Predicting cardiovascular disease (Supervised classification)

Esther Olabisi-Adeniyi

Department of Integrative Biology

University of Guelph, Guelph, Ontario, N1G 2W1, Canada

ABSTRACT

Cardiovascular disease (CVD) can be caused by several genetic and lifestyle factors. Nevertheless, machine learning (ML) algorithms can be used to predict CVD and assess the relevance of different risk factors. This project aims to predict CVD based on age, cholesterol, smoking status, alcohol consumption, and physical activity status using gradient boosting, k nearest neighbor, random forest, ridge, and decision tree classifiers. Feature selection algorithms (SelectKBest and RFE) indicate that age, cholesterol, and smoking are the most important features for predicting CVD. The gradient boosting and ridge classifiers were the best models based on F1 scores and learning plots. The ridge classifier had the lowest computational cost and is, therefore, the preferred option for scaling up to larger datasets. These findings could prompt early prevention and intervention of CVD.

Key words (maximum 6 keywords):

machine learning, cardiovascular disease, gradient boosting classification, ridge classification, feature selection

List of abbreviations:

ML: machine learning, CVD: cardiovascular disease, CV: cross-validation, RF: random forest, KNN: K nearest neighbor classification, GBC: gradient boosting classification, DTC: decision tree classification, WHO: world health organization, RFE: recursive feature elimination.

1. INTRODUCTION

Cardiovascular disease (CVD) is a broad term that includes all diseases of the heart and blood vessels. It is the leading cause of death globally, accounting for an estimated 17.9 million deaths per year according to the World Health Organization (WHO) (WHO, 2021). Risk factors for CVD include smoking, high blood pressure, high cholesterol, diabetes, obesity, physical inactivity, and hereditary factors (Benjamin et al., 2019). While some risk factors such as family history cannot be changed, many can be modified through lifestyle changes and medications. Moreover, over 80% of CVD diseases are preventable (Salah and Srinivas, 2022); therefore, an understanding of the relationship between CVD and its risk factors could prompt early prevention and intervention. A machine-learning algorithm that predicts the presence of CVD could effectively model the connection between risk factors and the disease enabling individuals to assess their CVD risk and take appropriate action by improving their lifestyle or seeking medical counsel. Several studies have employed machine learning algorithms to examine the relationship between CVD and its risk factors (Pal et al., 2022; (Salah and Srinivas, 2022). Likewise, this project aims to adopt ML models in predicting CVD occurrence based on several risk factors: age, cholesterol, smoking status, alcohol consumption status, and physical activity status. The selected ML models are gradient boosting, k nearest neighbor, random forest, ridge, and decision tree classifiers. By building the aforementioned models to predict CVD, this study aims to answer the following questions:

(i) Which of the classification models would have the best performance when predicting CVD?

(ii) Which risk factors will emerge as the most relevant ones for predicting CVD?

2. MATERIALS AND METHODS

Data collection

This project uses a 'cardiovascular disease dataset' from *Kaggle.com* (The Devastator, 2023). The dataset records 12 different health metrics of 70, 000 individuals, including their cardiovascular disease status (presence or absence of CVD). The following measurements were taken: age, cholesterol, smoking status, alcohol consumption status, physical activity status, glucose level, gender, height, weight, systolic blood pressure, diastolic blood pressure, and CVD status. Meanwhile, only the first five measurements of the aforementioned list were selected to be features for predicting the latter measurement, CVD status. The reasoning behind this manual feature selection was based on an assumption that we could have a scenario where these five measurements were the only recorded features; and if so, "what would our results be?" However, it has now become clear that this manual feature selection step was not ideal and that all features should have been included in the study.

According to the source of the dataset, the variable measurements were recorded from individuals during medical examinations (Ulianova, 2019). Some of the variables were objectively considered (age, gender, height, and weight) while some were medically assessed (cholesterol, glucose level, systolic blood pressure, and diastolic blood pressure). The rest were recorded as provided by the patient. For the six variables used for this project, cholesterol was a categorical variable, age was numerical, and others were binary variables.

Data processing

Quality control of the raw dataset involved a series of steps carried out using Python programming language and libraries. Although the dataset appeared to be free of non-standard characters, I did not scan every single cell for such characters. Therefore, the most practical option was to import the dataset through "utf-8" encoding. I also used the *Pandas* package to analyze the dataset for missing values but none was found. Considering that, 6 out of 12 columns were selected for this project, the chances of having duplicate rows and confounding records was relatively high. Consequently, 31,475 duplicate rows were found and removed; 38,525 remained. I further inspected the remaining dataset for confounding records, that is, records with the same input values but different CVD statuses. Confounding records were detected and removed, narrowing the dataset down to 29,886 records. The ‘age’ variable was the only variable with a wide numerical range; therefore, it was examined for outliers using ±2IQR as the cutoff. No outliers were detected. By employing functions from the *Pandas* library, all binary variables were converted to dummy variables while ‘cholesterol’ was converted to an ordinal variable on a scale of 1 to 3 indicating cholesterol level: 1= normal, 2=above normal, 3=well-above normal.

Problem definition

In order to predict the target variable, CVD occurrence, based on the input variables, the ML models will be designed to solve a supervised classification problem. Each model will be trained with input variables and their corresponding CVD outputs. Afterwards, the models will be tested with unseen data.

Statistical analyses

The refined dataset contained 29,886 complete records (CVD+ : 15,917, CVD- :13,969). *Scikit-learn*'s splitting function was used to split the dataset into training and testing sets at a ratio of 80:20 ratio respectively. Afterwards, I applied two feature selection algorithms on the training set. I expect the two algorithms to give the same results; otherwise, only one set will be used. SelectKBest with *f_classif* was utilized in selecting the three most relevant features for output prediction. This option was used because it is designed for classification problems. The function computes ANOVA F-values for each feature and high scores indicate the best features. The second feature selection algorithm was Recursive Feature Elimination (RFE) with logistic regression. For this method, the least relevant features are removed during iterations of the logistic regression model (Guyon et al., 2002). I requested the best three features for RFE also.

Machine learning modelling

Ridge classification was chosen for this project because it is a regularization technique that efficiently handles the issue of multicollinearity between input and output variables, as is the case for this study (Arashi et al., 2021). I utilized the ridge classification function from *scikit-learn* which first converts output values from their binary state to {-1, 1} before applying ridge regression techniques (sklearn.linear_model.Ridgelassifier, 2023). The default settings for ridge classifier were used during training and testing.

Decision tree classification (DTC) is another powerful algorithm used for supervised learning problems. It works by recursively splitting the data into subsets based on the most significant features that provide the best separation between classes

(Bahzad Taha Jijo and Adnan Mohsin Abdulazeez, 2021). DTC can be used to predict categorical, numerical, and binary variables. S*cikit-learn's* DTC function was applied to the training and testing sets of this project. Parameters were set as follows: 'gini impurity' for quality criterion and 'best' for splitting strategy.

Gradient boosting classification (GBC) iteratively builds a series of decision trees, where each new tree is trained to correct the errors of the previous ones (Alexey Natekin and Knoll, 2013). The final model is an ensemble of these weak learners, with each tree weighted according to its performance on the training set. Gradient boosting is known for its ability to handle complex datasets and provide relatively accurate estimates of the target variable (Alexey Natekin and Knoll, 2013). I applied *scikit-learn's* GBC function during training and testing with 100 as the number of boosting stages.

The fourth classifier used in this project was k nearest neighbor classifier (KNN). KNN is popularly used for classification and regression problems. This model first computes the distance between the new sample and training sample and then identifies the K nearest data points for that new sample (Wang, 2019). Ultimately, a class label is assigned to the sample based on the majority class of those neighboring points. The chosen K for training and testing with KNN was five.

Random forest classification (RF), like GBC, incorporates several models in building an ensemble. It works by creating multiple decision trees on randomly sampled subsets of the data and features (Kulkarni and Sinha, 2013). The algorithm then combines the predictions of all the decision trees to obtain the final prediction. Contrary to the weighted technique used in GBC, RF uses a bagging method whereby the generated models are independent of each other (Kulkarni and Sinha, 2013). The number of trees for the RF function from *scikit-learn* was set to 100.

Subsequently, the ML models were evaluated based on five metrics: k-fold cross-validation (CV), balanced accuracy, F1-score, precision score, and learning curves. Firstly, a 10-fold CV of the training dataset was done with each classifier, after which the average values for balanced accuracy, F1-score, and precision score were computed to compare the models. A similar process was repeated for the testing dataset but with a *scikit-learn* function for predictions. Two types of learning curves were generated for each model. Test-training learning curves were plotted to analyze fitting during training and CV while 'training samples-fit time' curves (scalability of the model) were generated to analyze the computational costs required by each model.

3. RESULTS AND DISCUSSION

Feature selection

This project employed the use of SelectKBest and RFE in selecting the most important features for predicting CVD. Results for the two algorithms demonstrate that age, cholesterol, and smoking, in the order listed, are the three most important features for predicting CVD (Table 1). These results are consistent with previous findings of a ML study that highlights age and smoking as important predictors for CVD (Salah and Srinivas, 2022).

ML performance based on evaluation scores

Each classification model was evaluated by tenfold CV averages of balanced accuracy, F1, and precision scores. All three metrics rank the models equally during training (Table 2); however, since F1 score considers class imbalance, we will prioritize it during this

analysis. GBC was the top performer during training with an F1 score of 0.634 and ridge follows closely with 0.633 (Table 2). KNN, RF, and DTC have 0.569, 0.556, and 0.554 respectively. Upon testing with unseen data, GBC and ridge maintained their positions (0.639 and 0.638 respectively) (Table 2). The ML study by Salah and Srinivas also discovered that a gradient boosting model, XGBoost had the best performance when predicting CVD based on risk factors (2022), thereby corroborating my findings. GBC likely performed better than the other models because of its ensemble and boosting techniques, which are ideal for complex datasets like the mixed-variable dataset for this study. Ridge also performed approximately as well as GBC, probably because of its efficiency when multicollinearity is involved.

ML performance based on learning plots

In addition to analyzing F1, balanced accuracy, and precision for model performance, we can also examine learning and scalability curves. The learning curves for DTC, RF and KNN indicate that a great amount of overfitting exist between their training and CV scores (Fig 1). The low F1 score rankings of these models in addition to their high levels of overfitting suggest that the models require some modifications such as hyperparamter tuning in order to fit this dataset better. GBC and ridge also had overfitting and underfitting respectively, albeit relatively low. Moreover, GBC's curves show a rapid convergence of training and CV scores, further suggesting that GBC may be the better model than ridge (Fig 1).

The 'training samples-fit time' plots can help to make inference about the computational power required by each model (Fig 1). By understanding the computational costs of a model, we can infer how well the model would perform with similar cardiovascular

disease datasets of larger sizes. Fit time (measured in seconds) for each model is as follows (Fig. 1): ridge (0.09s), KNN (0.125s), DTC (0.6s), GBC (10.8s), and RF (32s). Although all fit times in the plots continue to increase as more samples are added such that none of them reach a plateau, the ridge classifier emerges with the lowest fit time and subsequently, the lowest computational cost. KNN and DTC have low fit times as well but their fitting performance has been weak so far. GBC, compared to its close second, ridge, requires more computational power, suggesting that in as much as GBC has had the best performance thus far, ridge may be the preferred option for scaling up to larger datasets. RF's high fit time is most likely because the model builds a large amount of trees when fitting samples. RF's computationally expensive model and weak performance makes for a bad CVD predictor in this study.

4. CONCLUSIONS

Despite the CVD still being a deadly illness with multiple risk factors, the potential for machine learning to predict this disease holds great promise in reducing its future incidence and fatalities. According to the findings of this project, which are in line with previous research outcomes, age, smoking habits, and cholesterol, are important predictors of CVD in ML (Salah and Srinivas, 2022). While these important predictors do not explicitly represent a causative or associative relationship with CVD, their roles as influential factors can contribute to other ML research and clinical studies for CVD. Furthermore, GBC was found to be the best performing classification model for predicting CVD, with ridge classifier as its close second. Ridge classification model is presented as the preferred option for larger datasets. Future work on this project should explore the applied models for potential improvements by employing several modifications such as hyperparameter tuning and feature engineering.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Dan Tulpan of the Department of Animal Biosciences for his instruction and support throughout this machine-learning course.

AUTHORS' CONTRIBUTIONS

All authors improved and contributed to the editing of the manuscript. All authors read and approved the final manuscript.

DISCLOSURES

The authors declare no real or perceived conflicts of interest.

LITERATURE CITED

Arashi, M., Mahdi Roozbeh, N.R. Hamzah, and M. Gasparini. 2021. Ridge regression and its applications in genetic studies. PLOS ONE. 16:e0245376–e0245376. doi:https://doi.org/10.1371/journal.pone.0245376. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245376

Alexey Natekin, and A. Knoll. 2013. Gradient Boosting Machines, A Tutorial. ResearchGate. Available from: https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial

Bahzad Taha Jijo, and Adnan Mohsin Abdulazeez. 2021. Classification Based on Decision Tree Algorithm for Machine Learning. ResearchGate. Available from: https://www.researchgate.net/publication/350386944_Classification_Based_on_Decision_Tree_Algorithm_for_Machine_Learning

Benjamin, E. J., P. Muntner, A. Alonso, Márcio Sommer Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das, F. N. Delling, Luc Djoussé, Mitchell S.V. Elkind, J. F. Ferguson, M. Fornage, L. C. Jordan, S. S. Khan, B. M. Kissela, K. L. Knutson, and T. W. Kwan. 2019. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation. 139. doi:https://doi.org/10.1161/cir.0000000000000659.

Guyon, I., J. Weston, S. D. Barnhill, and Vladimir Vapnik. 2002. Machine Learning. 46:389–422. doi:https://doi.org/10.1023/a:1012487302797

Kulkarni, V., and P. Sinha. 2013. Random forest classifiers: A survey and future research directions. Available from: https://adiwijaya.staff.telkomuniversity.ac.id/files/2014/02/Random-Forest-Classifiers_A-Survey-and-Future.pdf

Pal, M., S. R. Parija, G. Panda, Kuldeep Dhama, and R. K. Mohapatra. 2022. Risk prediction of cardiovascular disease using machine learning classifiers. Open Medicine. 17:1100–1113. doi:https://doi.org/10.1515/med-2022-0508.

Salah, H., and S. Srinivas. 2022. Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. Scientific Reports. 12. doi:https://doi.org/10.1038/s41598-022-25933-5.

Wang, L.-S. 2019. Research and Implementation of Machine Learning Classifier Based on KNN. IOP conference series. 677:052038–052038. doi:https://doi.org/10.1088/1757-899x/677/5/052038. Available from: https://iopscience.iop.org/article/10.1088/1757-899X/677/5/052038

World Health Organization. 2021. Cardiovascular diseases (CVDs). Available from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

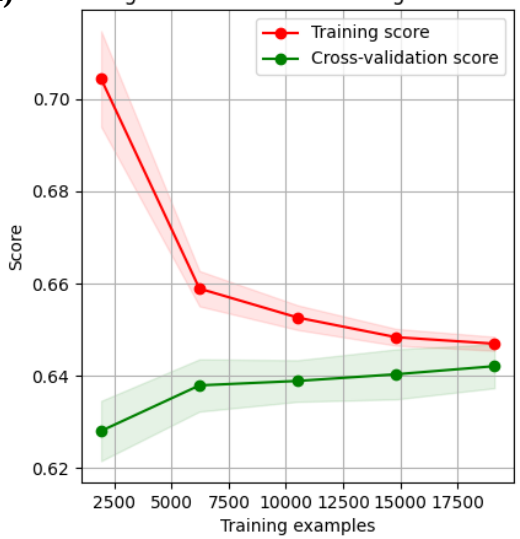
sklearn.linear_model.Ridgelassifier. 2023. scikit-learn. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridgelassifier.html

The Devastator. 2023. Risk Factors for Cardiovascular Heart Disease. Kaggle.com. Available from: https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas

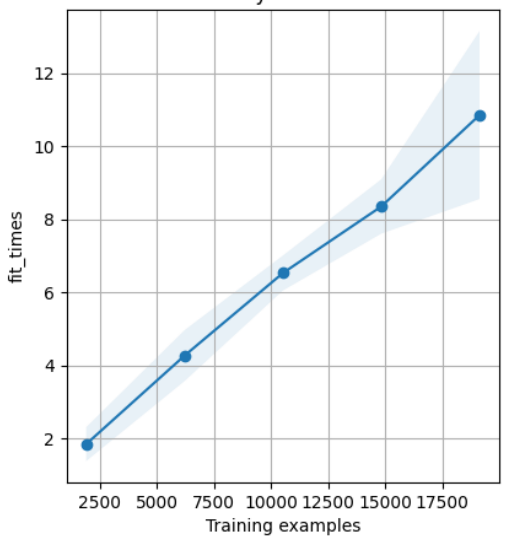
Ulianova, S. 2019. Cardiovascular Disease dataset. Kaggle.com. Available from: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

LIST OF FIGURES

(a)

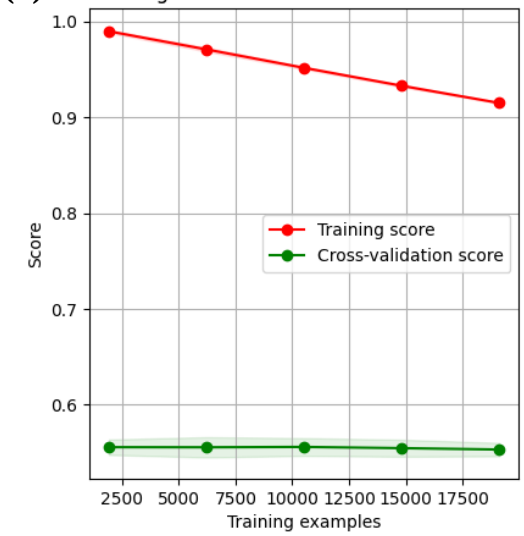

Learning Curves Gradient Boosting Classification
Training score
Cross-validation score
Score
0.70
0.68
0.66
0.64
0.62
2500 5000 7500 10000 12500 15000 17500
Training examples

(f)

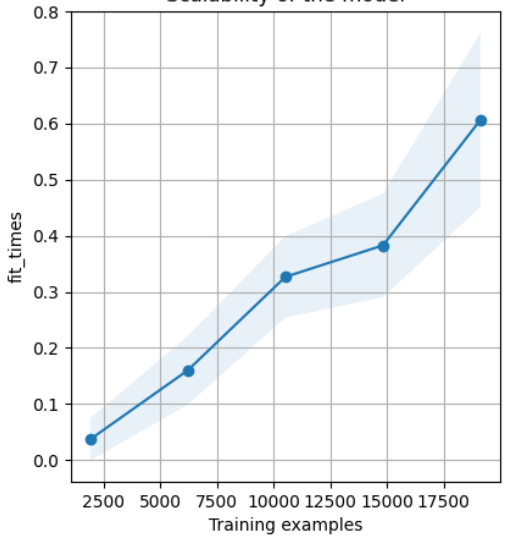

Scalability of the model
fit_times
12
10
8
6
4
2
2500 5000 7500 10000 12500 15000 17500
Training examples

226

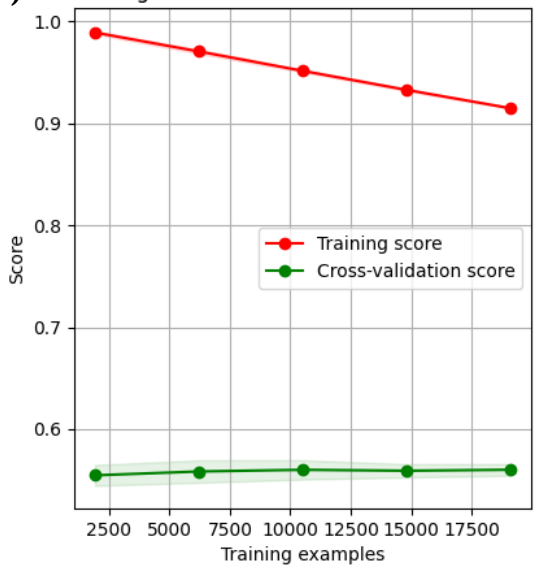
(b)


Learning Curves Decision Tree Classification
Training score
Cross-validation score
Score
1.0
0.9
0.8
0.7
0.6
2500 5000 7500 10000 12500 15000 17500
Training examples

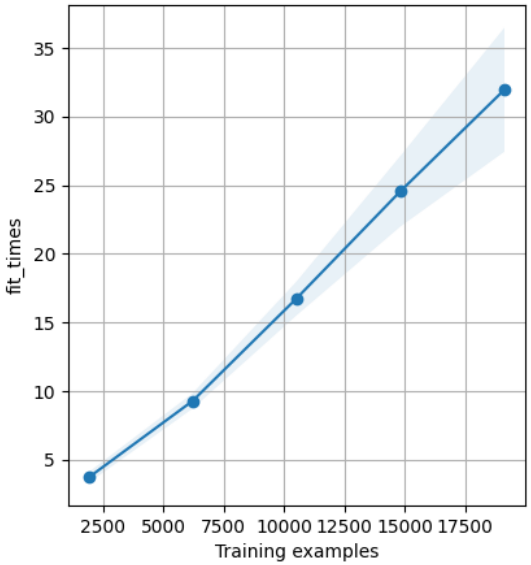
(g)


Scalability of the model
fit_times
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2500 5000 7500 10000 12500 15000 17500
Training examples

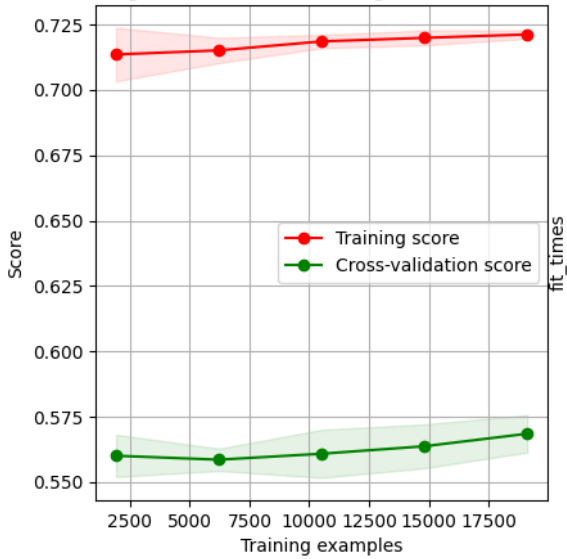
(c)


Learning Curves Random Forest Classification
Score
Training examples
Training score
Cross-validation score
1.0
0.9
0.8
0.7
0.6
2500 5000 7500 10000 12500 15000 17500

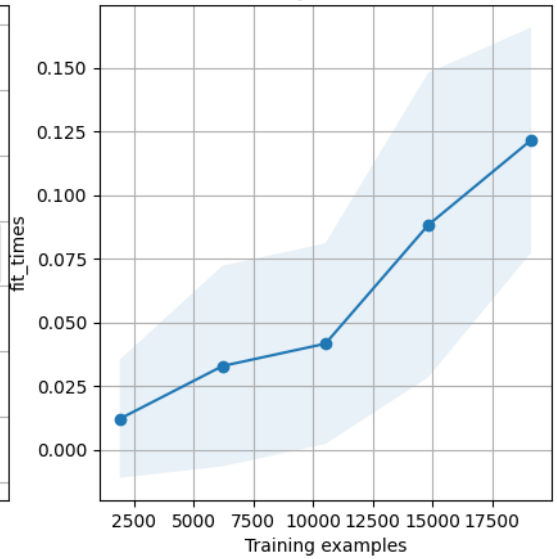
(h)


Scalability of the model
fit_times
Training examples
35
30
25
20
15
10
5
2500 5000 7500 10000 12500 15000 17500

(d)


Learning Curves K-Nearest Neighbour Classification
Score
Training examples
Training score
Cross-validation score
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
2500 5000 7500 10000 12500 15000 17500

(i)


Scalability of the model
fit_times
Training examples
0.150
0.125
0.100
0.075
0.050
0.025
0.000
2500 5000 7500 10000 12500 15000 17500

227

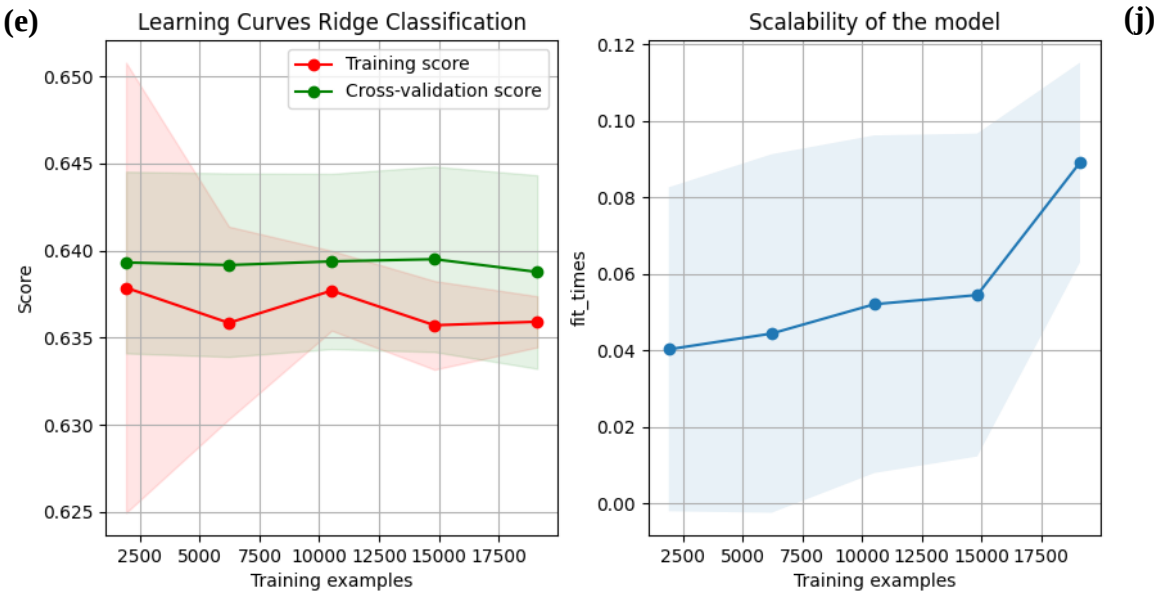


Fig. 1 a-e. Learning curves showing how each model's accuracy score changes depending on the number samples in the training set. A relatively high level of overfitting can be observed in **(b, c, d)** DTC, RF, KNN while **(a)** GBC overfits minimally**.** The ridge classifier **(e)** is underfitting the training set**.** The scalability plots demonstrate how the time (seconds) for fitting the training set changes as number of samples increase **(f-j)**. All the scalability plots show an increase in fit time as more samples are added. However, **(j)** ridge has the lowest fit time at about 0.09s while **(h)** RF has the highest at about 32s.

LIST OF TABLES

RFE

Rank	Feature	ANOVA F-values
1	Age	1371.4424
2	Cholesterol	1567.2153
3	Smoking	61.1316
4	Alcohol	31.0252
5	Physical Activity	17.9860

(a)

SelectKBest

Rank	Feature
1	Age
1	Cholesterol
1	Smoking
2	Alcohol
3	Physical Activity

(b)

Table 1. Feature selection rankings based on **(a)** RFE and **(b)** SelectKBest algorithms. The two algorithms rank the features similarly. Age, cholesterol, and smoking are the most important features for predicting CVD.

Training scores				
Rank	**Algorithm**	**Balanced accuracy**	**F1 Score**	**Precision**
1	GBC	0.635	0.634	0.639
2	Ridge	0.633	0.633	0.634
3	KNN	0.569	0.569	0.569
4	RandomForest	0.556	0.556	0.556
5	DTC	0.555	0.554	0.554

(a)

Testing scores				
Rank	**Algorithm**	**Balanced accuracy**	**F1 Score**	**Precision**
1	GBC	0.639	0.639	0.642
2	Ridge	0.638	0.638	0.638
3	RandomForest	0.563	0.564	0.564
4	DTC	0.562	0.562	0.562
5	KNN	0.543	0.543	0.543

(b)

Table 2 (a). Evaluation scores for each model's fitting of the training set. Highest to lowest performance (1 - 5). **(b).** Evaluation metrics for each model's performance when predicting testing dataset. Highest to lowest performance (1 - 5).