Aalto University
School of Business

A Predictive Analysis of Cardiovascular Disease Risks

Group 16: Duy Hoang, Duong Mai, Thu Nguyen, Ngan Do, Matias Sairanen, Vy Dang
20.10.2024

Introduction & Objective

Cardiovascular disease is a serious health threat to anyone which requires up-to-date and as close to accurate as possible diagnosis. With increasing awareness of well-being, the demand in the healthcare sector for detailed health analysis and diagnosis is emphasized. In a business setting, firms specializing in health data analytics can provide insights to other companies or customers using predictive models on cardiovascular risk. Therefore, the business problem that we want to address is the need for accurate and actionable cardiovascular risk predictions to improve health outcomes and reduce costs.

By analyzing data collected by the Behavioral Risk Factor Surveillance System (BRFSS) in 2021 for the CDC and developing predictive models, we aim to refine an early disease identification capability to prevent customers from suffering any fatal accidents. Therefore, the first party to benefit from the model is the customers themselves. Moreover, from the perspective of a company in the medical sector, it is of significant value to be able to detect potential customers and provide personalized service.

Data & Methodology

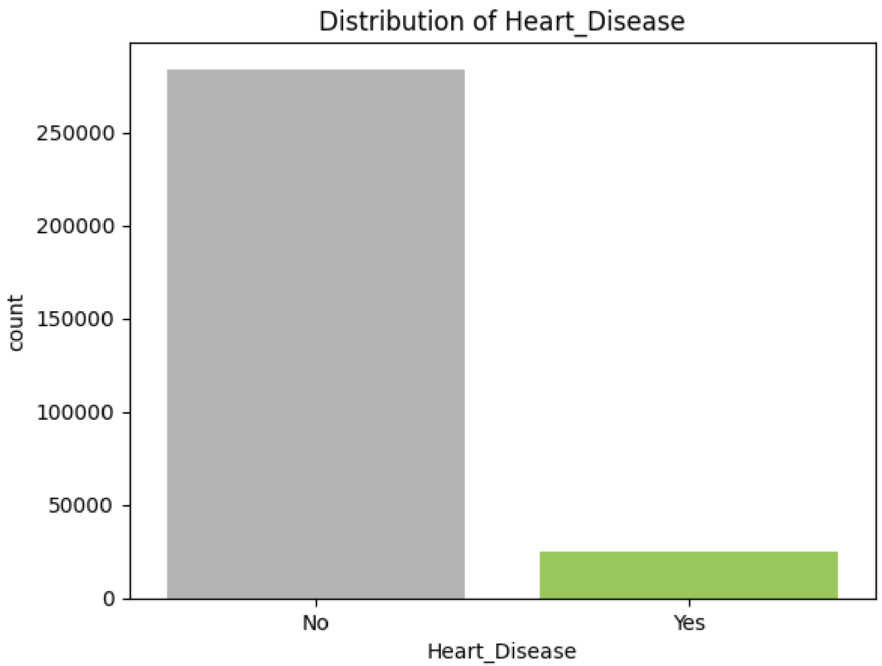
The dataset consists of **308854 health records** with **19 features** that may influence cardiovascular disease such as General Condition, Exercise, Age, Other Cancer, Food Consumption, and so on.

	mean	std	25%	50%	75%
Height_(cm)	170.615249	10.658026	163.00	170.00	178.00
Weight_(kg)	83.588655	21.343210	68.04	81.65	95.25
BMI	28.626211	6.522323	24.21	27.44	31.85
Alcohol_Consumption	5.096366	8.199763	0.00	1.00	6.00
Fruit_Consumption	29.835200	24.875735	12.00	30.00	30.00
Green_Vegetables_Consumption	15.110441	14.926238	4.00	12.00	20.00
FriedPotato_Consumption	6.296616	8.582954	2.00	4.00	8.00

	count	unique	top	freq
General_Health	308854	5	Very Good	110395
Checkup	308854	5	Within the past year	239371
Exercise	308854	2	Yes	239381
Heart_Disease	308854	2	No	283883
Skin_Cancer	308854	2	No	278860
Other_Cancer	308854	2	No	278976
Depression	308854	2	No	246953
Diabetes	308854	4	No	259141
Arthritis	308854	2	No	207783
Sex	308854	2	Female	160196
Age_Category	308854	13	65-69	33434
Smoking_History	308854	2	No	183590

Statistics Summary of Variables

Data cleaning:
No missing values
No features were removed initially
Categorize Age to 5 groups
Set Heart_Disease as the target



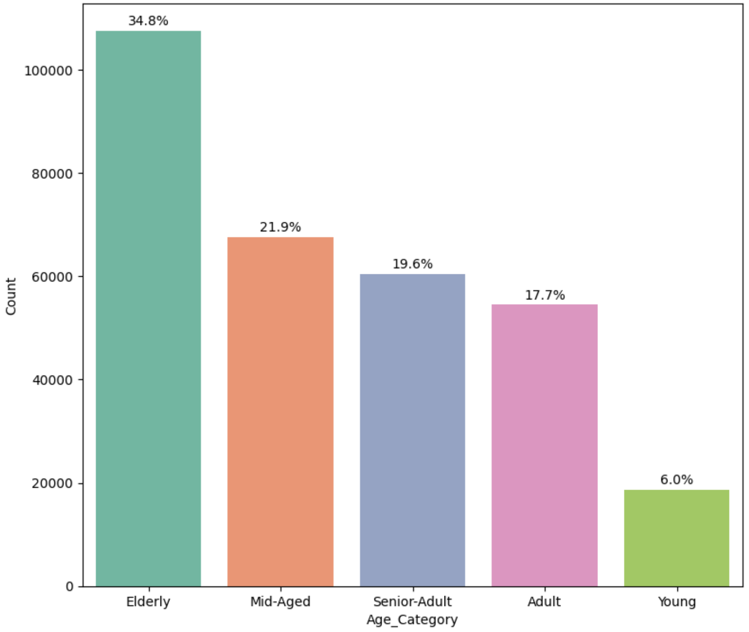
Distribution of Heart Disease

Exploratory Data Analysis – Univariate Analysis

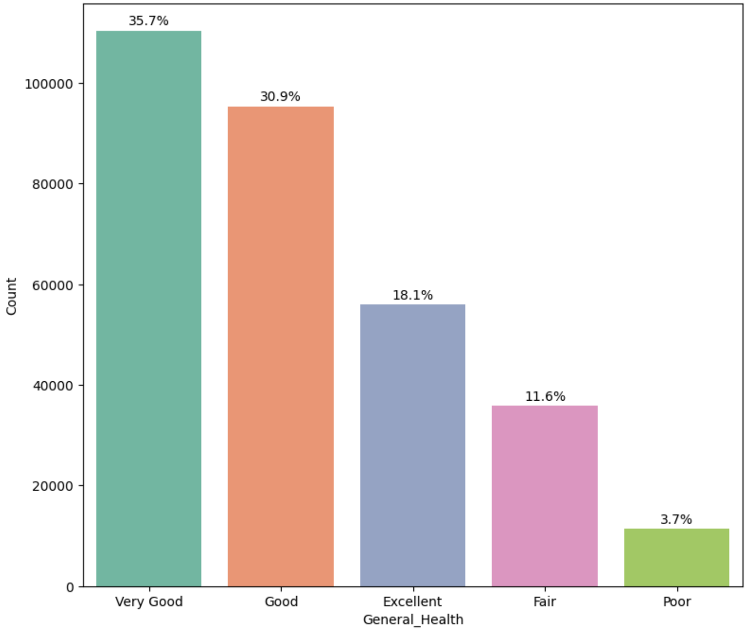
56.7% of the residents surveyed are middle-aged and elderly.

The majority of people have very good general health and check-ups within the past year.

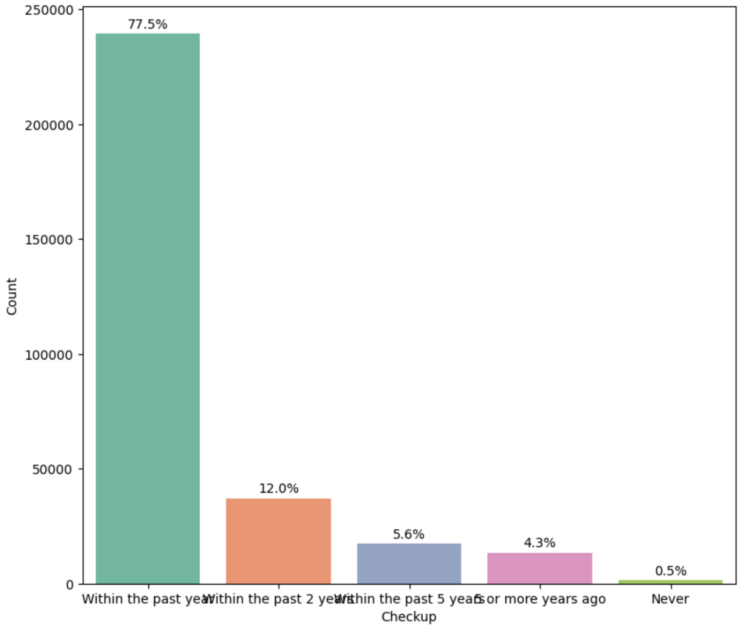
Approximately 20% do not exercise or have depression, and 10% have either skin or other cancer.



Distribution of Age



Distribution of General Health



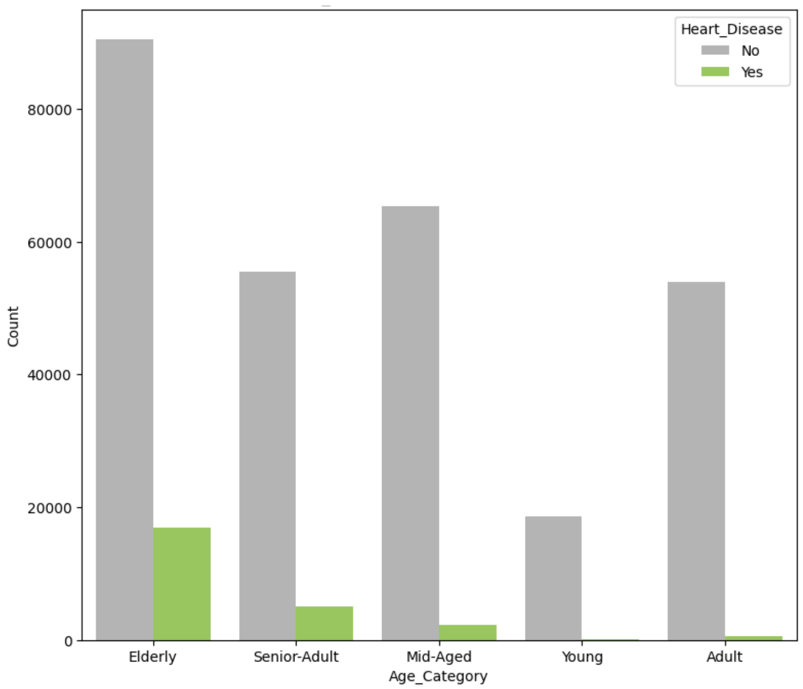
Distribution of Checkup

Exploratory Data Analysis – Bivariate & Multivariate Analysis

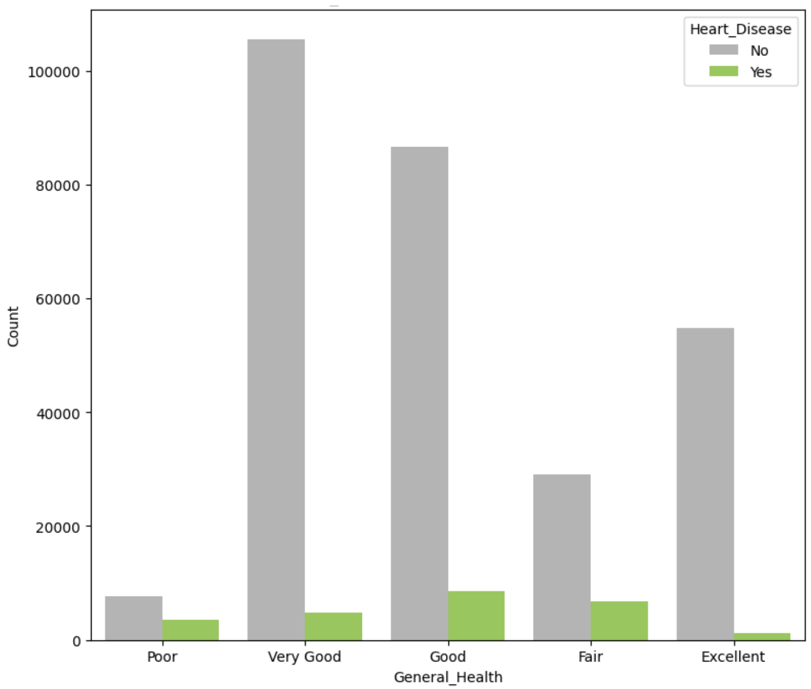
Variables are plotted in relation to heart disease to understand its distribution.

The **elderly group** as well as **poor and fair health conditions** experience higher cardiovascular risk.

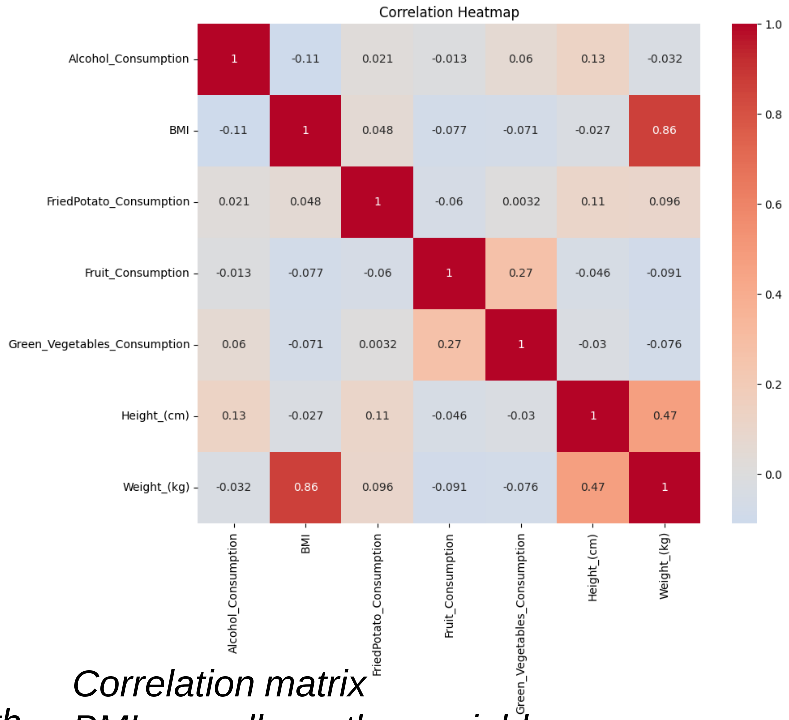
The matrix does **not reveal a strong correlation** between continuous variables.



Heart Disease Distribution by Age



Heart Disease Distribution by General Health



Correlation matrix
BMI as well as other variables are re-checked, and the matrix remains the same

Data Preprocessing

To prepare for predictive analysis, we do **the following steps**:

- **Weight_(kg) and Height_(cm) are removed** since they are highly correlated with BMI
- **Map** Heart_disease to 0 (No) and 1 (Yes)
- Do **one-hot encoding** for binary variables (Sex, Skin Cancer, Diabetes, Arthritis, Depression, and Exercise)
- Map variables General Health, Age, and Checkup
- **Classify BMI** into different categories
- **Define the target variable and features**
- **Split the dataset into training and testing** (80 – 20)
- **Resample** with **SMOTE** for the minority class and random undersample for the majority class

```
[17] # BMI Category
     df['BMI_Category'] = pd.cut(df['BMI'], bins=[0, 18.5, 24.9, 29.9, np.inf], labels=['Underweig


     bmi_mapping = {
         'Underweight': 0,
         'Normal weight': 1,
         'Overweight': 2,
         'Obesity': 3
     }

     df['BMI_Category'] = df['BMI_Category'].map(bmi_mapping).astype(int)

     # Mapping for Diabetes
     diabetes_mapping = {
         'No': 0,
         'No, pre-diabetes or borderline diabetes': 0,
         'Yes, but female told only during pregnancy': 1,
         'Yes': 1
     }
     df['Diabetes'] = df['Diabetes'].map(diabetes_mapping)

     # One-hot encoding for Sex
     df = pd.get_dummies(df, columns=['Sex'])
     df[['Sex_Female', 'Sex_Male']] = df[['Sex_Female', 'Sex_Male']].astype(int)

     # Convert remaining categorical variables with "Yes" and "No" values to binary format for cor
     binary_columns = ['Heart_Disease', 'Skin_Cancer', 'Other_Cancer', 'Depression', 'Arthritis',
```

Building models

Three models are developed to predict whether a person has cardiovascular disease.

- **Decision Tree:**

```
model_tree = DecisionTreeClassifier(random_state=1234, max_depth=8, min_samples_leaf=10)
```

- **Logistic Regression:**

```
model_logreg = LogisticRegression(penalty='elasticnet', solver='saga', l1_ratio = 1, C=0.0001,
random_state=1234, max_iter=20000, class_weight='balanced')
```

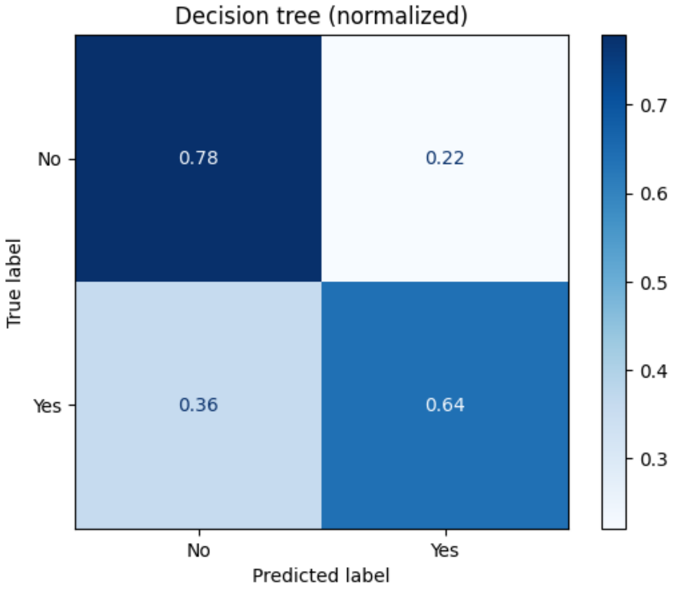
- **Random Forest:**

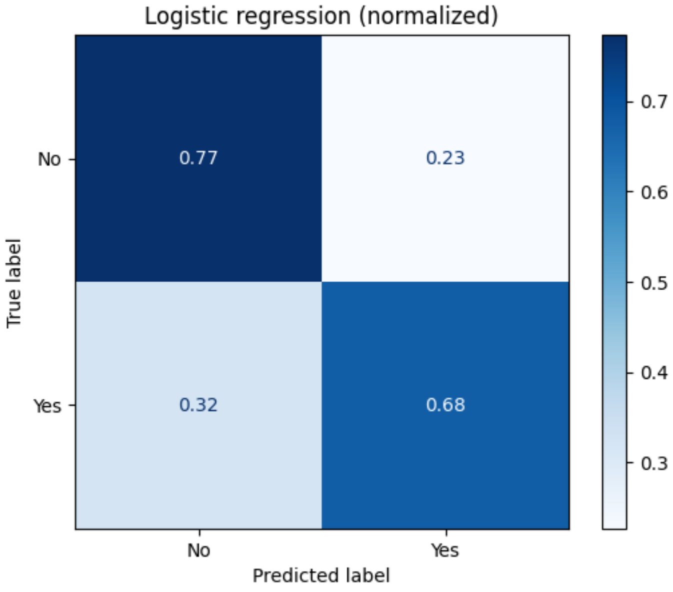
```
model_rf = RandomForestClassifier(n_estimators=200, max_depth=7, min_samples_leaf=20,
class_weight='balanced’)
```

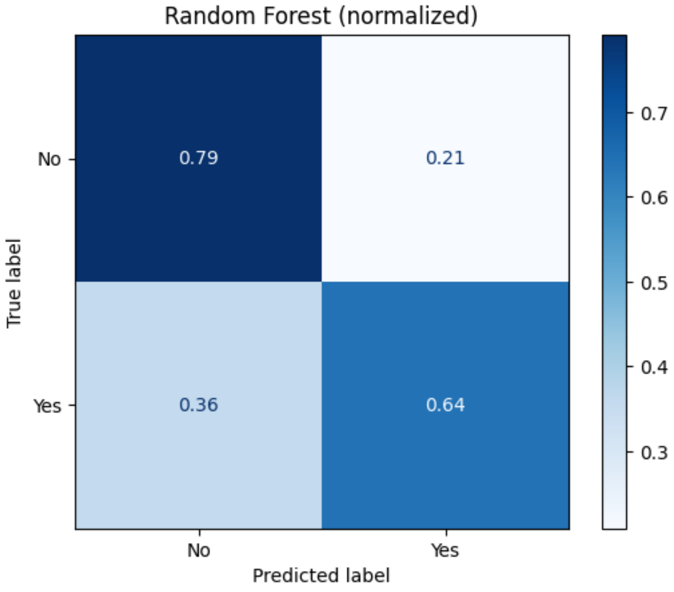
The maximum depth of the trees as well as other hyperparameters are tested so that the models are not underfitting and overfitting.

Results
Model Evaluation – Prediction Accuracy

The **confusion matrices** show that all three models **predict the 0 instances** (no heart disease cases) **better than the 1 instances** (heart disease cases), even with the use of techniques like SMOTE and class_weight='balanced.



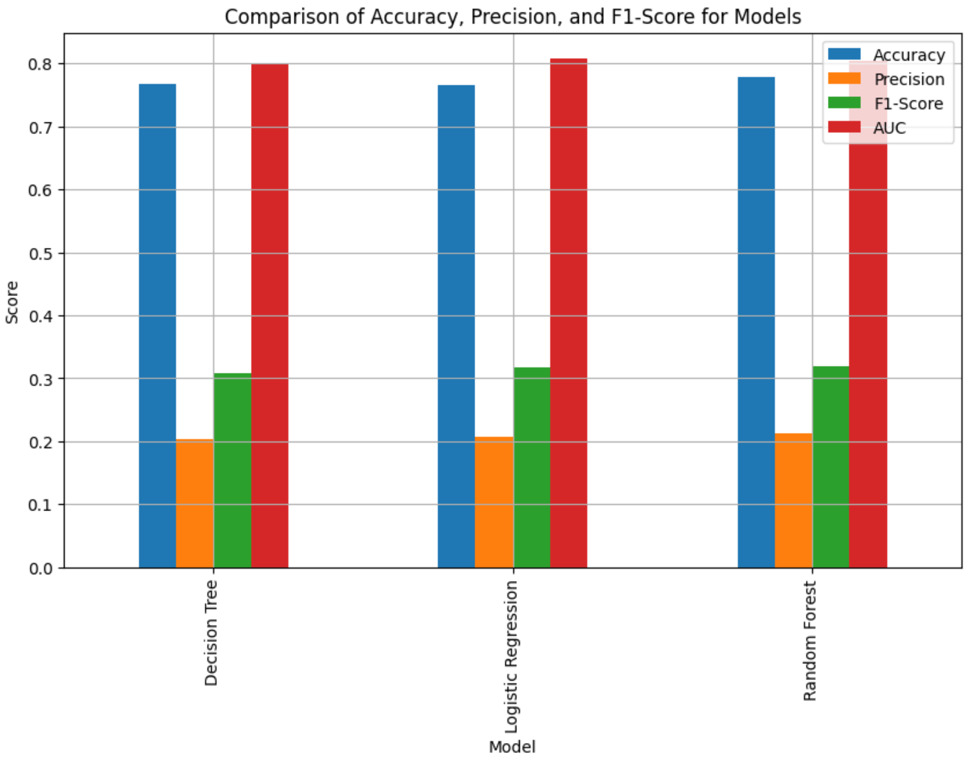




Results
Model Evaluation – Cross-model Comparison

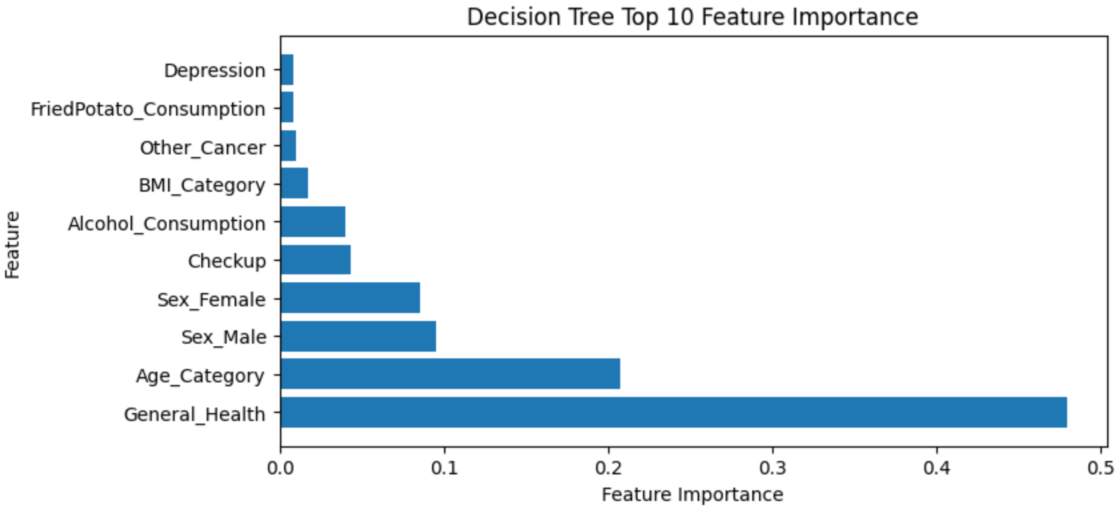
This is due to the **significant class imbalance** in the dataset. Hence, metrics like **precision, recall, F1-score, and the area under the ROC curve (AUC)** are not high in this specific case, and are more informative than truly conclusive of the models' performance.

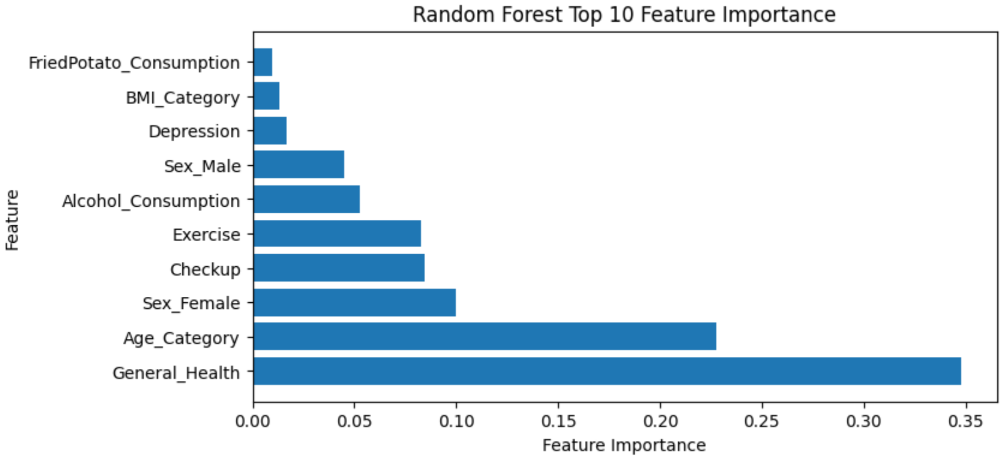
Nevertheless, based on those metrics, the three models perform relatively similarly in terms of precision accuracy. The method random forest seems to slightly dominate.

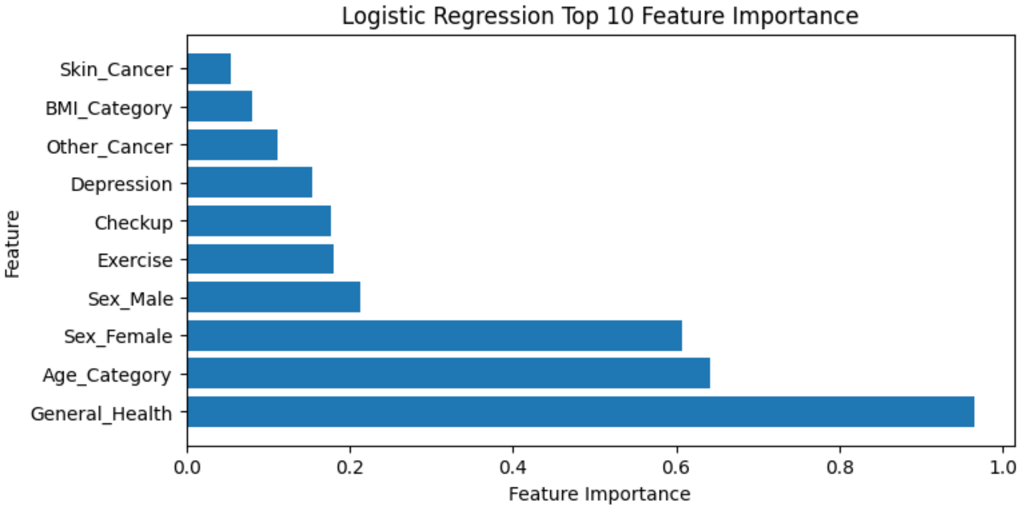


Results
Model Evaluation – Feature Selection

Similarly, there is **little difference regarding feature selection** among the models. The biggest weights are put on general health and age.





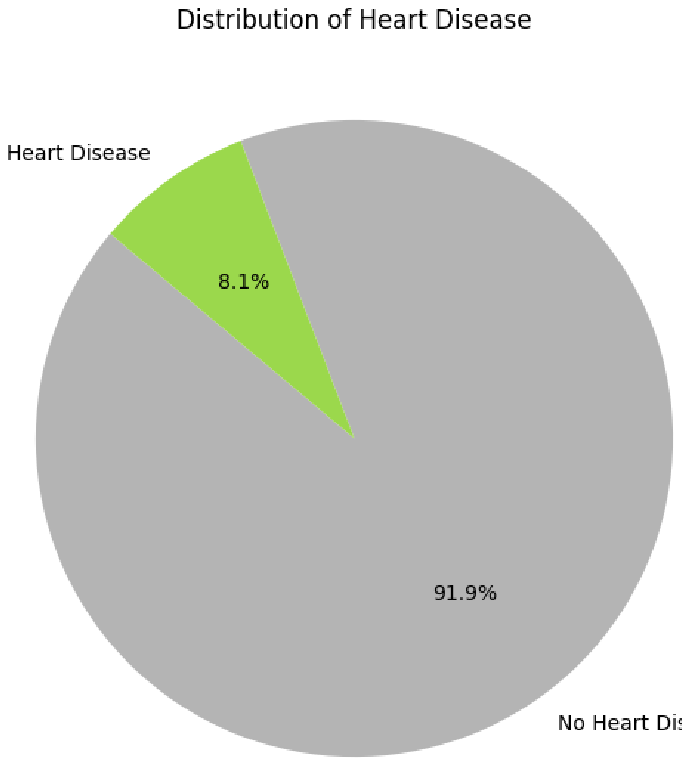


Further Discussion
Limitation and Solutions

To improve the precision of our predictions, it is of the utmost importance **to address the problem of class imbalance** in our data.

Besides what has been applied, other possible solutions could be **tune model threshold**, **ensemble methods**, and **hyper-parameter tuning**.

However, as those applications are **costly and resource-intensive**, their application in practical scenarios requires careful consideration.



Further discussions
Application

With further tuning and a more balanced data set, the predictive models attempted could be employed in the medical field to **offer earlier detection of cardiovascular disease risks**.

These applications promise immense **economic benefits** to both the susceptible individuals (since preventive treatment is less costly) and the limited capacity of health services. The direct benefits to individuals in the society are **healthier hearts, longevity, and increased well-being**.