Yes': 1) target variable, indicating the presence or absence of heart disease. The model utilizes 18 variables: 1) categorical variables including 'Checkup', 'Exercise', 'Skin_Cancer', 'Other_Cancer', 'Depression', 'Diabetes', 'Arthritis', 'Sex', and 'Smoking_History'; 2) ordinal variables including 'General_Health' and 'Age_Category'; 3) numerical variables such as 'Height_(cm)', 'Weight_(kg)', 'BMI', 'Alcohol_Consumption', 'Fruit_Consumption', 'Green_Vegetables_Consumption', and 'FriedPotato_Consumption'. The dataset shows a significant imbalance, which will be elaborated on in the EDA part.

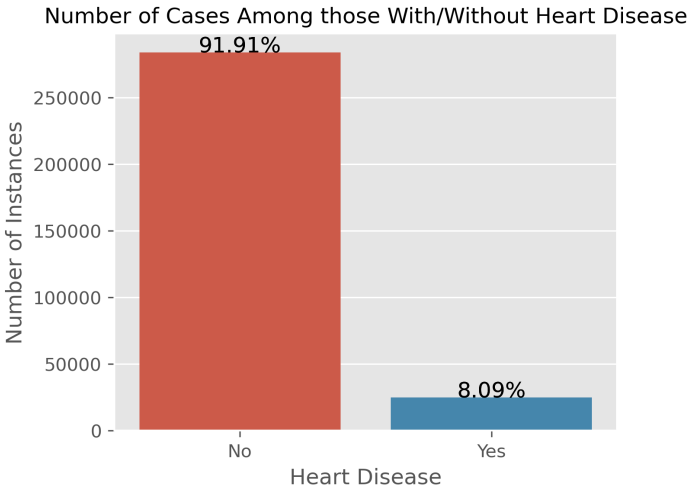
Previous Work

The dataset is widely used on Kaggle. Many Kaggle notebooks tackle its imbalance with SMOTE, achieving high metric scores such as an AUC of 0.94, precision of 0.92, and recall of 0.95 [4]. However, considering the downsides of the SMOTE approach, including artificially altering the data, this project addresses the imbalance by utilizing weights during model training. The final model of this project attains a recall of 0.92, comparable to previous work. However, the precision is notably lower than the SMOTE-adjusted models.

EDA

Target Variable

The target variable used in this project is the binary categorical variable "Heart Disease," which records whether or not heart disease exists in an individual.

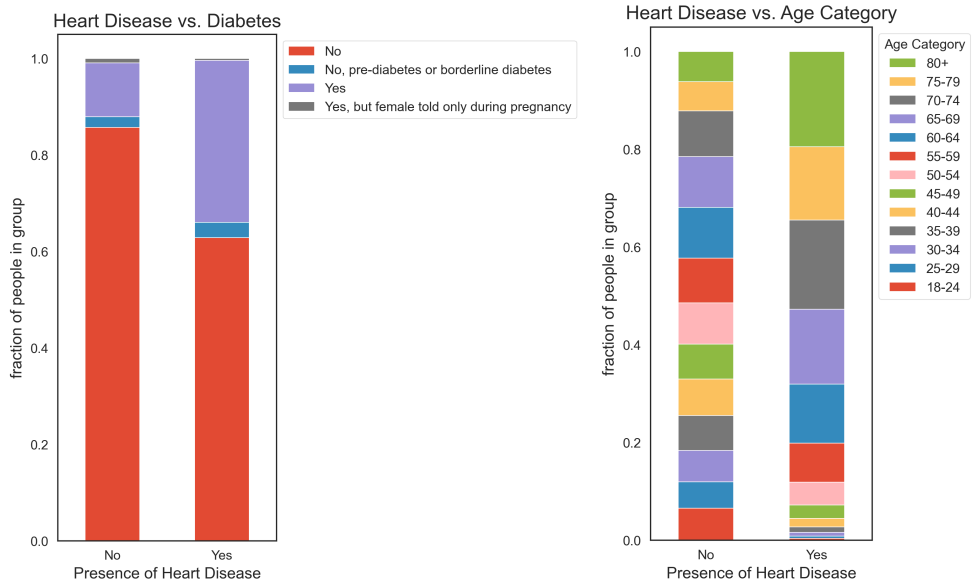


<Fig 1: Visualization of imbalance target variable>

The dataset in this project is imbalanced, as shown in Figure 1, with the positive class (individuals with heart disease) representing only 8% of the whole set, indicating the need to utilize methods like stratification and weights in later steps.

Feature Analysis

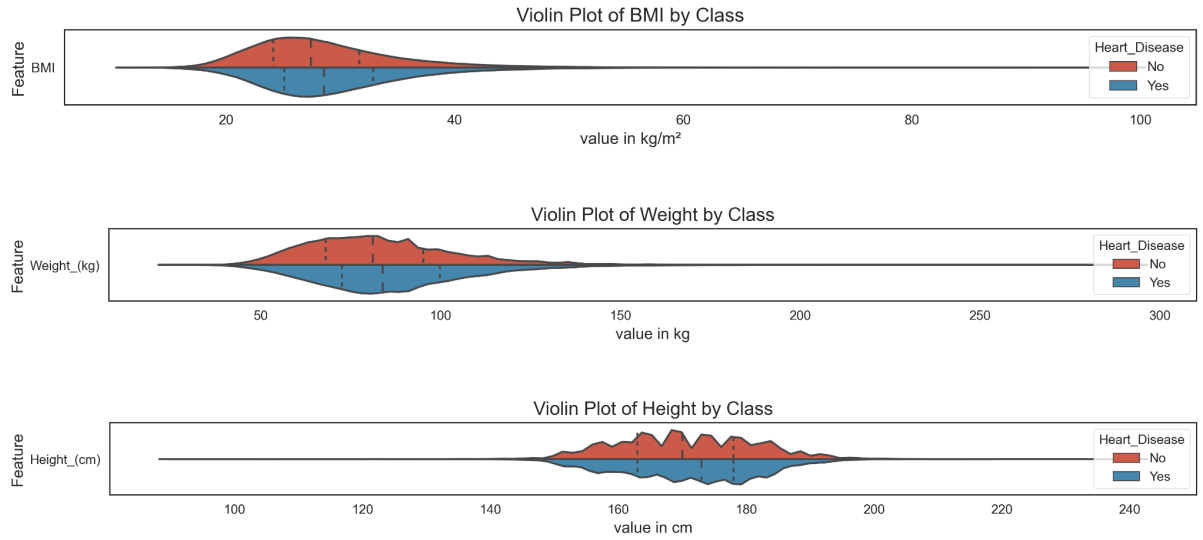
I utilized .value_counts() for each categorical variable to count the unique values within each category.



<Fig 2: Sample categorical variable visualizations>

Figure 2 is a sample of stacked bar plots, illustrating how categorical variables are visualized. The plots show the distribution variations across subcategories within each class for each categorical feature. The plot on the left details the number of subcategories within the 'diabetes' feature and their proprotions. On the right, the stacked bar plot represents the ordinal feature 'age_category' and differences in the distribution of each subcategory across classes. Both plots demonstrate that the distribution of related subcategories varies significantly between individuals with and without heart disease and explain how columns will be encoded in preprocessing.

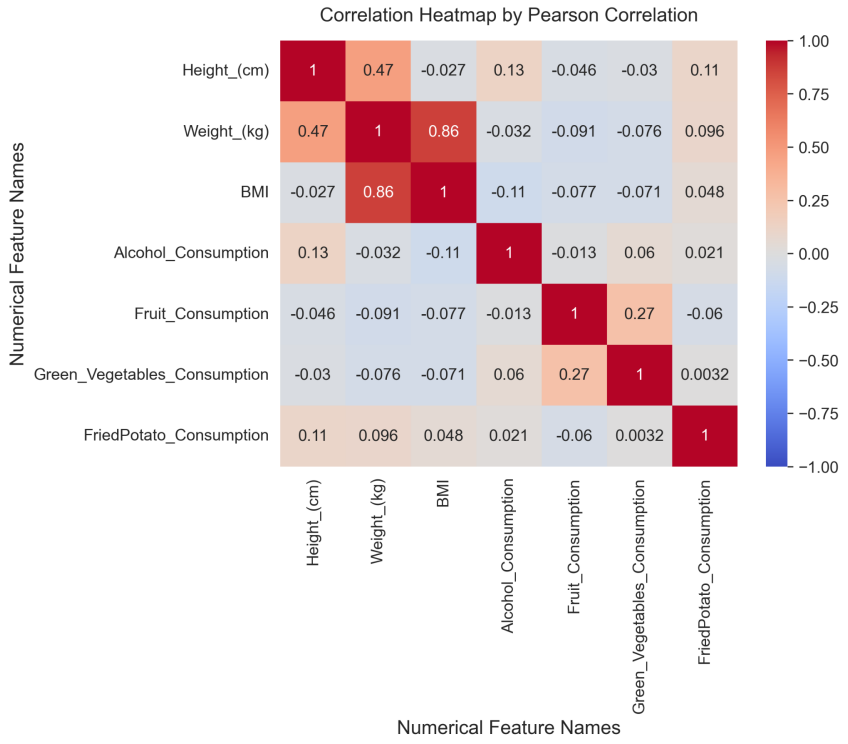
I applied .describe() for the numerical variables to compute mean, minimum, maximum, and standard deviation to understand each variable. Violin plots were generated to compare the distributions of numerical variables segmented by class visually.



<Fig 3: Sample categorical variable visualization>

Figure 3 illustrates the distribution of body measurement data for individuals with and without heart disease. The overall distribution pattern is consistent between the two classes. However, there is a subtle but noticeable shift to the right for those diagnosed with heart disease, suggesting a higher average in the measurements within this group.

A correlation heatmap of all the numerical variables is created to see the correlation between each of the features.



<Fig 4: Correlation Heatmap of numerical features>

Figure 4 indicates that all other features exhibit relatively low correlations besides BMI and weight. The correlation heatmap employs the Pearson correlation coefficient, which measures linear correlation; The nature of how BMI is calculated using the formula, $BMI = weight(kg)/[height(cm)]^2$, explains the strong linear correlation of 0.86 between weight and BMI. Hence, the lack of a strong correlation between height and BMI is reasonable. However, considering that weight could indicate certain body traits, I have decided to retain all features for further analysis.

Methods

Data Preprocessing

Data Splitting

The dataset is split into train, validation, and test sets using a 60-20-20 approach. Due to the imbalanced nature of the dataset, the `train_test_split` method is applied with stratification by 'y' to maintain the ratio of the target variable in the split subsets consistent with its ratio in the entire dataset.

Data Preprocessing

All original data points were preserved without any feature engineering, and the dataset did not contain any missing values. A preprocessor was established, including OneHotEncoder for categorical features, OrdinalEncoder for ordinal features, and StandardScaler for numerical features. This preprocessor is designed to be used in GridSearchCV to fit and transform the X_train set and transform validation and test sets.

ML Pipeline

Four distinct machine learning algorithms were trained: Logistic Regression, Random Forest Classifier, XGBoost Classifier, and SVM Classifier. A unique hyperparameter grid was assigned to each model for exploration using GridSearchCV. (Table 1)

Logistic Regression	Penalty: ‘elasticnet’ l1_ratio': [0, 0.25, 0.5, 0.75, 1] C: [0.01, 0.1, 1, 10, 100] class_weight: [None, 'balanced']
Random Forest Classifier	max_depth: [1, 3, 10, 30, 100] max_features: [0.25, 0.5, 0.75, 1.0] class_weight': [None, 'balanced', 'balanced_subsample', {0: 1, 1: 10}, {0: 1, 1: 15}]
XGBoost Classifier	n_estimators: [100,200, 500], max_depth: [1, 3, 10, 30], gamma: [0.1, 1, 10], learning_rate: [0.01], scale_pos_weight: [1, 11, 12]
SVM Classifier	C: [0.01, 0.1, 1] gamma: [0.001, 0.01] svc__class_weight: ['balanced']

<Table 1: Hyper-parameter Grid>

Considering the imbalanced nature of the dataset, weights were tuned in the training of each model. Parameters are chosen to prevent overfitting or underfitting while maximizing performance on the dataset used. For Logistic Regression: the penalty was set to 'elasticnet', with l1 ratios spanning from 0 to 1 to save time while considering both L1 and L2 regularizations individually.

Training Process

Due to the imbalanced nature of the dataset and the model's intended purpose – not for strict medical diagnosis but to predict heart disease based on easily accessible factors as an initial indicator – it is crucial to minimize false negatives. The model aims to inform individuals whether further medical examination might be necessary. Therefore, recall was chosen as the metric to measure the model's performance to ensure that no potential cases of heart disease are unpredicted, emphasizing the correct identification of all positive cases.

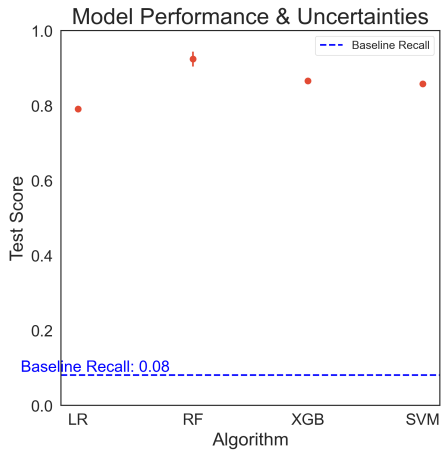
Each algorithm underwent training with five different random states, beginning with the data-splitting process. A pipeline was constructed for Logistic Regression, Random Forest Classifier, and SVM Classifier that combined the preprocessor with the respective model. This pipeline was then fed into GridSearchCV. A predefined split was utilized in GridSearchCV to ensure cross-validation was conducted on the prepared validation set instead of employing a KFold strategy. For the XGBoost Classifier, the training, validation, and testing datasets were manually preprocessed using the established preprocessor. Since the SVM Classifier is computationally expensive, the model used in this project is based on 40% of the entire dataset that is split from the original dataset with stratification.

The best model, associated train, validation, and test sets were selected for each algorithm based on the highest test recall score from five fixed random states. This method ensures the reproducibility of the result, which is crucial in evaluating the uncertainties in recall due to data splitting and the inherent randomness in methods like Random Forest.

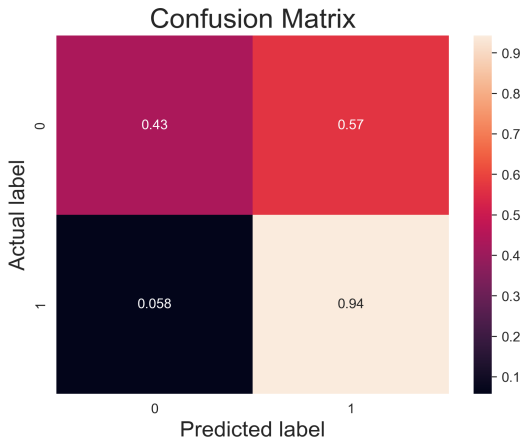
Results

Model	Mean Test Recall Score	Standard Deviation	Number of std above the baseline
Logistic Regression	0.79	0.008	88.18
Random Forest Classifier	0.92	0.02	42.37
XGBoost Classifier	0.87	0.004	182.74
SVM Classifier	0.86	0.005	144.96

<Table 2: Model Training Results>



<Fig 5: Model Performance and Unvertainties>



<Fig 6: Test set confusion matrix using Random Forest>

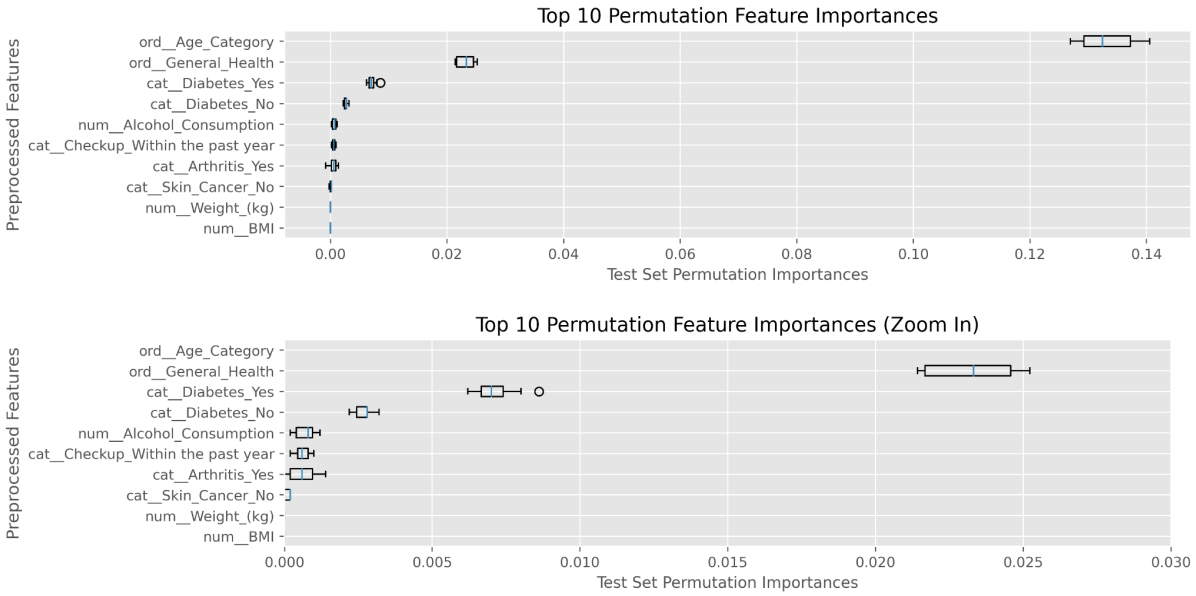
Table 2 presents the mean values for the best model of each Machine Learning algorithm. The baseline recall for the entire dataset is defined by the percentage of class 1 (positive class) in the dataset. All models outperformed this baseline, with the Random Forest Classifier as the best overall performer based on the mean recall score of 0.92 (Figure 5). According to the confusion matrix (Figure 6), this model correctly predicts 94% of the positive cases. However, the model only predicts 43% of the actual class 0 (negative class), indicating a trade-off between recall and precision.

Global Feature Importance

To understand which features are most influential in the best-performing model, the following three metrics were used to evaluate the global feature importance.

Permutation Importance

This approach determines feature importance by observing how much the model's performance decreases when each feature is randomly shuffled; the greater the decrease, the more important the feature.

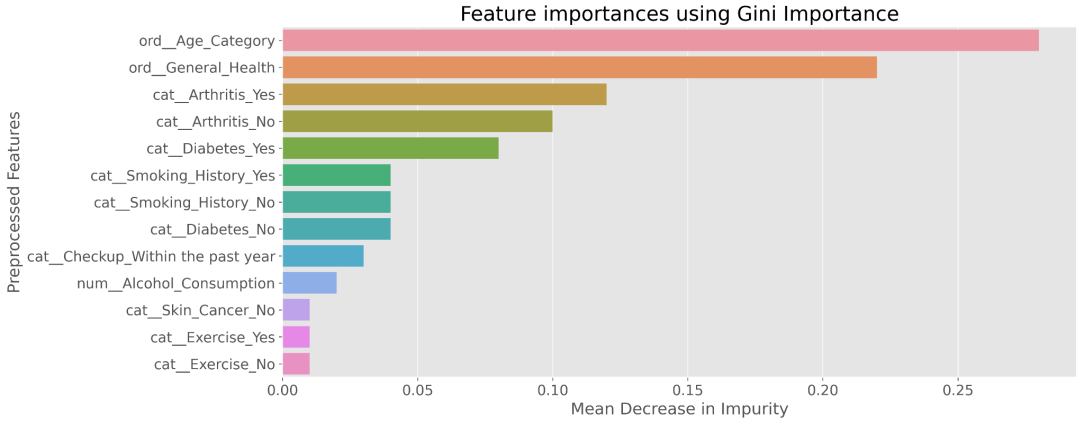


<Fig 7:Top 10 Permutation Feature Importance>

Figure 7 shows that age category, general health, and the presence of diabetes in an individual are determined as the most critical features by permutation importance.

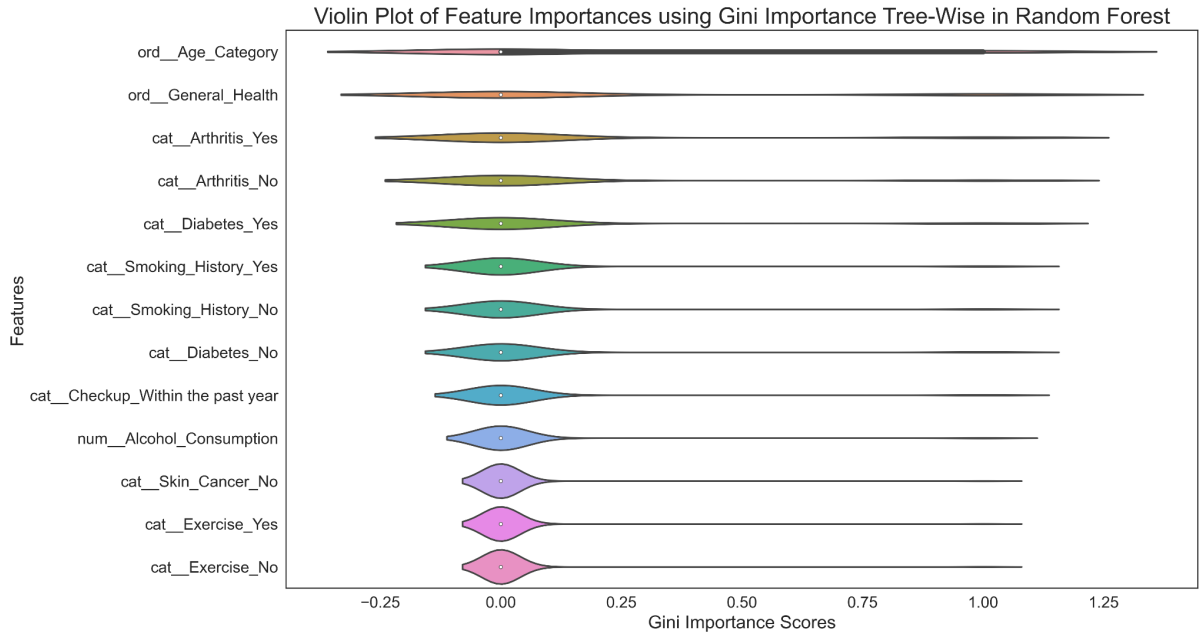
Gini Importance

This approach is specific to the random forest approach, measuring feature importance based on its contribution to reducing uncertainty or impurity in node splits.



<Fig 8:Top 10 Features with Gini Importance>

Figure 8 indicates that, age category, general health, and the presence of arthritis in an individual are determined by Gini importance as the most essential features.

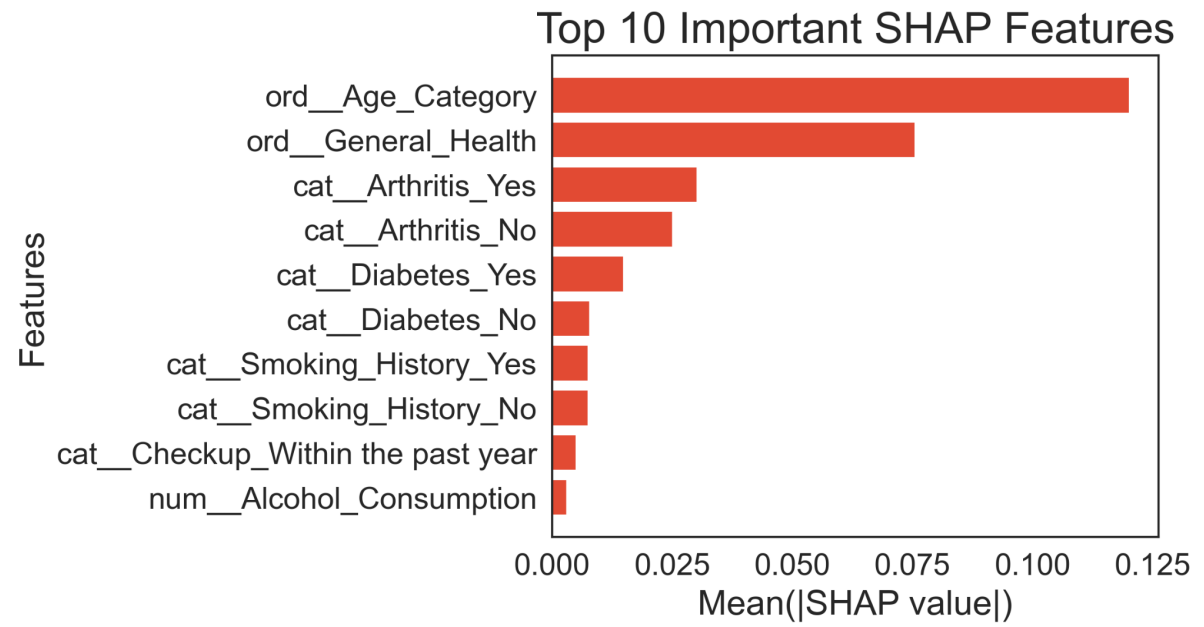


<Fig 9: Standard Deviation of Gini Importance Scores Across Trees>

Figure 9 shows that feature importances across individual trees in the Random Forest model reveal significant variation. The standard deviations suggest the randomness of the algorithm but also the need for more consistency in how features contribute to the model's predictions from one tree to another.

SHAP Global Importance

This approach uses the average absolute SHAP values across all instances for each feature to quantify its overall impact on the model's predictions.



<Fig 10: Top 10 Shap Importance Features>

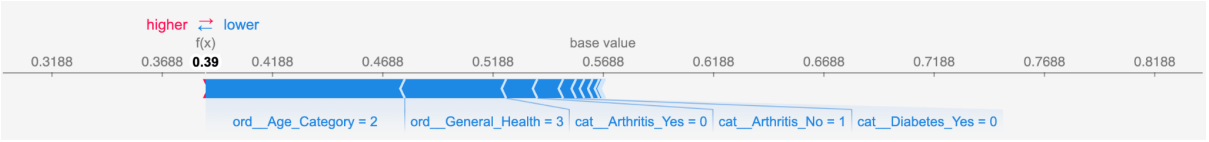
Figure 10 indicates that, age category, general health, and the presence of arthritis in an individual are among the most important features determined by the SHAP approach.

Age category, general health, and the presence of arthritis and diabetes consistently ranked highest in importance metrics, indicating their stable and significant contribution to the model's predictions. Some features,

including body measurement data and the habit of exercise, are sometimes ranked the least important in global feature importance, which is quite surprising.

SHAP Local Feature

SHAP local importance explains the contribution of each feature to the prediction of an individual instance. The following is a prediction on the point with index = 20000, where the true label is 0.



<Fig 11: SHAP local feature Explanation of Class 1>

In this example, the base value of 0.57 reflects the model's default prediction for the target variable being in class 1 in the absence of feature influences. Figure 11 demonstrates that age category, general health, and the presence of arthritis are major contributing features, consistent with the global feature importance analysis. They push the prediction negatively from the base value to a 0.39 model output, indicating the model's strong confidence in assigning this data point to class 0.

Outlook

To enhance and provide a more comprehensive view of this project, an initial step is to train the SVM Classifier using the entire dataset to give a more holistic and parallel assessment of the performance of all models used in this project.

Specifically for the chosen Random Forest Classifier, the comparatively large standard deviation of the test and Gini importance scores suggests that the model's performance is unstable. Thus, a broader range of hyperparameters could be tuned to enhance the model's stability.

The global feature importance analysis indicates that some features in the dataset contribute minimally to the model's output. Given that the BRFSS dataset contains numerous other features, selecting a subset of more influential factors based on background knowledge and literature reviews regarding heart disease could enhance the model's performance and produce more meaningful prediction results.

References

[1] World Health Organization. (2020, December 9). WHO reveals leading causes of death and disability worldwide: 2000-2019. Retrieved from https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019
[2] World Health Organization. (2021, June 11). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[3] Alphiree. (n.d.). Cardiovascular Diseases Risk Prediction Dataset. Retrieved from https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset
[4] Faheem, W. (n.d.). Cardiovascular Disease Prediction (AUC 0.93). Retrieved from https://www.kaggle.com/code/waleedfaheem/cardiovascular-disease-prediction-auc-0-93