RESEARCH

Introduction to Focus Areas in Bioinformatics Project Week 1

Sina Glöckner[*†], Christina Kirschbaum[†], Swetha Rose Maliyakal Sebastian[†] and Gokul Thothathri[†]

[*]Correspondence:
sina.gloeckner@fu-berlin.de
Institute for Informatics, Freie Universität Berlin, Takustr. 9, Berlin, DE
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Goal of the project: Use any three classifiers and compare and evaluate them, to diagnose coronary heart disease.

Main results of the project: We observed that Random Forest provided the highest accuracy after comparing results of the k-nearest Neighbours algorithm, Support Vector Machine and Random Forest for the given model .

Personal key learnings: Machine Learning algorithms can identify the risk of a disease given any medical data.

Estimation of the time: 16 hours

Project evaluation: 2

Number of words: 1260

1 Scientific Background

One of the leading causes of disease and mortality among the world's population is cardiovascular disease. The most important subject in the domain of clinical data analysis is cardiovascular disease prediction. However, because of multiple contributing risk factors such as diabetes, high blood pressure, elevated cholesterol, abnormal pulse rate, and many others, it can be difficult for medical doctors to detect cardiac disease, because of these constraints data mining or machine learning is not favoured by scientists.

The data obtained from the Cleveland database [1, 2] includes 303 patients which are categorised under 14 distinctive parameters such as age, sex, chest pain to predict Coronary artery disease (CAD). CAD is the narrowing of the heart's arteries do to fat build up or because of plaque. Due to this the heart doesn't get enough blood when a person is carrying out strenuous work, which causes chest pain and breathlessness resulting in a myocardial infarction (heart attack). [3]

2 Goal

The main purpose of the project was to train and compare the data using three classifiers for the diagnosis of Coronary Artery Disease.

3 Data and Preprocessing

For $n = 303$ patients, this data set defines 14 attributes. There is information about each patients age and sex. Furthermore, the patients chest pain is divided into four types, typical angina, atypical angina, non-anginal pain and no chest pain at all. Some data about the patients blood is available as well. This includes resting

blood pressure, serum cholesterol, fasting blood sugar. The blood and haemoglobin electrophoresis screening tests are represented as normal, fixed defect, or reversible defect. And the number of major blood vessels coloured by fluoroscopy is given. Additionally, the electrocaridagraphic results were classified into normal, a ST-T wave abnormality, and hints for left ventricular hypertrophy by Estes' criteria. More information about the patients heart condition is given in the maximum heart rate. The electrocardiographic results during exercise ere reflected if there was a ST depression induced, and further analyzed as upsloping, flat or downsloping.[2] Information if the patients suffers from exercise induced angina are featured, too. The "target" field indicates whether or not a patient has cardiac disease. It has a value between 0 (no presence) to 4 (present).

The null values were eliminated by replacing them with valid values. Only in "thal" and "ca" null values existed. They were replaced with the means for each column. The values in the target column were transformed to 1 from 2 to 4. This steps were done by using the library pandas [4]. Afterwards, the target column was asigned as label, while the remaining 13 columns were defined as features. Lastly, the dataset was split in a training set and in a test set with sklearn [5].

4 Methods

The three used classifiers are the k-nearest neighbours classifier, the support vector machine, and the random forest classifier. They were applied to the data set and compared with the Python library sklearn [5].

4.1 K-Nearest Neighbour

The k-nearest neighbor classifier is premised on the idea that similar occurrences should have equivalent labels in classification or target values in case of regression. This algorithm is effective, but it can actually understand sets which are not linear with definite boundaries and regression features that are immensely complex. The method, on the other hand, is resource expensive and inclined to generalization or overfitting. [6]

4.2 Support Vector Machine

Support vector machine (SVM) is a typical classification approach combined with regression analysis for the given data. The classifier is a supervised learning strategy that evaluates data by splitting it into two distinct categories. The outcome of this algorithm is a layout map in which the arranged dataset with the boundaries between the categories is maximum. [7]

A kernel is used to incorporate the support vector model in training the dataset. An input feature space is transformed into the specified form by a kernel. The kernel function is a method used by an algorithm. The kernel converts the dimensional input storage from low to high in this case. To define it in another way, it turns complex undividable problems into simple dividable problems by introducing additional dimensions to them. It is essential in situations concerning arbitrary models. The kernel trick aids in the development of a more precise solution for the given model. [8]

4.3 Random Forest

Random forests is an approach for supervised learning, it is used for classification and regression because of the large number of decision trees involved in the process [7]. This classifier is a collection of multiple random decision trees that are less sensitive to the training data, since we use many trees it gets the name Random Forest [9]. Random Forest chooses random rows from the dataset to create a new training dataset using Bootstrapping, here the features used are ca, thal, target. Then a new data point is created and then passed through each tree one by one and the predictions are recorded. All the predictions are then combined and then since it is a classification problem, majority voting is used, and this process of combining results from multiple models is called aggregation. The combined process of bootstrapping and aggregation is called bagging. The random feature selection helps to reduce the correlation between the trees [10].

5 Results

For the KNN classifier, a Sensitivity of 65.38% and a Specifcity was 71.43% calculated. The AUC score was 0.70 and the F measure reached 0.64. The overall accuracy falls at 68.85%.

The Sensitivity of SVM was lower at 26.92% but a higher 80.00% Specifity could be found. The AUC score was 0.66 and the F measure 0.35. The overall accuracy is slightly less than kNN at 57.38%.

Lastly the random forest model reached a Sensitivity of 57.69% and a Specifity of 88.57%. It has the highest AUC score with 0.80 as well as the highest F measure 0.67. This also results in the highest overall accuracy at 72.13%.

A comparison for all accuracies is visible in Figure 1. The ROC curves for all algorithms can be seen in Figure 2.

6 Discussion

A comparative study was performed for the RF, KNN, and SVM classifiers based on their accuracy in the detection of coronary artery disease to help identify patients with high risk factors. With the results obtained for this Coronary Artery Disease Prediction, we can conclude that the Random Forest method outperforms the SVM and KNN algorithms.

This project has some key data-science elements to be addressed. First we took real-time data, followed by all preliminary data-preprocessing steps. With the known algorithmic knowledge, we performed a disease prediction with the model. Three distinct machine learning techniques were implemented to get the accuracy scores. The project omits all bioinformatics requisites by seeing the input as merely as a typical data scientist but not as a bioinformatical entity.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Christina Kirschbaum and Sina Glöckner implemented the code in Python. They also worked on the report. Swetha Rose Maliyakal Sebastian wrote the report. Gokul Thothathri researched different classifiers and worked on the implementation.

References

1. Dua, D., Graff, C.: UCI Machine Learning Repository (2017). http://archive.ics.uci.edu/ml
2. Andras Janosi, M.P.R.D. William Steinbrunn: Heart Disease Data Set (1988). https://archive.ics.uci.edu/ml/datasets/heart+disease
3. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K.H., Lee, S., Froelicher, V.: International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology **64**(5), 304–310 (1989). doi:10.1016/0002-9149(89)90524-9
4. Reback, J., McKinney, W., jbrockmendel, den Bossche, J.V., Augspurger, T., Cloud, P., gfyoung, Sinhrks, Hawkins, S., Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., Garcia, M., Schendel, J., Hayden, A., Saxton, D., Jancauskas, V., McMaster, A., Battiston, P., Seabold, S., patrick, Dong, K., chris-b1, h-vetinari, Hoyer, S., Gorelli, M.: Pandas-dev/pandas: Pandas 1.1.5. doi:10.5281/zenodo.4309786. https://doi.org/10.5281/zenodo.4309786
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
6. Asogbon, M., Omisore, O., Samuel, O., Ojokoh, B., Dahunsi, O.: Comparative analysis of knn and svm classifiers for students' academic performance prediction. (2016)
7. Hariharan, K., Vigneshwar, W.S., Sivaramakrishnan, N., V, S.: A comparative study on heart disease analysis using classification techniques. International Journal of Pure and Applied Mathematics **119**, 13357–13365 (2018)
8. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, Fourth Edition. Academic Press, ??? (2009)
9. Ho, T.K.: Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–2821 (1995). doi:10.1109/ICDAR.1995.598994
10. Phan, T.-N., Kappas, M.: Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. Sensors **18**, 18 (2017). doi:10.3390/s18010018

Figures

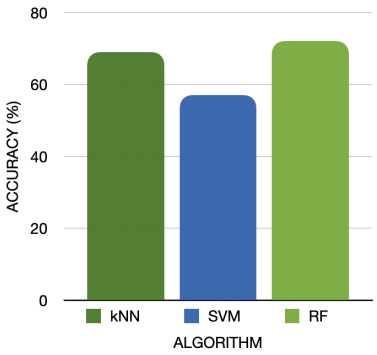

Figure 1 Accuracy Graph The diagram shows a comparison of the accuracy for all three classifiers.

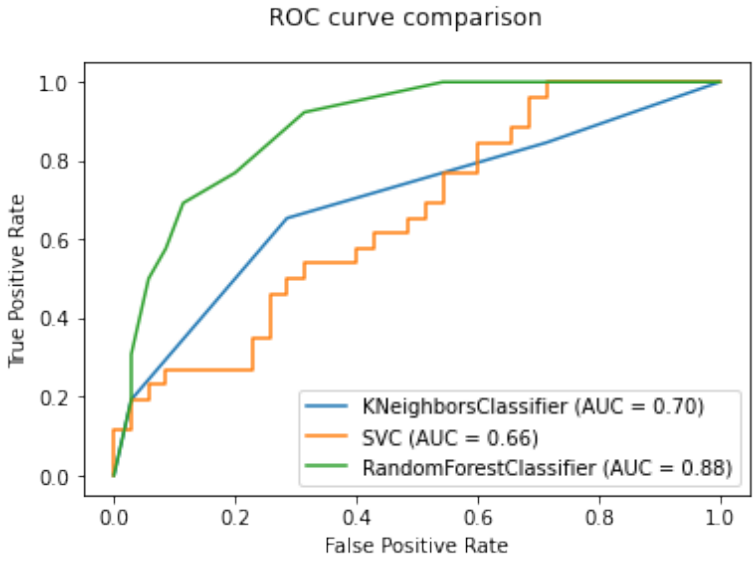

Figure 2 ROC Graph The graph shows a comparison of the ROC curves of all three classifiers.

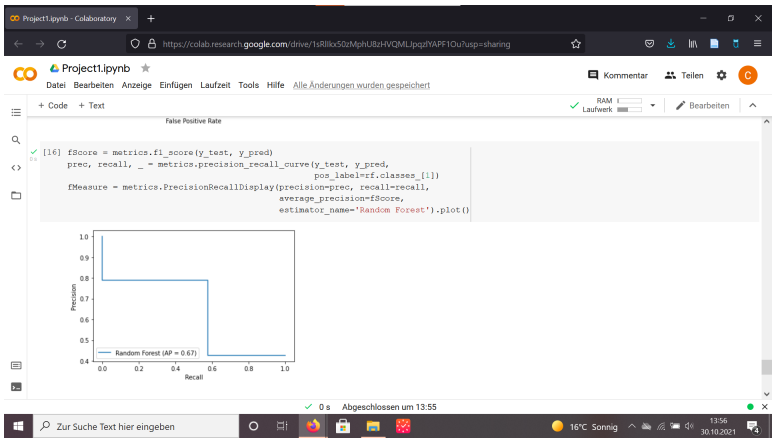

Figure 3 Screenshot from Run