St. Clair College Applied Arts and Technology

DAB 304 Healthcare Analytics

Predicting Cardiovascular Risk Using Supervised Machine Learning Algorithms

Group Number: 44

Group members and IDs:

Hari Sai Palem	Manoj Reddy Lenkala	Naveen Kumar Tedla	Sunil Kumar Vidam
0747511	0753802	0753623	0735027

Abstract:

One of the significant causes of mortality around the world is heart disease, which is interchangeably called CVD, short for cardiovascular disease. These diseases are identified through medical tests, symptoms. Traditional paperwork for the medical records is rapidly going to be extinct and most of the medical institutions use digitalized data and while entry of these records manually, we cannot guarantee precision due to human errors. Also, the mortality rate due to this disease can be decreased if diagnosed earlier. This can be done using machine learning. The objective of this study is to predict the presence/absence of cardiovascular disease based on objective features like age, gender, weight and height, examination features obtained through medical test like cholesterol, glucose levels, systolic and diastolic blood pressures, and subjective features like smoking, alcohol, and physical tasks using machine learning algorithms such as Logistic Regression (LR), K Nearest Neighbours classifier (KNN), Support Vector Machine (SVM) classifier, Gaussian Naïve Bayes (GNB) classifier, Random Forest (RF) Classifier and XGBoost (XGB) Classifier. Two different split ratios have been used i.e., 75:25 and 80:20. The XGBoost model with best parameters achieved highest accuracy, random forest model with default parameters has the lowest accuracy for both the splits. The best model is Random forest model with best parameters as it has lower number of false negatives (type 2 error) than the XGBoost model with best parameters.

Introduction:

Cardiovascular disease (CVD) is one of the principle causes of death in men and women around the world. Under this umbrella, many diseases are included such as coronary artery disease, arrhythmias, and congenital heart defects. Patients are affected by cardiovascular disease mainly due to high cholesterol levels, blood pressure, diabetes, and consumption of alcohol. CVD is generally identified by symptoms such as chest pain/pressure, shortness of breath, irregular heartbeat, and medical examination features. As per the World Health Organization, in 2013, around 17 million people died due to cardiovascular diseases, of them 3 million died even before reaching 60 years age. It is believed that 90% of these deaths would have been preventable, if they were diagnosed in advance and the patients have followed appropriate measures and healthy habits. Though medical professionals/doctors use their knowledge to examine the records of the patient and detect the risk of cardiovascular disease, precision is not guaranteed due to human errors. Machine learning in healthcare aids the medical professionals in identifying the cardiovascular disease and treating the patients in better way.

The main purpose of this study is to predict the presence/absence of the cardiovascular disease in the patients, provided with objective, examination, and subjective features using the supervised machine learning algorithms such as KNN Classifier, SVM Classifier, Logistic Regression, Naïve Bayes Classifier, Random Forest Classifier, and XGBoost Classifier.

Related Work:

There exist numerous researches on this data. In this section, we would like to discuss one of the researches carried out on the same cardiovascular data we have chosen for this study. The title of the research is ‘Heart Disease Detection Using Machine Learning’ and the link to this research is provided in the references section. The authors of the research are Chithambaram T, Logesh Kannan N, and Gowsalya M from Vellore Institute of Technology. The authors created a correlation matrix and found that the correlation is strong between Cholesterol and glucose features. After segregating the dataset into independent features set and target data, scaling is performed on the independent features data using the StandardScalar(). Then, a comparison of machine learning models; SVM Classifier, Decision Trees Classifier, Random Forest Classifier, and KNN Classifier, is done to find the better performing model. The KNN Classifier has taken long time to process as this is a large dataset. This model has obtained an accuracy of 63.4%. Random Forest Classifier has been utilized to improve the accuracy. The accuracy of this model is 71.4%. The accuracy gained by the decision tree classifier model is 68.4%. The SVM Classifier model is built using two kernels; linear, gaussian. The SVM classifier with linear kernel has an accuracy of 72.5%, and with gaussian kernel the accuracy is 86.2%. It is mentioned that for the range of 1 to 11 neighbours, the KNN Classifier’s accuracy stands at 69.8%. Then the confusion matrices are plotted for SVM classifier and Random Forest Classifier. The rate of false positive and false negatives is higher in random forest classifier. After comparing the models, SVM classifier model built using ‘gaussian’ kernel is reported as the best model for predicting the cardiovascular risk in patients.

Dataset:

For this study, we have chosen the Cardiovascular Disease dataset to classify the presence and absence of Cardiovascular Disease in the patients. This dataset is free to access in the Kaggle community. Svetlana Ulianova published this dataset. There are 70000 samples with 11 features and a target. These features are categorized as; Objective, Examination, and Subjective features. Objective features are factual information like Age, Gender, Height, Weight. Examination features are the results obtained through the medical tests undergone by a patient i.e., Systolic and Diastolic blood pressures, Cholesterol, Glucose. Finally, the Subjective features are the information provided by the patients themselves like do they smoke, drink, and are they physically active. The target feature cardio which is the presence and absence of Cardiovascular Disease in the patients.

Data Preprocessing:

After reading the dataset into a dataframe, it is checked for missing values. No missing values are found in the dataframe. Then, the ID column is dropped from the dataframe and, is checked for duplicates. There are 3211 duplicate samples and these samples are dropped from the dataframe. The age feature has number of days since birth in the dataframe. It is converted to years by dividing the feature by 365. The gender feature is binary, 1 is for female and 2 is male. This feature is encoded by replacing 1, 2 with 0, 1 i.e., 0 is female and 1 is male. Then the dataframe is checked for outliers. Most of the outliers are found in the examination features, Systolic blood pressure and Diastolic blood pressure. All the outliers are removed.

Exploratory Data Analysis:

For the exploratory data analysis, a correlation matrix is plotted, and it is observed that there is a strong positive correlation between the systolic blood pressure and diastolic blood pressure and weak negative correlation between cholesterol and weight. A bar chart is plotted for the target feature and it is observed that there is a slight difference in samples for both the classes which need not be considered as a class imbalance problem. A boxplot is plotted for age and cardio (target) features. Patients who are elder have high risk of cardiovascular disease. A boxplot is plotted for age, gender, and cardio features. Women who are elder have high risk of cardiovascular disease. In another boxplot, weight and cardio are plotted and it is found that patients with overweight are more prone to the cardiovascular disease. Then, a clustered column chart is created for the features; cholesterol, glucose, smoke, alcohol, and activity. Most of the patients have cholesterol, glucose on level one. Also, there are more users with smoking, alcohol consumption habits and most of the users are physically active. In order to gain more insights, samples with patients having cardiovascular risk are selected. Then, the distribution of age is observed. Two boxplots are plotted; one for age and cardio features and another is for weight and cardio. From the first boxplot, it is observed that patients aged between 50 to 60 years are higher and in the second boxplot, patients weighing between 65 to 85 Kgs are more in number. Now, a bar chart is plotted for gender feature to find the number of samples for each of the genders. There are more females with the risk of cardiovascular disease than males. Then, alcohol consumption rate gender wise is shown by a clustered column chart. Females who consume alcohol are less in number than the males. Another clustered column chart is plotted to show the number of patients with smoking habit gender wise, and it is observed that more males are smokers than females. Also, there are females exhibited more physical activity than the males.

Methods:

First, the dataframe is divided into two sets; independent features set and target feature set. We have used two different split ratios; 75:25 (default) and 80:20, of the training and testing sets to build the models. Then, independent features set of data is scaled using the MinMaxScalar() to normalise it within the range (0,1).

Logistic Regression:

Even though the model name has regression in its name, Logistic Regression is a supervised classification algorithm. This model is good for a dataset with binary class.

For 75:25 and 80:20 splits of the data, the logistic regression model is imported and fit using the training data. Then, the metrics such as accuracy, precision, recall, and f-measure are computed. A confusion matrix is also plotted for the model built using 75:25 split of the data.

KNN Classifier:

KNN Classifier is a simple supervised machine learning algorithm where a value is picked for k (number of neighbours) and the data points are classified based on the Euclidean distance from test data point to each data point.

For KNN classifier model, we have considered 9 neighbours and the model is fit using the training data. Then, the testing set is passed to the model for prediction. After that, accuracy, precision, recall, f-measure are computed for both splits of the data, and a confusion matrix is plotted for the model built using default split (75:25) of the data.

Support Vector Machine (SVM) Classifier:

A Support Vector Machine classifier is one of the supervised machine learning algorithms that utilizes a boundary to separate the classes.

To build SVM classifier model for both default split and 80:20 split of the data, we have utilized the ‘linear’ kernel and assigned 1 for C. The model is built by fitting the training set. Then, accuracy, precision, recall, and f-measure are computed. A confusion matrix is also plotted for the model built using 75:25 split of the data.

Naïve Bayes Classifier:

Naïve Bayes Classifier is one of the supervised machine learning algorithms that works on the principle of conditional probability.

This classifier functions at low computation cost and works efficiently on large datasets. The following steps have been followed for building the models using 75:25 and 80:20 splits of the data. The Naïve Bayes classifier is imported and is fit with training data. Then, the testing set is passed to the model for prediction. Then, accuracy, precision recall, and f-measure of the model are computed, and a confusion matrix is plotted for the model built using the default (75:25) split of the data.

Random Forest Classifier:

Random Forest Classifier is an ensemble machine learning algorithm that generates a set of decision trees from randomly selected samples of the data.

First, the random forest classifier is imported, and the model is trained with the training set. Accuracy, precision, recall, and f-measure of the model are computed, and a confusion matrix is plotted for both default and 80:20 splits of the data. Then, hyper parameter tuning is implemented on both the random forest models by passing a parameter grid and 10 cross folds using GridSearchCV(). The accuracies of the models are computed, best parameters are obtained, and confusion matrices are plotted for the models.

XGBoost Classifier:

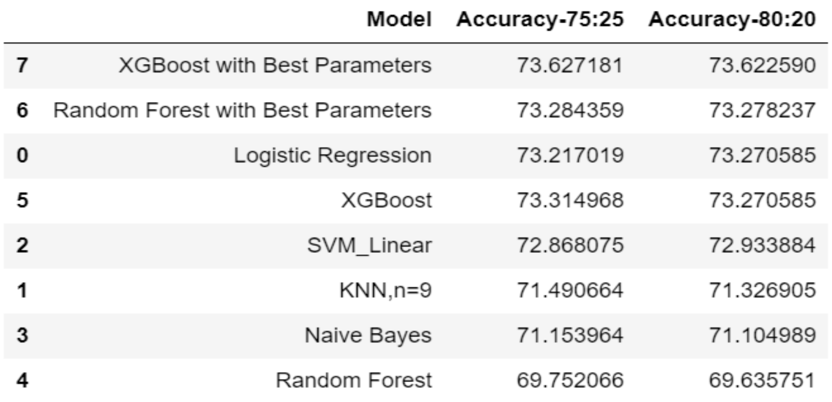
XGBoost Classifier is one of the supervised ensemble machine learning algorithms which utilizes a gradient boosting.

For both the splits (default and 80:20), the models are built in the following way. The XGBoost classifier is imported, and the model is fitted using the training set. Then, accuracy, precision, recall, and f-measure are computed, and confusion matrices are plotted for the model. Then, random forest classifier models using hyper parameter tuning are built by passing parameter grid and using 10 cross folds to GridSearchCV(). The accuracies of the models are computed, best parameters are reported, and confusion matrices are plotted for the models.

After building all the models, a dataframe with models and their accuracies for 75:25 and 80:20 splits of the data, is created and sorted in descending order of testing accuracies for 80:20.

Results:

The image below displays the dataframe that contains the names of the models we have built and their accuracies for the 75:25 and 80:20 splits of the data.



	Model	Accuracy-75:25	Accuracy-80:20
7	XGBoost with Best Parameters	73.627181	73.622590
6	Random Forest with Best Parameters	73.284359	73.278237
0	Logistic Regression	73.217019	73.270585
5	XGBoost	73.314968	73.270585
2	SVM_Linear	72.868075	72.933884
1	KNN,n=9	71.490664	71.326905
3	Naive Bayes	71.153964	71.104989
4	Random Forest	69.752066	69.635751

For 75:25 split:

The XGBoost classifier model with best parameters has obtained an accuracy of 73.63%. The best parameters for the model are {'max_depth': 5, 'min_child_weight': 2}. The accuracy for the XGBoost classifier model with default parameters is 73.31%. The random forest classifier model with best parameters for the model obtained an 73.28% and the best parameters are {'max_depth': 10, 'max_leaf_nodes': 50, 'min_samples_split': 50}. The accuracy for the random forest classifier model with default parameters is 69.75%. The logistic regression model has exhibited an accuracy of 73.22%. The accuracy of the model built using SVM classifier is 72.87%. The KNN Classifier model built using 9 neighbours has an accuracy of 71.5%. The accuracy of the Naïve Bayes classifier model is 71.15%.

For 80:20 split:

For this split, the accuracies for XGBoost classifier model with best parameters and default parameters are 73.63% and 73.27% respectively. The best parameters are {'max_depth': 10, 'max_leaf_nodes': 50, 'min_samples_split': 50}. The accuracies for Random forest classifier model with best parameters and default parameters are 73.28% and 69.64% respectively and the best parameters are {'max_depth': 10, 'max_leaf_nodes': 50, 'min_samples_split': 50}. The accuracies for the models; logistic regression, SVM Classifier, KNN Classifier, and Naïve Bayes classifier, are 73.27%, 72.93%, 71.33%, and 71.1% respectively.

Discussion:

The objective of this project is to predict the risk of cardiovascular disease in patients based on medical test features such as cholesterol and blood pressure, subjective features like age, gender, weight and height, and factual features like alcohol consumption, smoking, and physical activity. The highest accuracy is obtained by the XGBoost Classifier model with best parameters in case of both the split ratios followed by the Random Forest Classifier model with best parameters. The lowest accuracy is reported for the Random forest classifier model with default parameters. The first challenge we faced is performing Exploratory Data Analysis. For better understanding of the data and to find more insights, we have extracted the samples with the risk of cardiovascular disease and plotted charts for features like age, gender, alcohol consumption and smoking. Another challenge is identifying outliers in all features and removing them from the dataframe. For both the split ratios of data, the XGBoost classifier and Random Forest classifier models with default parameters have not exhibited much accuracy. Improving the accuracy of those models using GridSearchCV() has been another challenge for this project.

Conclusion:

Though the accuracy is higher for the XGBoost classifier model with best parameters in 80:20 split ratio, the number of false negatives is lower for the Random forest classifier model with best parameters. So, random forest classifier model with best parameters is the best model for predicting the presence/absence of cardiovascular disease risk. This prediction helps the doctors in treating the patients with better care and patients to take measures in advance to reduce the risk of getting affected by cardiovascular disease. Provided few more examination features for this dataset, the models can be enhanced with more accuracy.

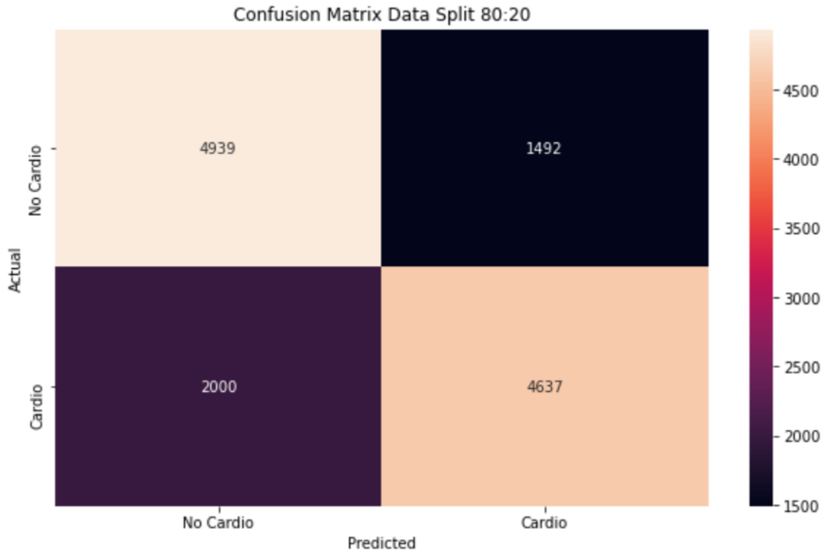


Fig: Confusion matrix for best model (Random Forest with best parameters).

Contributions:

Student	ID	Contribution
Hari Sai Palem	0747511	I have performed Data Preprocessing (checked for missing values, duplicates, outliers) and Exploratory Data Analysis, added comments to the jupyter notebook and helped in documenting the Abstract, Introduction and Related work sections of the project report.
Manoj Reddy Lenkala	0753802	I have built the KNN classifier and SVM classifier model. I helped in documenting Dataset, Data preprocessing, and EDA sections of the report.
Naveen Kumar Tedla	0753623	I have built the random forest classifier and XGBoost and performed hyper parameter tuning to improve the accuracy and obtain best parameters. I helped in documenting Methods section of the report.
Sunil Kumar Vidam	0735027	I have built the logistic regression, Naïve Bayes classifier models. I have helped in documenting the Results, Discussion, and Conclusion sections of the project report.

References:

- Dataset Source: *https://www.kaggle.com/sulianova/cardiovascular-disease-dataset*.
- GridSearchCV for hyper parameter tuning: *https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html*.
- Scikit learn: *https://scikit-learn.org/stable/*.
- Cardiovascular disease facts: *https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118*.
- Related Works: *https://assets.researchsquare.com/files/rs-97004/v1/6a23d55b-f6a9-44e7-aa6f-eebec260402c.pdf*.

Appendices:

- DAB 304 - Healthcare Analytics Final Project - Group 44.ipynb

  DAB 304 - Healthcare Analytics Final Project - Group 44.ipynb

  ➢ This file contains all the code for reading the data, EDA, Data Preprocessing and models building and evaluation.
- Dataset File:

  CVD_risk.csv

  ➢ Contains all the features and target data for Cardiovascular disease.