Project Report

Done By:
Maroun Abou Boutros
Cynthia Obeid
Elie Nakad
Georges Farah

Supervisor:
Tiphaine Viard

Table of Contents

Introduction 4

Distribution of work 4

Methodology 4

Data Visualization 5

Data Cleaning 7

Classification 8

Cross-Validation 9

Decision Tree Interpretation 9

Limitations and Future Improvements 10

Conclusion 11

References 11

Table of Figures

Figure 1: Pie chart representing the proportion of people with/without heart disease 5
Figure 2: Correlation matrix of features 6
Figure 3: Histogram of proportion of people with/without heart disease with respect to their sex.. 7
Figure 4: Histogram of the proportion of females/males 7
Figure 5: Trained decision tree.. 10
Figure 6: Histogram of proportion of people with heart disease in function of age 10

Introduction

Cardiovascular diseases are the second most common cause of death in France at 25% of total deaths. Most of these diseases can be avoided by taking preemptive measures. This project aims to study the relationship between everyday life habits and the general health of an individual on the one hand and the risk of cardiac complications on the other to answer the following question:
Could heart disease be preempted by observing someone's lifestyle?

Distribution of work

To complete this project, tasks were divided among team members in the following manner:

- Cynthia Obeid: found the dataset and did the cleaning
- Elie Nakad: did the data visualization
- Georges Farah: trained the model with grid search
- Maroun Abou Boutros: SMOTE and model evaluation and cross-validation

All team members contributed to writing the project report.

Methodology

To answer this question, a dataset was used from Kaggle containing the following information:

- BMI: Body Mass Index which is a person’s weight in kilograms divided by the square of height in meters. The normal range for a healthy individual should be between 18.5 and 24.9.
- Smoking: “Yes” if the person has smoked at least 100 cigarettes in his entire life and “No” if not.
- AlcoholDrinking: “Yes” if adult men having more than 14 drinks per week and adult women having more than 7 drinks per week, and “No” if not.
- Stroke: “Yes” if the person had a stroke before and “No” if not.
- PhysicalHealth: The number of days during the past 30 days in which a person’s physical health was not good.
- MentalHealth: The number of days during the past 30 days in which a person’s mental health was not good.
- DiffWalking: “Yes” if the person has serious difficulty walking or climbing stairs and “No” if not.
- Sex: Male or Female

- AgeCategory: Fourteen-level age category.
- Race: Person’s ethnicity
- Diabetic: “Yes” if the person has diabetes and “No” if not.
- PhysicalActivity: “Yes” if the person exercises and “No” if not.
- GenHealth: Either “Poor”, “Fair”, “Good”, “Very good”, or “Excellent”.
- SleepTime: Hours of sleep per day.
- Asthma: “Yes” if the person has asthma and “No” if not.
- KidneyDisease: “Yes” if the person has kidney disease and “No” if not.
- SkinCancer: “Yes” if the person has skin cancer and “No” if not.
- HeartDisease: “Yes” if the person has heart disease and “No” if not.

Using this data, different machine learning models will be trained based on different algorithms and varying hyperparameters to try to predict if a person will have heart disease with the best performance.

Data Visualization

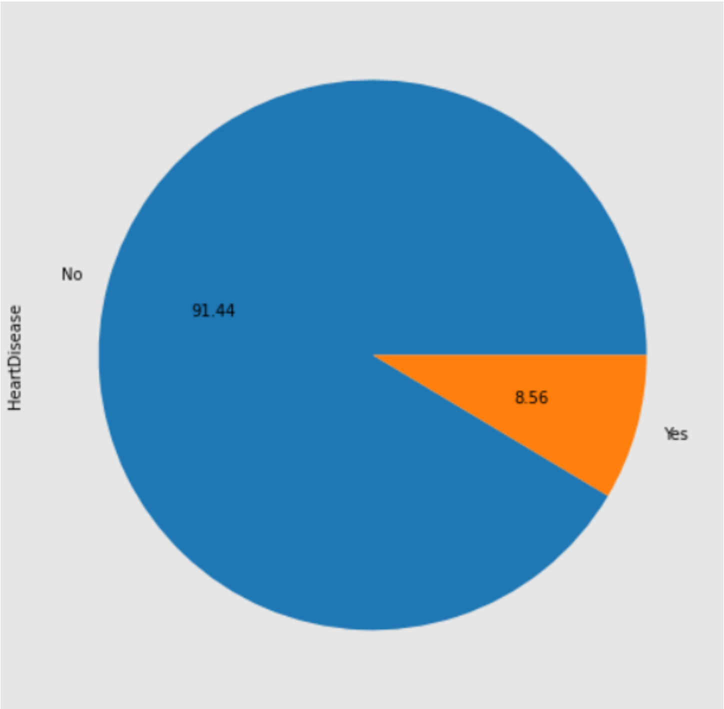


Figure 1: Pie chart representing the proportion of people with/without heart disease

By visualizing the data, a first and important observation is that it is imbalanced. In fact, people with no heart disease make up 91.44% of the dataset. This can have negative effects on the performance of the trained model since when splitting the data into training and testing datasets, we might not have a fair representation of the minority group. To remedy this problem we used SMOTE (synthetic minority oversampling technique) which is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesizes new minority instances between existing minority instances.

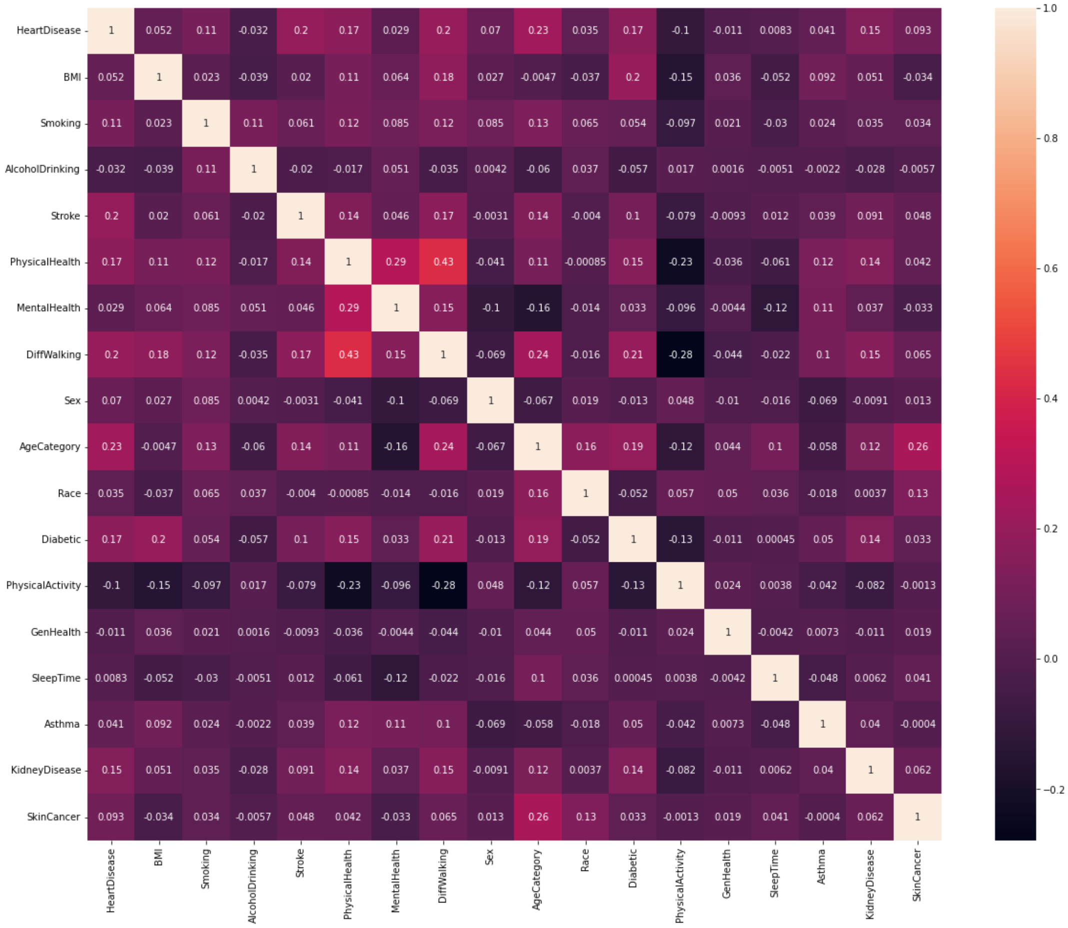


Figure 2: Correlation matrix of features

Additionally, a correlation matrix between the different features of the dataset has shown that there is very little correlation between them. And so doing feature selection or PCA is unnecessary.

And finally for each feature 2 histograms were drawn:

- one visualizing the number of people with heart disease for each value of the feature.

- another showing the percentage of heart disease by feature value.

Here are presented below such histograms for the ‘sex’ feature:

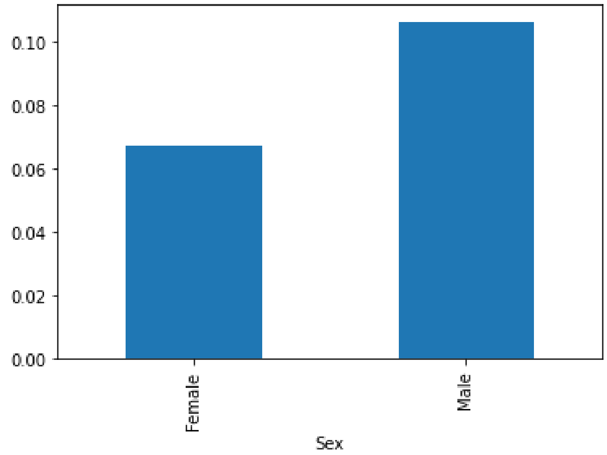


Figure 4: Histogram of the proportion of females/males with heart disease

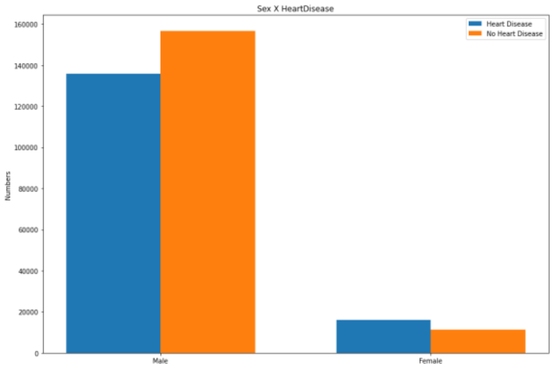


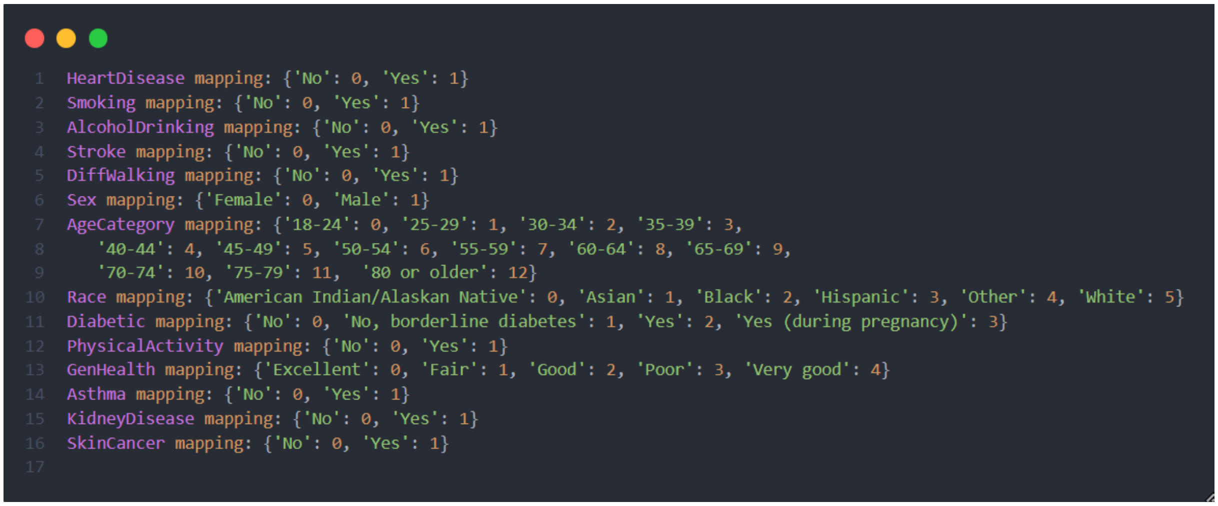
Figure 3: Histogram of proportion of people with/without heart disease with respect to their sex

Data Cleaning

To train a model, the dataset’s features must all be of numeric type and have no null values. The heart disease dataset has no null values. However, it has features of type string. All these features are of categorical type, so this can be easily resolved by transforming the data using label encoding. The types of the different features before cleaning are:

```
HeartDisease         object        AgeCategory          object
BMI                 float64        Race                 object
Smoking              object        Diabetic             object
AlcoholDrinking      object        PhysicalActivity     object
Stroke               object        GenHealth            object
PhysicalHealth      float64        SleepTime           float64
MentalHealth        float64        Asthma               object
DiffWalking          object        KidneyDisease        object
Sex                  object        SkinCancer           object
```

And so we perform label encoding with the following label for each category for each feature:



```
HeartDisease mapping: {'No': 0, 'Yes': 1}
Smoking mapping: {'No': 0, 'Yes': 1}
AlcoholDrinking mapping: {'No': 0, 'Yes': 1}
Stroke mapping: {'No': 0, 'Yes': 1}
DiffWalking mapping: {'No': 0, 'Yes': 1}
Sex mapping: {'Female': 0, 'Male': 1}
AgeCategory mapping: {'18-24': 0, '25-29': 1, '30-34': 2, '35-39': 3,
   '40-44': 4, '45-49': 5, '50-54': 6, '55-59': 7, '60-64': 8, '65-69': 9,
   '70-74': 10, '75-79': 11,  '80 or older': 12}
Race mapping: {'American Indian/Alaskan Native': 0, 'Asian': 1, 'Black': 2, 'Hispanic': 3, 'Other': 4, 'White': 5}
Diabetic mapping: {'No': 0, 'No, borderline diabetes': 1, 'Yes': 2, 'Yes (during pregnancy)': 3}
PhysicalActivity mapping: {'No': 0, 'Yes': 1}
GenHealth mapping: {'Excellent': 0, 'Fair': 1, 'Good': 2, 'Poor': 3, 'Very good': 4}
Asthma mapping: {'No': 0, 'Yes': 1}
KidneyDisease mapping: {'No': 0, 'Yes': 1}
SkinCancer mapping: {'No': 0, 'Yes': 1}

```

Classification

To get the best model, several classification algorithms were tried and compared. The algorithms used are:

- Logistic regression
- Support vector machine
- Decision tree
- Random forest

For each of these algorithms the training was done using a grid search approach to find the best parameters and get the best results possible for each model. The results with the best parameter combination for each algorithm for the testing dataset are showcased in the table below:

algorithm	accuracy	precision	recall	f1_score	parameters
Logistic Regression	0.740	0.728	0.769	0.749	penalty: none solver: lbfgs
Support Vector Machine	0.740	0.724	0.777	0.750	Loss: squared_hinge penalty: l2
Decision Tree	0.860	0.864	0.855	0.860	Criterion: entropy Max_features: log2
Random Forest	0.881	0.902	0.855	0.878	Criterion: entropy Max_features: sqrt N_estimators: 10

Accuracy represents the ratio of correct diagnosis.

Precision represents the percentage of people who have heart disease among those who got diagnosed.
Recall represents the percentage of people who were diagnosed with heart disease among those who actually got heart disease.
f1_score is calculated using the following formula: $f1\_score = \frac{precision * recall}{precision + recall}$. f1_score sums the performance of the model by regrouping precision and recall.

It is best to increase the recall because this model is to predict the possibility of getting a cardiovascular disease for an individual. So it is better for recall to be high and so presume more people would have heart disease and so they would take preventive measures and therefore decrease the number of deaths.

According to the above results, Random Forest and Decision Tree seem to be the best algorithms with Random Forest surpassing Decision Tree by a small margin.

Cross-Validation

The scores obtained above have shown that random forest is an excellent model to predict heart disease. To verify the effectiveness of the random forest model, cross-validation is done. With 5 cross-validations the following results were obtained:
[0.7516703, 0.90639974, 0.90839413, 0.91088936, 0.9068187]
The cross-validation scores show that the random forest model is effective having high f1_scores.

Decision Tree Interpretation

The Decision Tree algorithm gave very good results, slightly lower than random forest. And since it is easily interpretable the top levels of the tree were plotted below in order to get a better understanding of the main causes of heart diseases by looking at the most important features.
By looking and the top of the decision tree we can observe that the most important feature is PhysicalActivity followed by PhysicalHealth and AlcoholDrinking. So doing sports and not consuming alcohol helps lower the risk of getting heart disease.

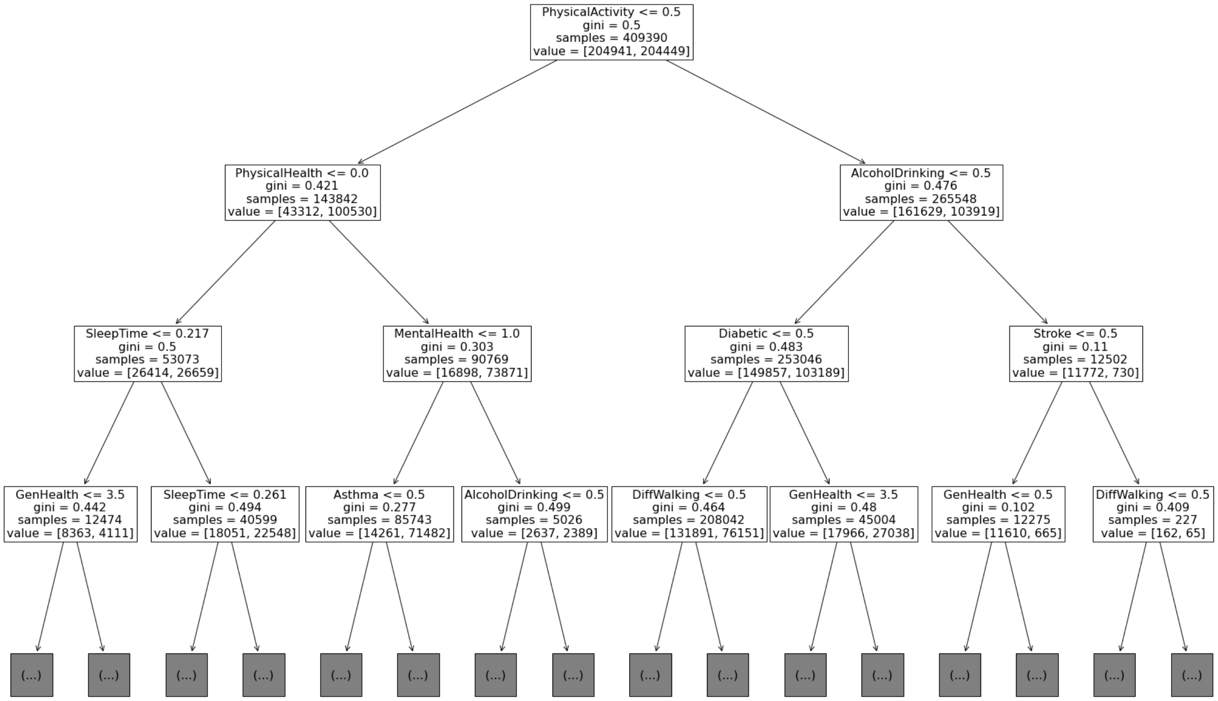


Figure 5: Trained decision tree

Limitations and Future Improvements

A major limitation was that the data was imbalanced. However, this reflects reality as only 7.2% of people are diagnosed with heart disease during their lifetime. But by visualizing the data, the percentage of heart disease diagnosis for people over 70 years old can be over 15% and so a more accurate model can be trained to predict heart disease for people aged more than 70 if more data for this age category was available.

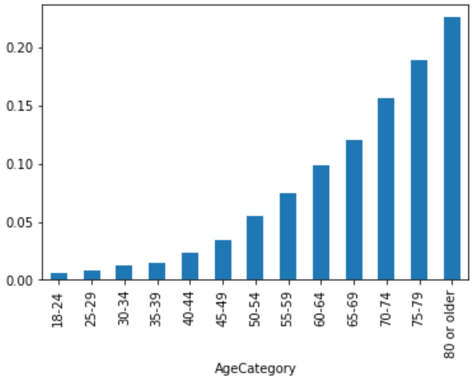


Figure 6: Histogram of proportion of people with heart disease in function of age

Under-sampling could have been used instead of SMOTE to see if better results could be had. This method can discard useful data that is necessary for building rule-based classifiers such as Random Forest and Decision Trees. Given the big size of the data, this problem could be avoided.

Another improvement could be using more classification algorithms to evaluate their performance on this dataset such as Gradient Boosting Classifier.

Conclusion

According to the project, heart disease can be predicted with high accuracy using different classification algorithms. In fact, all the algorithms give good results, however, the decision tree and random forest perform better with random forest better than the decision tree by a small margin. And so we conclude that it is indeed possible to foresee heart disease by looking at personal life indicators.

References

[1] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[2] https://analyticsindiamag.com/how-to-improve-the-accuracy-of-a-classification-model/

[3]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[6]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html