Predicting 10-year Coronary Heart Disease risk from the Framingham Heart study dataset

Johan van Nispen

January 29, 2024

Abstract

Coronary Heart Disease (CHD) is the major cause of death worldwide. An early diagnosis is crucial for treatment of the disease and the reduction of further complications. This study investigates the possibility to predict the 10-year CHD risk by using different machine learning models on the publicly available Framingham heart study dataset. After missing value imputation, outlier removal and class re-balancing using SMOTE, the machine learning models are trained and compared. A stacking ensemble using the basic learners Random Forest, Support Vector Machine and k-Nearest Neighbors achieved the highest accuracy (0.925). The lowest False Positive Rate was found for the Random Forest model (0.066) and the lowest False Negative Rate was found for the voting ensemble model (0.056).

1 Introduction

Coronary Heart Disease, also called Coronary Artery Disease, is a disease affecting the blood vessels of the heart. The disease is often caused by the build up of fatty deposits on the inside of the blood vessels, leading to a disruption in the supply of oxygen rich blood to the heart muscle. In 2015, CHD affected 110 million people, and resulted in 8.9 million deaths globally, which is 15.6% of all deaths, making it the most common cause of death globally[1]. If diagnosed early the disease can be treated, reducing the risk of further problems[2]. According to a World Health Organization (WHO) survey, only 67% of heart diseases can currently be predicted [KP16].

To help doctors diagnose the risk of developing CHD, recent advancements in the field of big data and machine learning (ML) can possibly be used to more accurately predict the risk of 10-year CHD by training ML models on medical patient data.

In this project, which is part of the course "Machine Learning" of the Open University in the Netherlands, it will be attempted to predict the 10-year CHD risk from the publicly available Framingham heart study dataset by using various classification ML algorithms. To guide the project, the following research questions have been formulated:

A. **Can 10-year CHD risk be predicted by using classification ML models?**

B. **How do different ML models compare in predicting 10-year CHD risk?**

C. **Can 10-year CHD risk factors be identified from ML model interpretation?**

This report is structured as follows: in section 2 the dataset and the choices made in data preprocessing will be presented. Section 3 will describe the choices made with regard to the different machine learning models. In section 4 the results will be presented, and in section 5 the conclusions will be drawn.

2 Data Analysis

The code used for data analyses and the machine learning steps described in this section and in rest of the report can be found in the accompanying Jupyter Notebook file which has been provided as part of the project deliverables.

[1] https://en.wikipedia.org/wiki/Coronary_artery_disease
[2] https://www.nhs.uk/conditions/coronary-heart-disease/treatment/

2.1 Dataset

The dataset used in this project is the Framingham Heart Study (FHS) dataset, which can be downloaded from Kaggle. The FHS dataset originates from a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham, Massachusetts, USA[3]. The dataset consists of 4240 records, each with 15 features related to potential risk factors in developing CHD, and 1 target feature indicating the 10-year CHD risk. A summary of the unprocessed dataset before the start of preprocessing is given in Table 1.

Feature	Description	Datatype	Missing values
Male	Patient gender (male/female)	Nominal	No
Age	Patient age	Discrete	No
Education	Patient education (1-4)	Discrete	Yes (105)
CurrentSmoker	Does the patient smoke?	Nominal	No
CigsPerDay	Number of cigarettes smoked per day	Discrete	Yes (29)
BPMeds	Is blood pressure medication is used?	Nominal	Yes (53)
PrevalentStroke	Did the patient have a stroke in the past?	Nominal	No
PrevalentHyp	Is the patient hypertensive?	Nominal	No
Diabetes	Does the patient have diabetes?	Nominal	No
TotChol	Total cholesterol level	Continuous	yes (50)
SysBP	Systolic blood pressure	Continuous	No
DiaBP	Diastolic blood pressure	Continuous	No
BMI	Body mass index	Continuous	Yes (19)
HeartRate	Heart rate	Continuous	Yes (1)
Glucose	Glucose level	Continuous	Yes (388)
TenYearCHD	**10-year CHD risk (target)**	**Nominal**	**No**

Table 1: Description of the Framingham heart study dataset (4240 records, 645 missing values)

As can seen from Table 1, from a total of 4240 records (rows) there are 645 missing values, so in worst case, assuming all missing values are on a different row, around 15% of the data could potentially be incomplete. The nominal and discrete data-types had already been converted to integer types in the unprocessed dataset (before downloading), and all continuous data-types had already been converted into floating point numbers. It was also checked the dataset does not contain any duplicates.

2.2 Preprocessing

In the first data preprocessing step, missing data values are either imputed or dropped, and in a second step outliers in the data are identified and removed.

The missing values of the features cigsPerDay, totChol, BMI, heartRate, glucose and education were replaced by the median value of the column. The median was chosen here to compensate for a slight skewness in some of the underlying feature distributions. In Figure 1 the distribution of the continuous and discrete features have been plotted. It was chosen to drop the missing values for the feature BPMeds. After this first preprocessing step the dataset still contained 4187 rows.

Figure 2 is a boxplot showing the outliers in the dataset. The figure shows that the majority of the outliers can be found above the upper whisker. Especially for totChol, sysBP and glucose there is a long tail with higher values. To remove outliers from the dataset it was chosen to apply the three-sigma rule. This rule will remove any value above or below three standard deviations from the mean of the distribution, which preserves 99.7% of the original data for a Gaussian distribution. After outlier removal the dataset still contained 3988 rows, so in total about 5.9% of the data rows were removed during the data preprocessing stage.

Figure 3 shows a pairwise scatterplot of the continuous features. From this figure, any direct correlation between two features can be noticed from the shape of the cloud of dots. Reading from the diagram, only the strong correlation between sysBP and diaBP catches the eye. The orange dots in the diagram represent the male population, the blue dots the female population. Looking at the feature distributions on the matrix diagonal, it does not appear that the distributions between male and female seem to differ significantly in either shape or mean value, with the exception of the mean BMI, which appears to be slightly higher for the male population. In total, the dataset has 43.7% males and 56.3% females, which is only slightly imbalanced.

[3] https://en.wikipedia.org/wiki/Framingham_Heart_Study

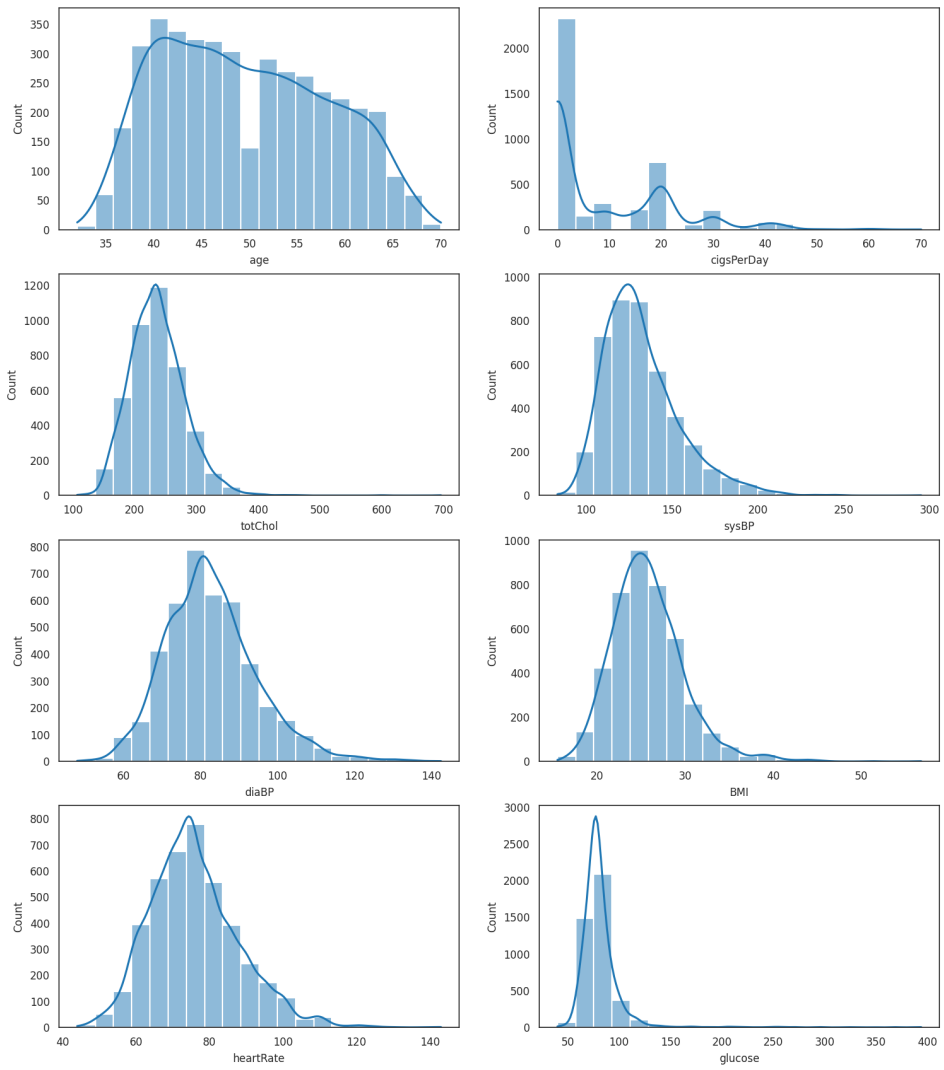


Figure 1: Distribution plots of continuous and discrete features from the FHS dataset.

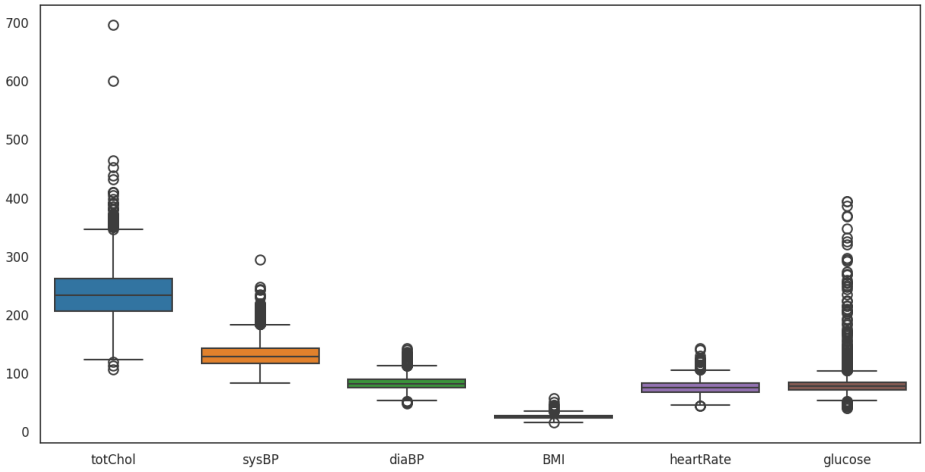


Figure 2: Boxplot showing data outliers in the FHS dataset.

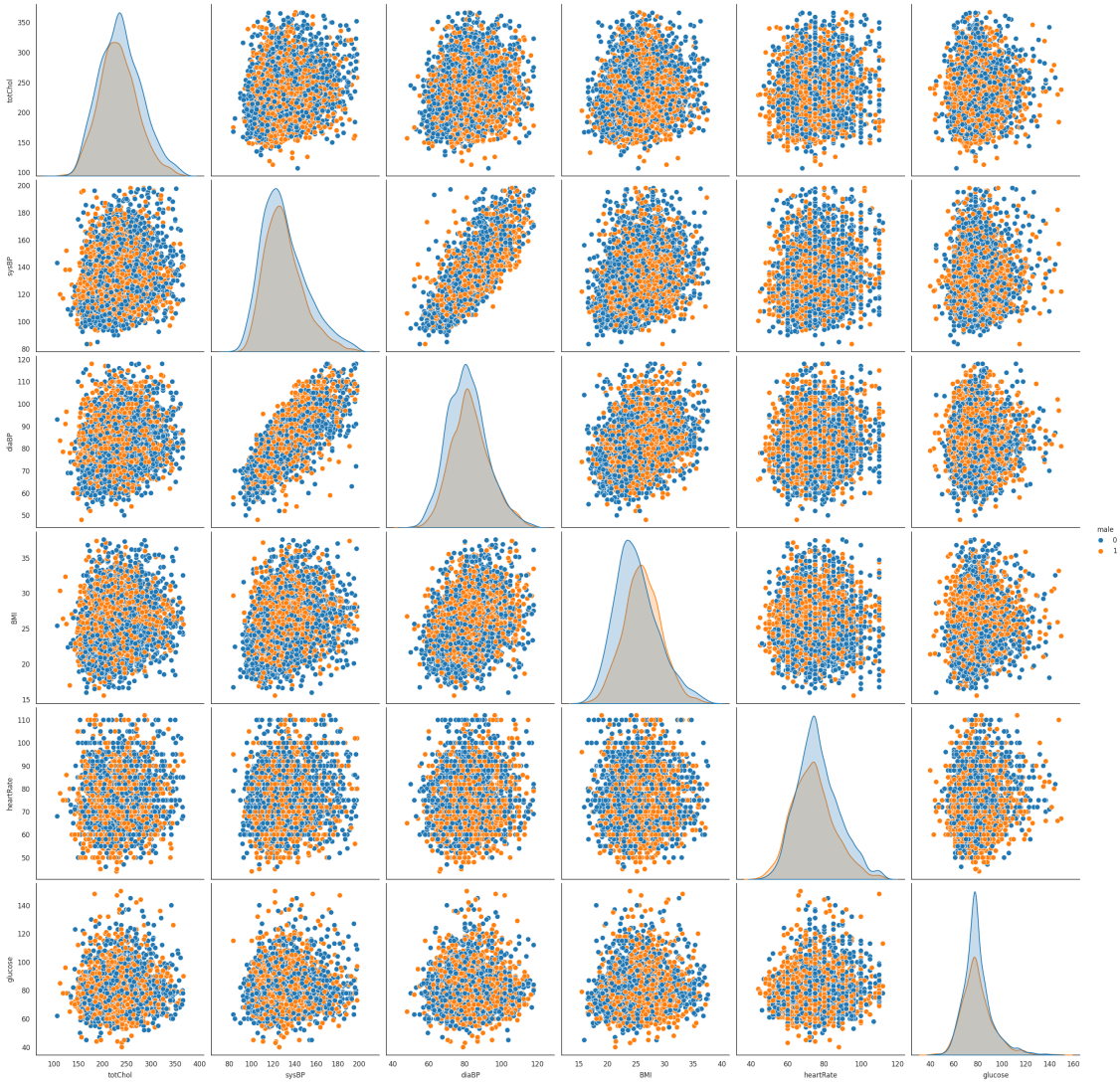


Figure 3: Pairwise scatterplot showing bivariate distributions in the FHS dataset.

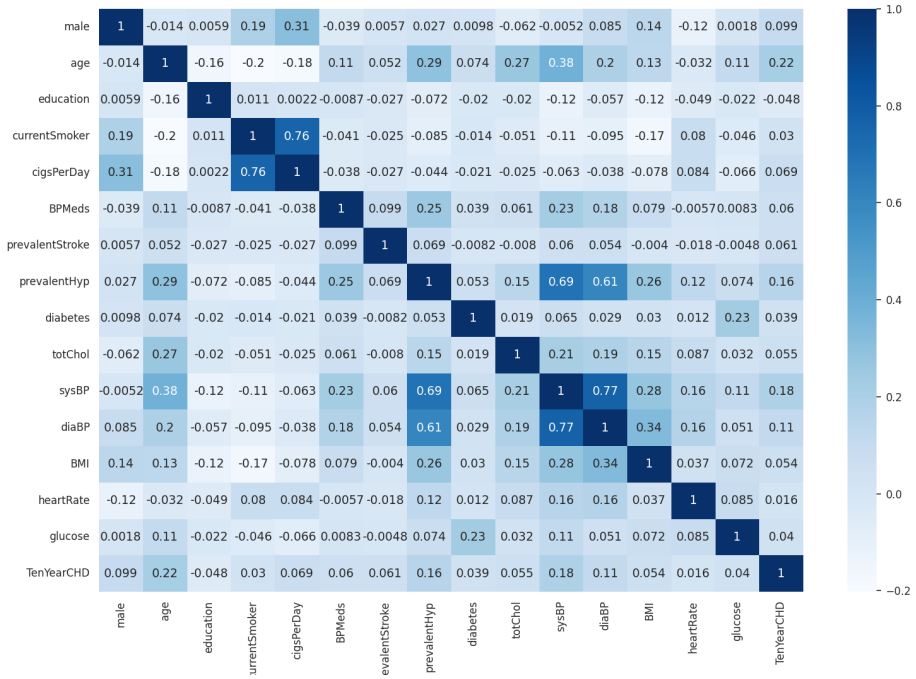


Figure 4: Correlation heatmap for the FHS dataset.

In Figure 4 the correlation heatmap between all features in the dataset is shwon. Correlations which can be read from this diagram are between the features cigsPerDay and currentSmoker, and also between sysBP and prevalentHyp, both of which are not surprising.

3 Methods

3.1 Re-Balancing, Scaling and Training and Testing split

Figure 5 shows the count of the 10-year CHD risk in the dataset. As can be seen from the figure, only 556 samples (13.3%) have a positive 10-year CHD risk, which makes the target class imbalanced. As a high degree of target class imbalance can lead to problems for the ML algorithms, it was chosen to apply the Synthetic Minority Over-sampling Technique (SMOTE) to re-balance the dataset [CBHK02]. After re-balancing the data, the target now contains an equal amount of samples for each class value.

As many ML models are also sensitive to data sample distances, each feature in the dataset was standardized to have a zero mean and a standard deviation of one. Before training, the dataset was split into a training and testing set using a ratio of 70:30 respectively. The split was performed using stratified sampling, assuring that the training and testing sets both contain an equal ratio of positive and negative samples from the target class (i.e. risk/no-risk).

3.2 Machine Learning Algorithms and Hyperparameter Optimization

Several machine learning algorithms were selected to predict the 10-year CHD risk, as well as three ensemble methods: boosting, voting and stacking. The following ML models were selected: Logistic Regression (LR), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Random Forest (RF). For the boosting ensemble method, the XGBoost algorithm was selected [CG16], and for both the voting and stacking ensemble a combination of the models SVM, kNN, and RF were selected as base-estimators.

Model hyperparameter optimization was done by performing a 5-fold cross-validated grid-search over a set of hyperparameters (model dependent), each within a predefined range.

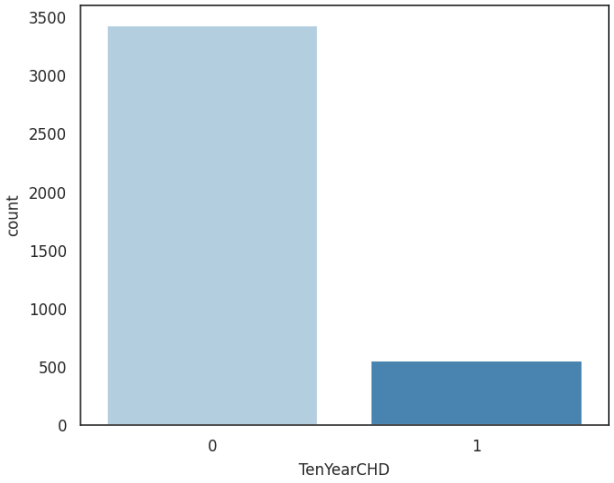


Figure 5: 10-year CHD risk count in the FHS dataset.

3.3 Performance Evaluation

After model training, the performance of the models was compared using various performance metrics: Confusion matrix, Accuracy, F1-score, the Area under Curve (AUC) of the Receiver Operating Characteristic (ROC)[4] curve, False Negative Rate (FNR) and False Positive Rate (FPR).

The confusion matrix for binary classification is a table of two rows and two columns, where the columns represent the model prediction and the rows represent the ground truth of the class label. An example of this is shown in Figure 6. True Positive (TP) and True Negative (TN) are correct model predictions, whereas False Positive (FP) and the False Negative (FN) are incorrect model predictions (i.e. errors).

Model predicts / Ground Truth	**Negative**	**Positive**
Negative	TN	FP
Positive	FN	TP

Figure 6: Confusion matrix for binary classification.

The model accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

The F1-score is defined as:

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{2}$$

For a perfect predictor the F1-score is 1 and the lowest score is 0.

The False Positive Rate (FPR), which is also called the 'false alarm rate', predicts a positive sample while actually it is negative, is defined as:

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

[4] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

The False Negative Rate (FNR), which is also called the 'miss rate', predicts a negative sample while actually it is positive, is defined as:

$$FNR = \frac{FN}{FN + TP} \tag{4}$$

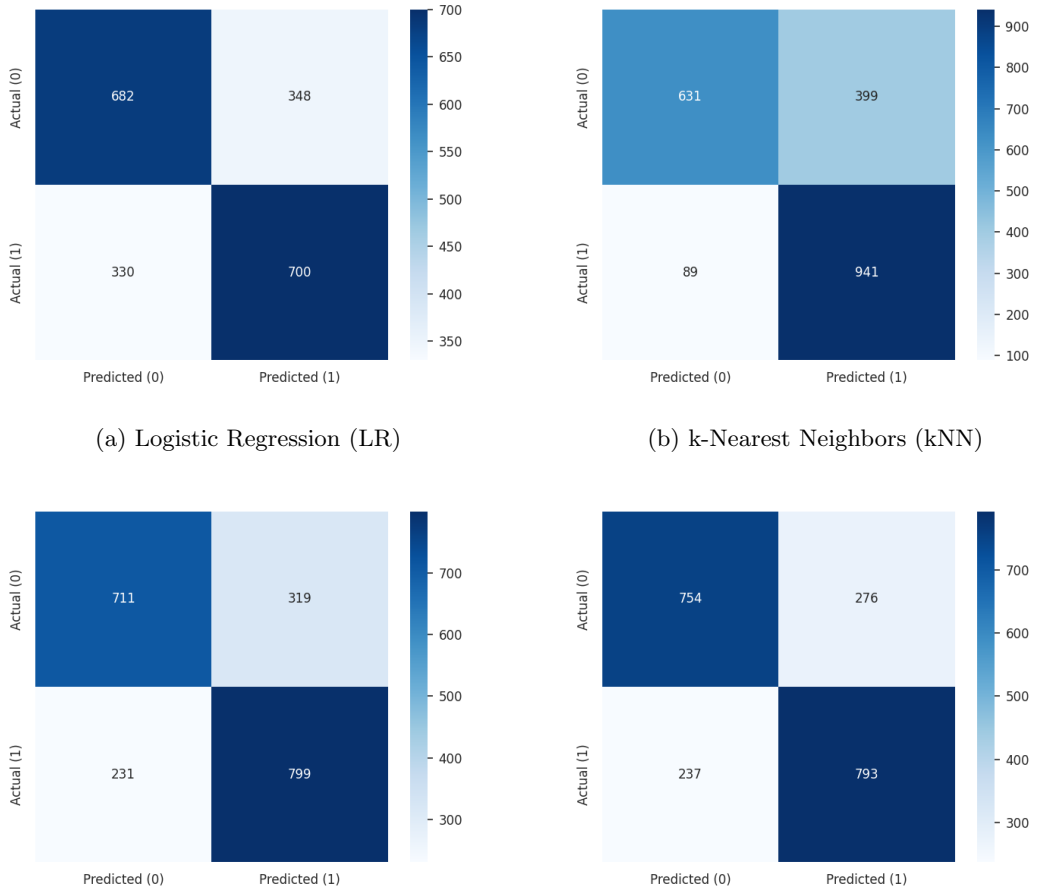
3.4 Model Interpretability

For the ML model with the highest performance score it was attempted to interpret the model by plotting the aggregated Shapley values of the dataset features. Shapley values provide an intuitive way to compute to what degree the different features contribute to a model prediction [RWB+22].

4 Results

4.1 Model Training

After model training, the confusion matrix showing the model prediction results on the testing set was plotted for each model. For the base models the result is shown in Figure 7. As can be seen from the figure, the majority of the base models have both a high number of false positives and false negatives, which can also be read back from the model performance metrics, summarized in Table 2. From these models, LR scores the lowest on accuracy (0.671), while the kNN model scores the highest (0.763). In general, all of these models have high false positive and false negative rates, with the exception of the kNN model, which has a significantly lower false negative rate (0.086).



(a) Logistic Regression (LR)

(b) k-Nearest Neighbors (kNN)

(c) Support Vector Machine (SVM)

(d) Artificial Neural Network (ANN)

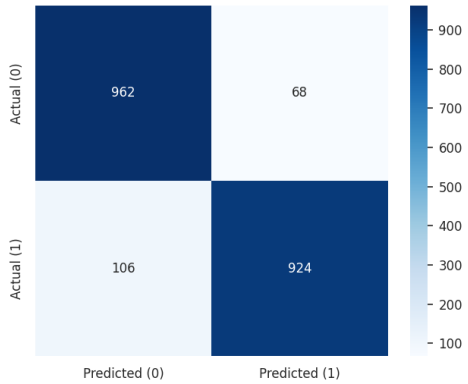
Figure 7: Confusion matrix (part I).

Looking at the confusion matrix for the ensemble methods (including the RF model), which is shown in Figure 8, we observe that the amount of false positives and false negatives decreases significantly as compared to the base models. The performance metrics for the ensemble models

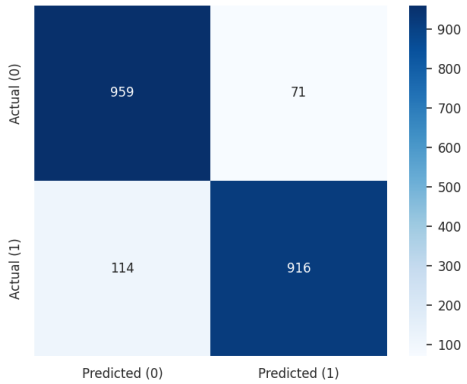
	Accuracy	AUC	F1-score	FNR	FPR
Logistic Regression	0.671	0.733	0.674	0.320	0.338
Support Vector Machine	0.733	0.817	0.744	0.224	0.310
k-Nearest Neighbors	0.763	0.856	0.794	0.086	0.387
Artificial Neural Network	0.751	0.837	0.756	0.230	0.268
Random Forest (Bagging)	0.916	0.971	0.914	0.103	**0.066**
Ensemble (Boosting)	0.910	0.962	0.908	0.111	0.069
Ensemble (Voting)	0.851	0.945	0.864	**0.056**	0.241
Ensemble (Stacking)	**0.925**	**0.979**	**0.925**	0.079	0.071

Table 2: Summary of model performance metrics.

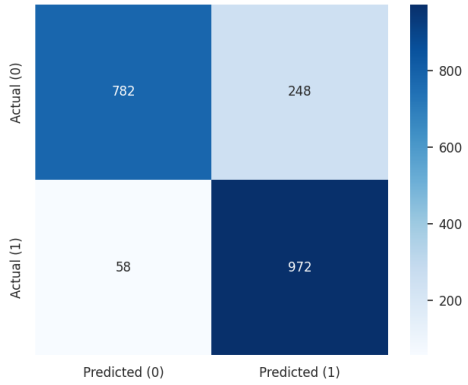
are also summarized in Table 2. In this case the best performing model with the highest accuracy is the stacking model (0.925), while the voting model has the lowest accuracy (0.851). For the ensemble models, the lowest false negative rate is found for the voting model (0.056), and the lowest False Positive Rate is found for the RF model (0.066). It should be noted that although the voting model has the lowest false negative rate, the number of false positives remains quite high, which helps explain the low accuracy score. To try and decrease the number of false positives, different combinations of base learners could be tried here. Although the stacking model does not have the lowest false positive and false negative rates, it is the highest scoring model overall, having the lowest number of errors in total.



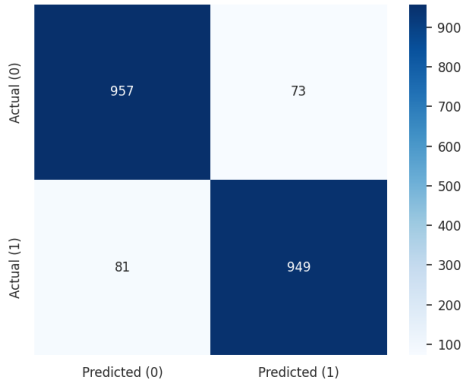
(a) Random Forest (RF)



(b) Ensemble (Boosting)



(c) Ensemble (Voting)



(d) Ensemble (Stacking)

Figure 8: Confusion matrix (part II).

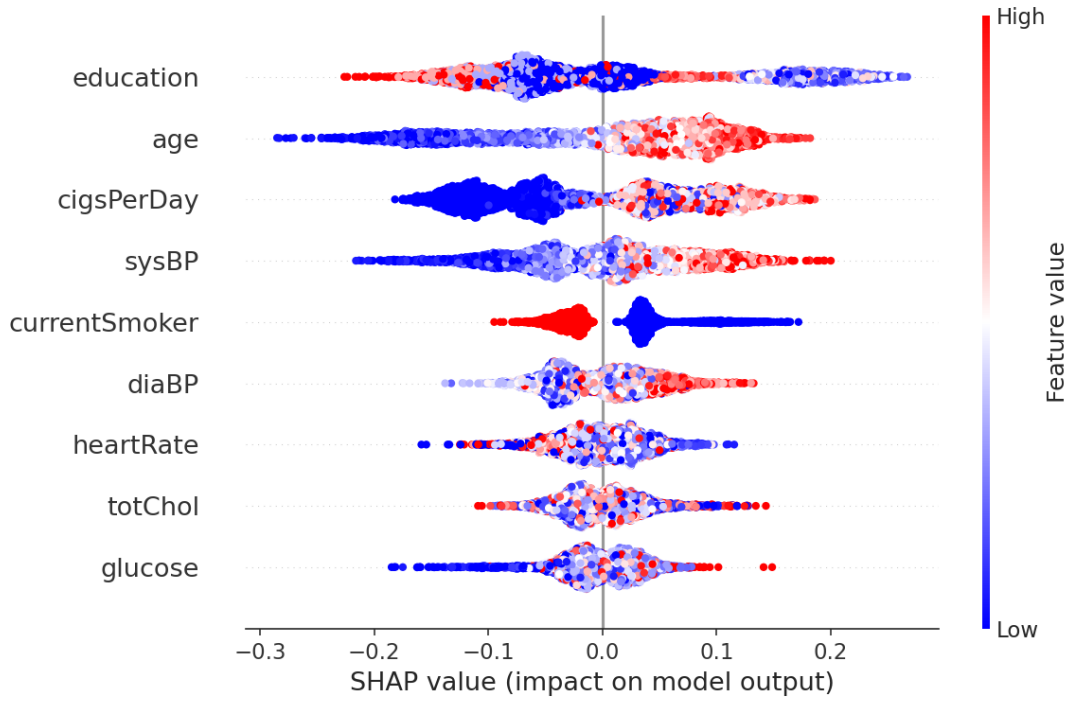


Figure 9: Shapley values for the RF model (contribution to 10-year CHD risk).

4.2 Model Interpretability

In Figure 9 the Shapley values for the RF model are shown. In the figure all the features have been ranked in order of total contribution to the 10-year CHD risk from high to low. So, a higher ranked feature on the y-axis of the figure indicates a higher total Shapley value summed over all datapoints. Per feature the figure also shows the contribution of each individual datapoint, where a datapoint with a high Shapley value is indicated by a red dot, while a low Shapley value is indicated by a blue dot. Reading from the figure a few points can be noticed.

First, the feature education has the highest total Shapley value, but the distribution within the feature is a mixture of high and low values, some positively contributing, and some negatively contributing to the 10-year CHD risk. In other words, it does not seem to indicate that a high or a low value of education alone directly contributes to the 10-year CHD risk. Although, for the extremes of the distribution (left and right tip) one could say that a low value adds to a increased 10-year CHD risk, while a high value for education lowers the contribution to the risk.

Second, the features age, cigsPerDay and sysBP (and also diaBP, but this is correlated with sysBP as was already seen in Figure 3) all have high values positively contributing to the 10-year CHD risk. This seems to indicate that high values for the features age, cigsPerDay and sysBP increases the chance of developing a CHD.

5 Conclusions and Discussion

In this project the possibility of predicting the 10-year CHD risk from a publicly available medical dataset was investigated. Starting with raw data in the preprocessing stage, missing data values were either imputed or dropped, and data outliers were removed. In order to facilitate the machine learning algorithms, in a second step the dataset was re-balanced and zero-centered, after which a selection of machine learning models was trained on the processed data. With the results of the training, the research questions as formulated in the introduction can now be answered:

A. **Can 10-year CHD risk be predicted by using classification ML models?**

   As was shown in the results section, basic ML models are able to predict the 10-year CHD risk much better than pure chance. When basic ML models are combined into ensembles, the prediction accuracy goes up even further.

B. **How do different ML models compare in predicting 10-year CHD risk?**

When looking at the basic ML models, the model with the lowest accuracy is the LR model, with an accuracy of 0.671. The best performing basic ML model is the kNN model, with an accuracy of 0.763. For the ensemble models, the lowest scoring model is the voting model, which achieves an accuracy of 0.851. The best performing ensemble model is the stacking model, which combined the RF, SVM and kNN models. For this model an accuracy of **0.925** is achieved. The lowest FPR ('false alarm rate') was found for the RF model (0.066), while the lowest FNR ('miss rate') was found for the voting model (0.056).

C. **Can 10-year CHD risk factors be identified from ML model interpretation?**

To interpret the ML models, the concept of Shapley values was used. When used on the RF model it appears that high values of the features **age**, **cigsPerDay** and **sysBP** contribute positively to the 10-year CHD risk. However, for the same model, it was also found that the feature education, which ranked highest in the cumulative feature contribution list, shows a much more diverse picture. Only at the extremes of the distribution does a low value directly contribute to the 10-year CHD risk (and a high value does exactly the opposite), while for the bulk of the distribution (i.e. the bulk of the population) the value can either contribute positively or negatively.

As was seen in section 2.1, the FHS dataset has quite a large number of missing values. If not dropped, these missing values were imputed using the median value of the feature. While simple, the major drawback of this method is that it does not consider the data point in relation to other data points. In order to further improve the ML models, it could be attempted to take data point relation into consideration, before imputing a value. In essence, if a data point belongs to a specific cluster, then imputing a value which more reflects the mode of this specific cluster is very likely to be a better guess for the missing value than the value which reflects the mode taken over the whole set of clusters (i.e. the whole population). A smarter way of imputing missing values has recently been investigated by Psychogyios et. al. [PIA22], which yielded quite promising results.

References

[CBHK02] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[KP16] V Kirubha and S Manju Priya. Survey on data mining algorithms in disease prediction. *International Journal of Computer Trends and Technology*, 38(3):124–128, 2016.

[PIA22] Konstantinos Psychogyios, Loukas Ilias, and Dimitris Askounis. Comparison of missing data imputation methods using the framingham heart study dataset. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5. IEEE, 2022.

[RWB+22] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*, 2022.