Cardiovascular Disease Prediction Model

A Classification Problem on an Imbalanced Dataset

Omi Singh

5/29/24

Contents

Summary

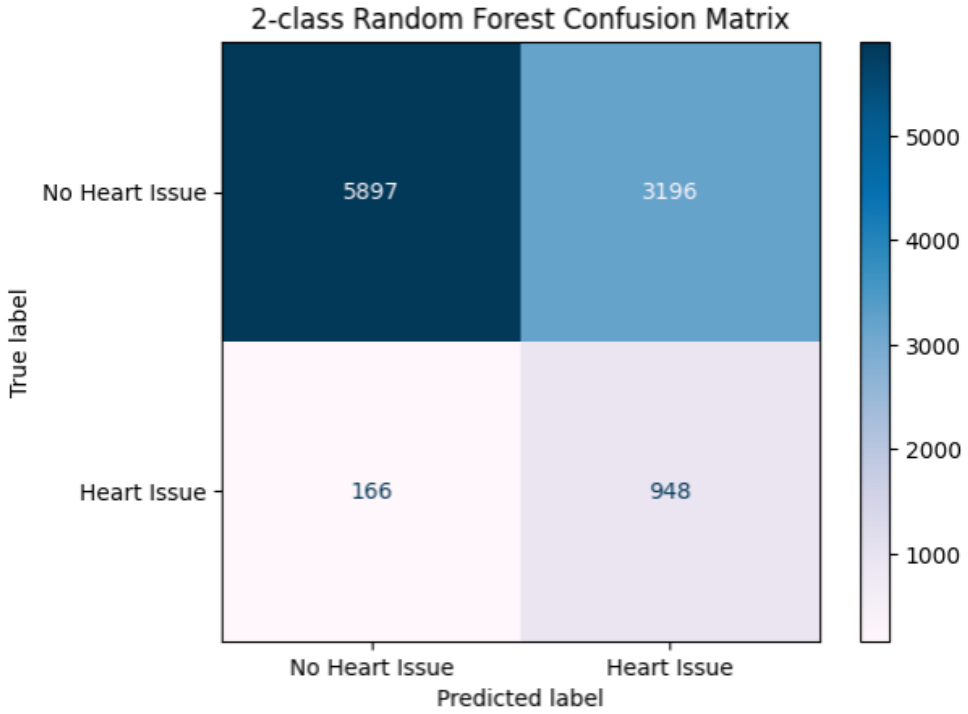
The project involved combining 5 common heart issues such as heart attack, stroke, or angina into one main target Cardiovascular event.

Several biographical and medical factors are taken in as inputs to be able to evaluate the risk of a potential cardiovascular event for the participant

The study is for members over 20 years old and is driven by 60 data files taken from the CDC website spanning over 2005 - 2016 period.

Based on approximately 35,000 participants the biggest driving factor in assessing the risk of a cardiovascular event is age, followed by a diabetes diagnosis, waist and household size, marital status, and the decision to quit smoking or not.

Out of the 1114 positive test cases, our prediction model correctly identified 948 as potential risk cases. Correctly catching risk cases 85% of the time. Leading to an inexpensive and solid initial screening method.



(Model performance during testing)

We have a strong classification model for initial screening. The questionnaire for screening will consist of 21 input factors which are revealed in further detail during the Process section of the report. Based on this questionnaire an initial risk assessment can be performed at home or a clinic.

Background

We live in an age where data is readily available, from smartphones to digital watches. The amount of information available at our fingertips is increasing, this project aims to make that information relevant by bringing in risk assessment features.

Taking in factors such as age, blood pressure, diabetes diagnosis, smoking habits, and others we can quickly assess if our lifestyle falls within a safe or high-risk group.

Being able to quickly analyze our heart health can lead to a more informed patient, ready to be proactive with medical needs or making changes to our habits or diet.

With 85% of risk cases being identified as a concern correctly, we can stay on top of our heart health within minutes at home or a medical clinic.

The Process

We have followed a standard Data Science Project Methodology. After identifying the goal of the project as a classification predictive model.

We have broken down the actual data steps into the following four components.

Problem Statement: Based on demographic and rudimentary medical information, create a cardiovascular risk assessment tool that takes a list of information and classifies participants as high-risk or not.

Data Steps:

- Data Cleaning
- Exploratory Data Analysis (EDA)
- Preprocessing Pipeline
- Model Evaluations and Selection

Overall, the sequence of steps was in the above order. However, as needed the data cleaning step was revisited during the EDA and preprocessing steps.

Data Collection and Cleaning

Our data source was a total of 60 files from the CDC website. As such, the first step was collecting and combining the data into one data frame or table. All initial column names needed an update.

Key data cleaning steps:

- As our target label is the presence of a cardiovascular event, participants 20 years or younger were removed due to the high number of missing target values in this age group. The 5 common heart issues were combined into one target column where the presence of any of the 5 issues was noted as a cardiovascular event.

- The number of days since a person quit smoking was present in two separate columns indicating days, weeks, or months followed by the number of units. This was combined into one final column with all values in the number of days.

- A key decision made at this point was regarding how to represent not applicable values. As an example, when it came to diabetes diagnosis age, participants who were never diagnosed with diabetes were represented using '-99' as the Not Applicable value.

At the end of the data-cleaning step, we had 34,022 entries with 29 columns (including our target label). We also note that only 11% of participants have experienced a cardiovascular event, making our data and problem an imbalanced one.

	0	1	2	3	4
SEQN	3.112700e+04	31128.00	31129.00	31130.0	31131.0
RIAGENDR	1.000000e+00	2.00	1.00	2.0	2.0
RIDAGEYR	5.397605e-79	11.00	15.00	85.0	44.0
RIDRETH1	3.000000e+00	4.00	4.00	3.0	4.0
INDFMIN2	4.000000e+00	5.00	10.00	4.0	11.0
DMDHHSIZ	4.000000e+00	7.00	6.00	1.0	4.0
DMDMARTL	NaN	NaN	5.00	2.0	1.0
BPXPULS	1.000000e+00	1.00	1.00	1.0	1.0
BPXSY1	NaN	100.00	104.00	NaN	144.0
BPXDI1	NaN	62.00	76.00	NaN	74.0
BMXBMI	NaN	17.45	26.53	NaN	30.9
BMXWAIST	NaN	62.80	97.80	NaN	96.0
LBDHDD	NaN	55.00	46.00	NaN	39.0
LBXTR	NaN	NaN	NaN	NaN	86.0
LBDLDL	NaN	NaN	NaN	NaN	49.0
LBXTC	NaN	129.00	170.00	NaN	105.0
DIQ010	NaN	2.00	2.00	2.0	2.0
DID040	NaN	NaN	NaN	NaN	NaN
KIQ022	NaN	NaN	NaN	2.0	2.0
MCQ160B	NaN	NaN	NaN	2.0	1.0
MCQ160C	NaN	NaN	NaN	2.0	2.0
MCQ160D	NaN	NaN	NaN	2.0	2.0
MCQ160E	NaN	NaN	NaN	2.0	2.0
MCQ160F	NaN	NaN	NaN	2.0	2.0
MCQ300A	NaN	NaN	NaN	2.0	1.0
SMQ020	NaN	NaN	NaN	2.0	2.0
SMQ040	NaN	NaN	NaN	NaN	NaN
SMQ050Q	NaN	NaN	NaN	NaN	NaN
SMQ050U	NaN	NaN	NaN	NaN	NaN

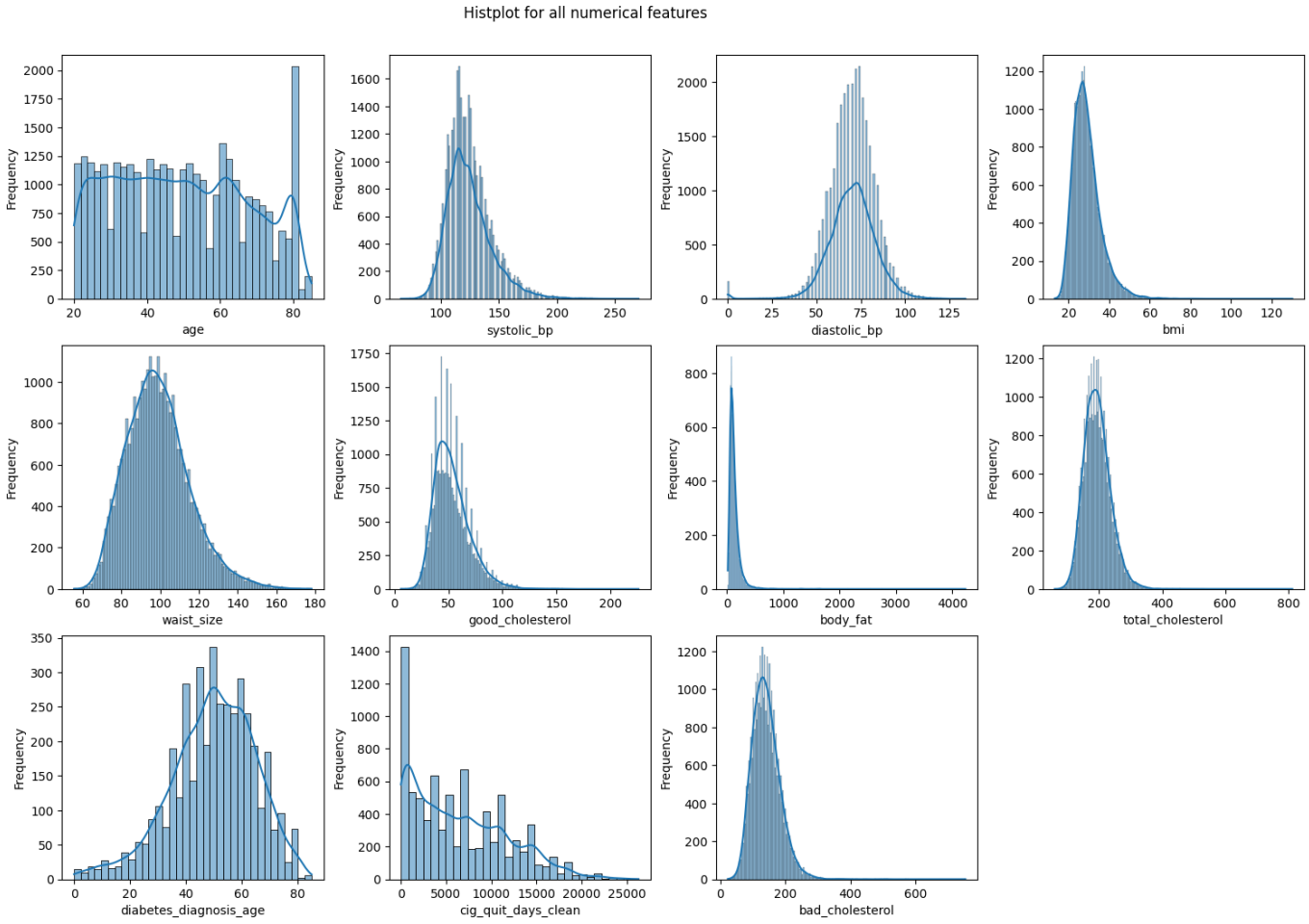
id	31130.0	31131.0	31132.0	31134.0	31136.0
gender	2.0	2.0	1.00	1.00	2.0
age	85.0	44.0	70.00	73.00	41.0
ethnicity	3.0	4.0	3.00	3.00	4.0
income	4.0	11.0	11.00	12.00	7.0
household_size	1.0	4.0	2.00	2.00	1.0
marital_status	2.0	1.0	1.00	1.00	5.0
bp_regularity	1.0	1.0	1.00	1.00	NaN
systolic_bp	NaN	144.0	138.00	130.00	NaN
diastolic_bp	NaN	74.0	60.00	68.00	NaN
bmi	NaN	30.9	24.74	30.63	NaN
waist_size	NaN	96.0	96.50	117.10	NaN
good_cholesterol	NaN	39.0	59.00	49.00	NaN
body_fat	NaN	86.0	65.00	195.00	NaN
total_cholesterol	NaN	105.0	147.00	186.00	NaN
diabetes_diagnosis	2.0	2.0	1.00	2.00	2.0
diabetes_diagnosis_age	-99.0	-99.0	63.00	-99.00	-99.0
kidney_fail	2.0	2.0	2.00	2.00	2.0
fam_heart_issues	2.0	1.0	2.00	2.00	2.0
cig_smoker	2.0	2.0	2.00	2.00	2.0
cig_quit	-99.0	-99.0	-99.00	-99.00	-99.0
heart_issue	2.0	1.0	2.00	2.00	2.0
cig_quit_days_clean	-99.0	-99.0	-99.00	-99.00	-99.0
bad_cholesterol	NaN	66.0	88.00	137.00	NaN
missing_body_fat_indicator	1.0	0.0	0.00	0.00	1.0

Exploratory Data Analysis

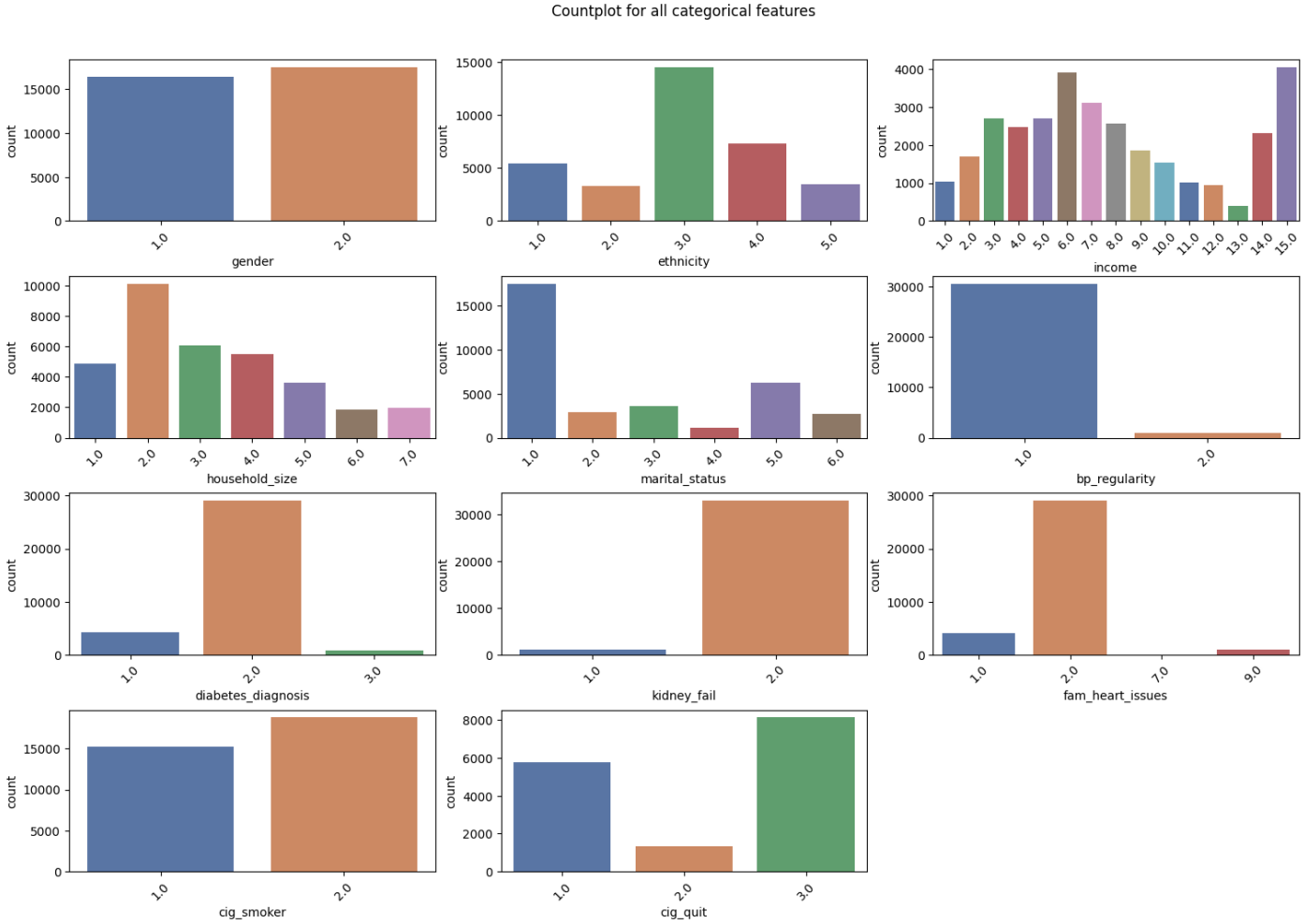
The goal of Exploratory Data Analysis (EDA) is to gain valuable intuition of the data. This step is crucial for building an understanding of the dataset and getting a feel of the big picture, as it is harder to appreciate these relations once we dive into modeling algorithms or cross-validations.

This step was carried out in 4 parts: Statistical Summary, Numerical Features, Categorical Features, and Correlations.

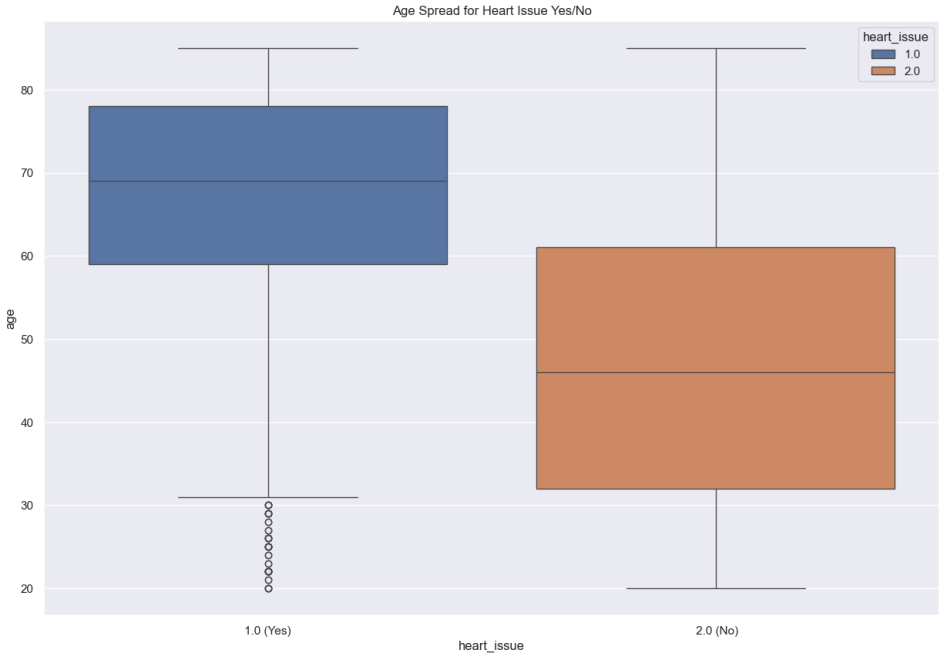
We noted the average age of participants at 50. However, the most common age of participants was around 80. We also noted that the number of days since a person quit smoking is right skewed with fewer and fewer members quitting smoking for a larger number of days.



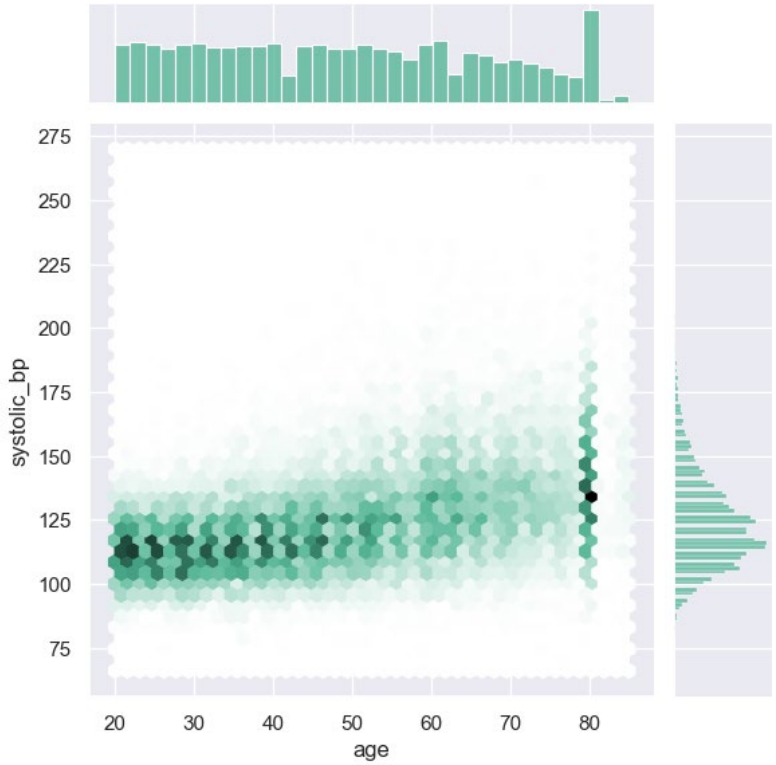
During the exploration of categorical features, we noted that we have a good mix of both genders. The most common ethnicity is 'Non-Hispanic White' while the most common marital status is 'Married'.



The correlation between age and heart issues jumped out in the pair plots. We also noticed a high correlation between some of the features, such as total cholesterol and bad cholesterol, or waist size and BMI. The median age of participants with heart issues lies in the late 60s, while the median age of participants without a heart issue lies in the 40s.



It was also interesting to note some other features that showed a correlation with age. As an example, the systolic blood pressure range tends to increase with age.



At the end of Exploratory Data Analysis, we dropped the body fat column due to a high number of missing values and the column showing a high correlation with total cholesterol. Post EDA we have 34,022 entries and 22 columns (including our target label).

Preprocessing Pipeline

The preprocessing step started with the splitting of available data into training and testing data. 70% of the data was used to train models while 30% of data was to test them.

After the split, the following preprocessing steps were completed.

- Imputation

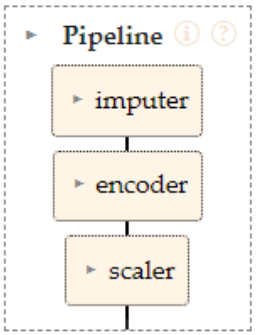
The missing values were imputed using the median for numerical features and mode for categorical features.
Given the even and high volume spread for both genders, missing median and mode values were computed after grouping entries by gender.
This was to account for certain differences in certain features like income and waist size between the two gender groups.

- Encoding

Encoding is used to convert categorical values into numerical features to prepare data for different models.
One Hot Encoding was used for ethnicity and marital status, as the categorical values are non-sequential.
On the other hand, for income brackets, Ordinal Encoding was used to retain the sequential nature of categorical values.

- Scaling

The data values were scaled using a Standard Scaler which sets the mean for each feature at 0 and scales them to a unit variance.



Modeling – Evaluation and Selection

The project is centered around identifying risk cases, as such missing a potential true positive will be a big drawback.
The need dictates a high recall score as such it has been used as the primary criterion.
We also created a custom evaluation metric to quickly note Recall and F1 score, while plotting Precision-Recall Curve, a confusion matrix, and a classification report.

Following classification modeling techniques were used.

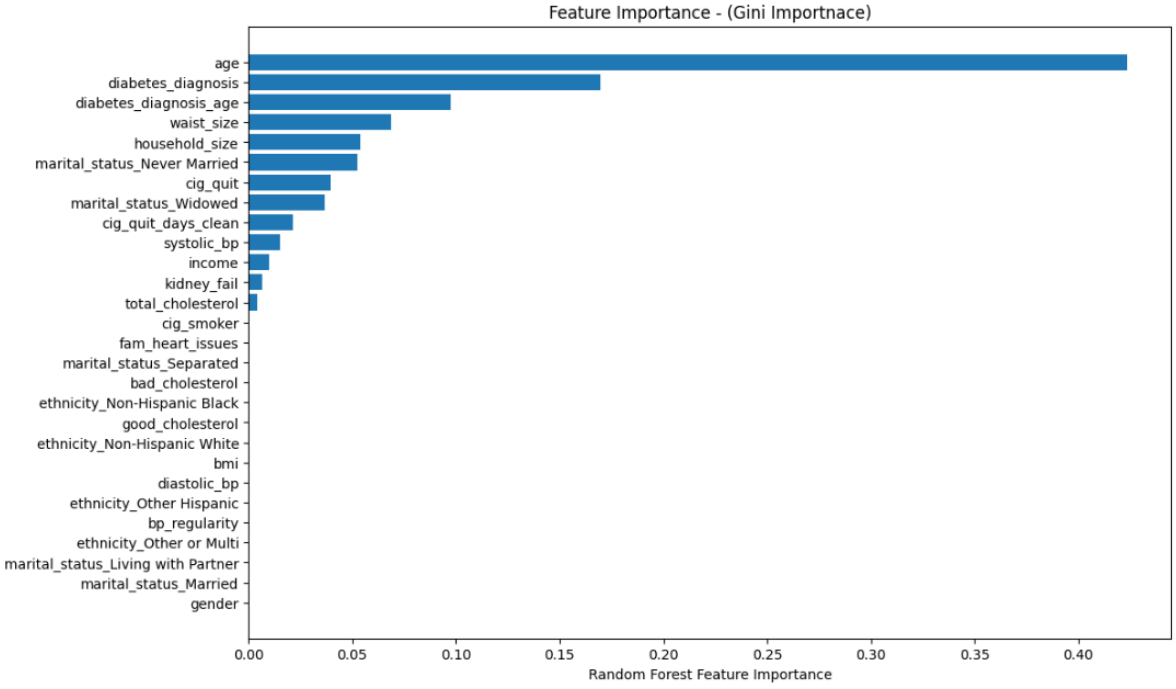
Classification Model	Recall Score
Logistic Regression	0.80
LightGBM Classifier	0.81
Support Vector Machine	0.83
Random Forest	0.85

We used the Grid Search technique to do cross-validation and hyperparameter tuning.
The following parameters led to the best Random Forest model.

```
print(best_rf_model.best_params_)
```

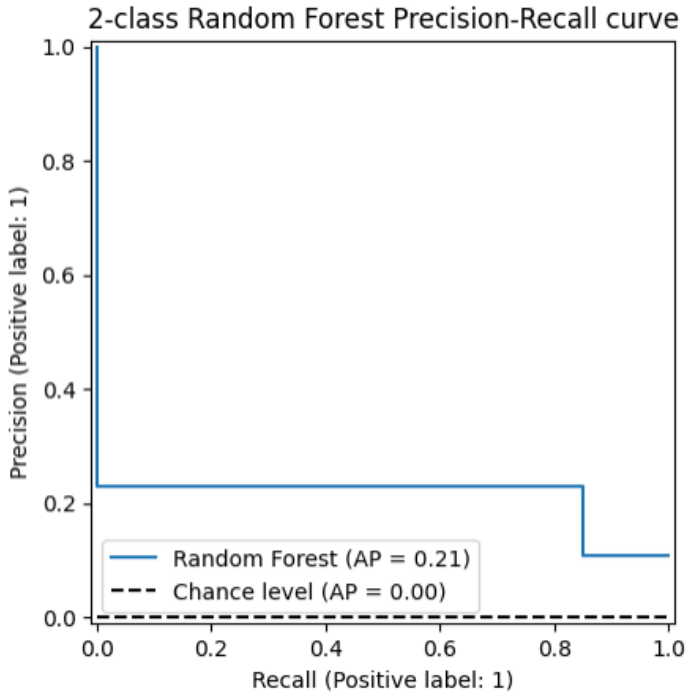
```
{'max_depth': 3, 'max_features': 'sqrt', 'max_leaf_nodes': 3, 'n_estimators': 25}
```

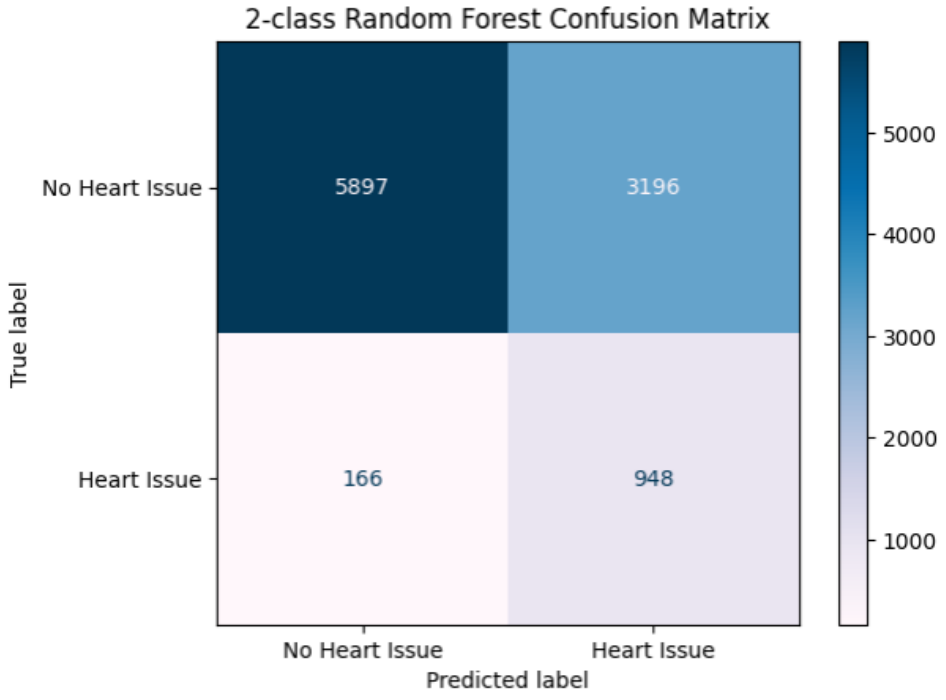
Plus, age was again noted as the most important feature of this model.
Due to one hot encoding certain marital status jumped ahead in importance scale, indicating overall the relevance of marital status.



Noting the test performance of our best Random Forest model below.

```
Random Forest Recall score: 0.85
Random Forest F1_score: 0.36
```



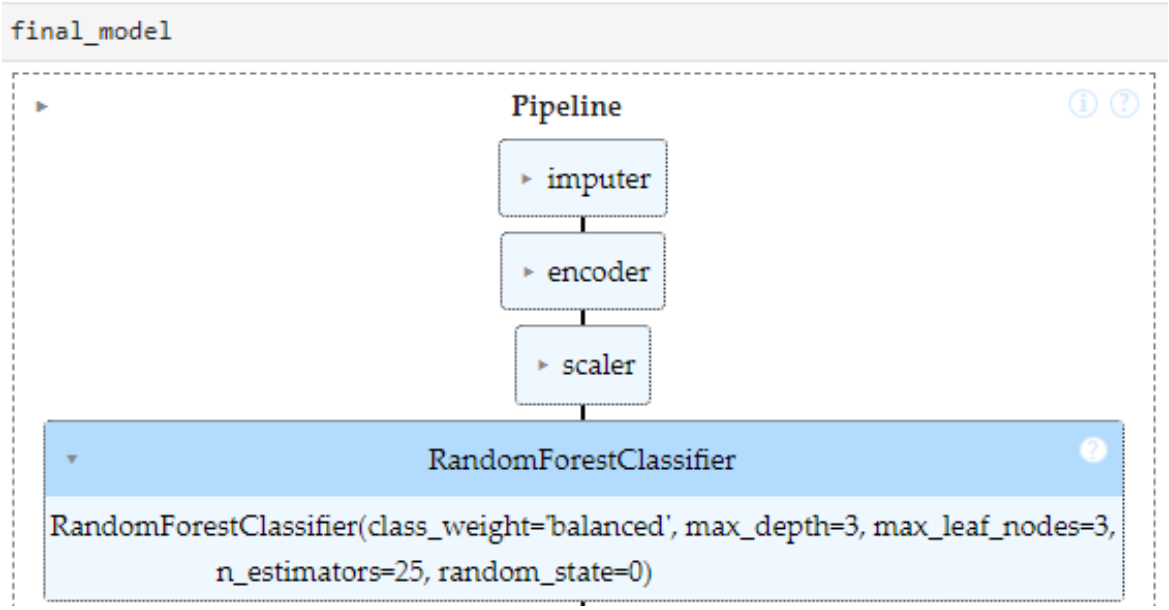


	precision	recall	f1-score	support
0.0	0.97	0.65	0.78	9093
1.0	0.23	0.85	0.36	1114
accuracy			0.67	10207
macro avg	0.60	0.75	0.57	10207
weighted avg	0.89	0.67	0.73	10207

Conclusion

After selecting the best model, we combine our preprocessing pipeline and classifier into one production-ready pipeline.

The new data can be entered for preprocessing and prediction straight away.



We ended with a model showing a recall score of 0.85 and an f1 score of 0.36.

This is a solid performance for an imbalanced data 1:9, this imbalance was handled using the class_weight parameter of sklearn classifiers.

There is definitely a future scope to this project, adding some additional features can lead to new insights while removing low-importance features, and limiting the number of inputs can make the model lighter and more user-friendly.