RESEARCH

Introduction to Focus Areas - Project 1

Dominik Bannwitz*, Florian Herzler, Maximilian Otto and Mariana Steffens

*Correspondence:
dominik.bannwitz@fu-berlin.de
Department of Mathematics and Computer Science, Free University of Berlin, Takustraße 9, 14195 Berlin, Germany
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Goal of the project: Perform an exploratory analysis and develop three classifiers for predicting the diagnose of heart disease.

Methods used in the project: As classification models, Logistic Regression, K-nearest neighbor and Decision Trees were implemented and evaluated.

Main results of the project: The Logistic Regression classifier was accurate at predicting the target groups with an accuracy of about 87%.

Personal key learnings: We implemented the classifiers Logistic Regression, K-nearest neighbor and Decision Trees for the first time.

Estimated working hours: 54

Possible improvements: Techniques to avoid overtraining of Decision Trees can be explored.

Keywords: Classifiers; Logistic Regression; K-nearest Neighbor; Decision Tree

1 Scientific Background

Computers are being used to support decision-making in more and more areas. In some, they have already taken over the decisions completely - in others, it is hard to imagine what the decision would be like without their support. Taking the example of autonomous driving, we can observe what decisions the computer has to make. The car's camera recognizes a traffic signal and must now decide whether it should stop at the intersection or whether it can drive over it. To be able to decide this, the car's computer must have previously learned what license plates exist and now classify the seen license plate as one of them. To understand how this works, we will introduce the concept of classification.

In data science, classification is the process of predicting which of a set of classes an observation belongs to. Many different algorithms for classification have been developed over the years. There are more "traditional" classification approaches such as k-nearest neighbor(KNN), logistic/linear regression, or decision trees, as well as more "modern" algorithms such as neural networks and deep learning. What all these methods have in common is that they must first be trained on a training data set before they can be evaluated on another test data set. Therefore, a considerable amount of data must be accumulated and categorized by some other program or expert. The data set must then be divided into a training and test data set, and a classifier must be chosen, considering its characteristics, benefits and limitations. The process of training the classifier often requires repetitions to adjust weights or other parameters until a certain accuracy is achieved or a set number of repetitions is reached. Once the classifier is trained, testing is required. To do this, we have to

run the learned classifier on the test dataset and then choose sensible evaluation parameters. For example, recall and precision or the F1-index, as well as the ROC-Curve (Receiver Operating Characteristic) and AUC (Area under Curve) are often used to evaluate the ability of the trained classifier.

2 Goal

The goal was to perform an exploratory analysis and develop three different classifiers for diagnosing heart disease based on the "Cleveland Heart Disease" data set, and then analyze and compare the performance of each classifier.

3 Data and Preprocessing

The data was collected from 303 patients referred for coronary angiography disease (CAD) at the Cleveland Clinic, in Cleveland, Ohio, between May 1981 and September 1984 [1]. The original data set is available at the "UCI Machine Learning Repository"[1] and contains 303 instances and 76 attributes of which 14 were used. The "goal" field refers to the presence of heart disease and is an integer value from 0 (no presence) to 4.
As a first step of the data preprocessing, the table was read as a data frame and all the values were converted to float, except the column "goal", which is categorical data and was converted to string for better visualization. The next step was to find out the amount of missing values in the data set and since it returned a small quantity, representing only 2% of all rows, the missing values were imputed as the mean-value of their respective column. The imputation of the missing values with the mean was sufficient in this data set, due to the non-significant changes in the variance.
The task was divided into two phases. In the first phase, it was required that the classifiers were trained based on the categorical data "goal", which could vary from presence (values 1,2,3,4) to no presence (value 0). In the second phase, the categorical data "goal" was transformed into a binary outcome "no heart disease vs. heart disease" (0 and 1) and each classifier was trained once more based on the new class labels.

4 Methods

For the purpose of exploratory analysis, descriptive functions for data frame, pair plots and heatmaps were used in order to have first insights of the data, understanding each variable and studying possible correlations between the attributes.
Then, in order to develop three classifiers to predict CAD and subsequently evaluate them, the data set was split into 80% training and 20% test set, using the inbuilt function `.sample()` of the `random` module in Python. Using the machine learning library `scikit-learn`[2], three basic classification solutions were implemented: Logistic Regression, K-Nearest Neighbor and Decision trees. All classifiers were trained using the same training set. For evaluation and comparison purpose,

[1]https://archive.ics.uci.edu/ml/datasets/Heart+Disease

a Confusion Matrix, Classification Report and ROC Curves were generated using the library `scikit-learn`.

5 Results

In the first phase of the task, it was observed that the data was unbalanced, since the amount of "no presence" values was significantly higher than the other categories referred to the presence of heart disease. This discrepancy can be seen in Figure 1.

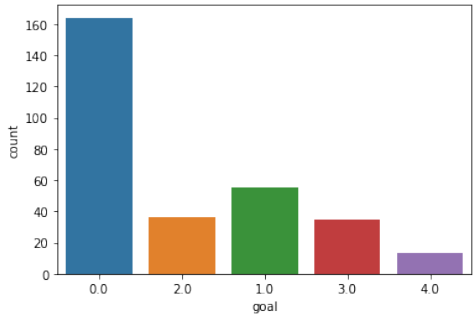

Figure 1 Imbalanced classification in the Phase 1

The classifiers were first trained on the imbalanced data and the results can be seen in 2. In terms of accuracy, the Decision Tree Classifier presented a train accuracy of 100%, which leads to the inference that this prediction model was overtrained because the test accuracy is much lower at 60,7%. The Logistic Regression and the K-nearest neighbor showed a similar result, with an accuracy of 64,9% (train) and 62,3% (test) and 67,4% and 57,4% (test), respectively. Considering the precision and recall outcome, the classifiers showed a similar pattern, with K-nearest Neighbor in the last position and the classifiers Logistic Regression and Decision Tree with a close performance.

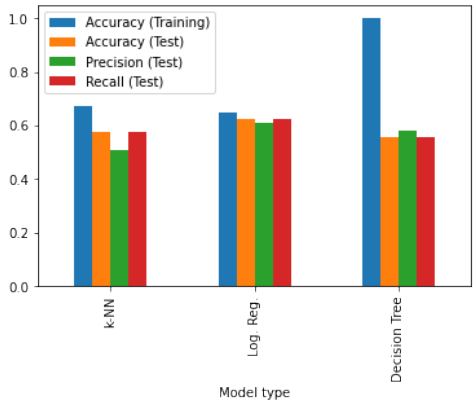

Figure 2 Performance of classifiers in Phase 1

In the second phase of the task, the data set was balanced, as the minority classes 1, 2, 3 and 4 were merged into only one class describing the presence of heart disease. Therefore, the category "goal" was converted into a binary class, in which values could take 1 for the presence of heart disease or 0 for no presence. The new classification can be seen in Figure 3.

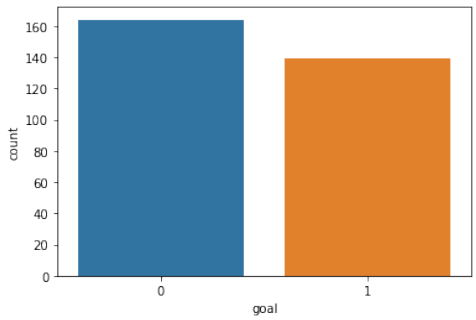

Figure 3 Balanced classification in Phase 2

The classifiers were trained based on the new target group, the results can be found in Figure 4. The classifiers perform better compared to the previous phase. The result of the accuracy during the training for the Decision Tree suggests that the model again overfitted the data, while K-nearest Neighbor and Logistic Regression presented accuracies of 83.9% and 84.7% during the training.

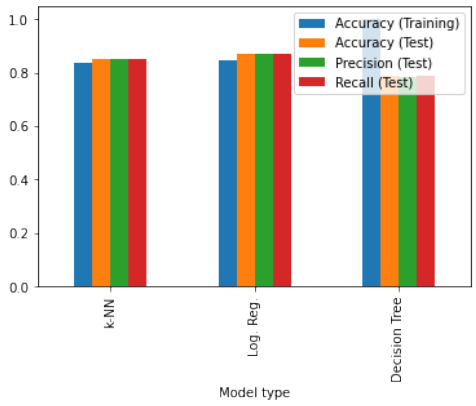

Figure 4 Performance of classifiers in the Phase 2

With regard to the test data set, the Logistic Regression provided the best accuracy around 86.9%, followed by K-nearest Neighbor with 85.2% and Decision Tree with 81.2%. In terms of precision and recall, the Logistic Regression presented a level of 87% and 87%, respectively, against 85% precision and 85% recall from K-nearest Neighbor and 82% precision and 82% recall from Decision Tree.

To evaluate and compare the performance of the classifiers, a ROC Curve was generated, which can be seen in Figures 5, 6 and 7. Comparing the ROC Curves, it is observed that the Logistic Regression performed better in relation to the other

classifiers, since it presented a curve closer to the upper left corner of the graph, indicating that the true positive rate is higher compared to the false positive rate.

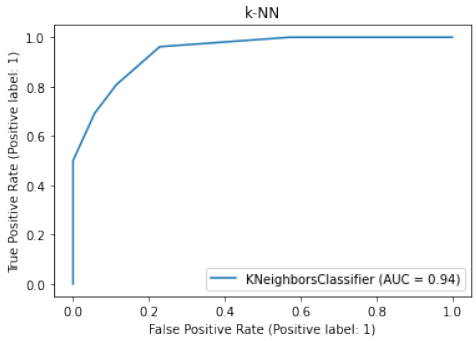

Figure 5 ROC Curve from the K-nearest Neighbor Classifier

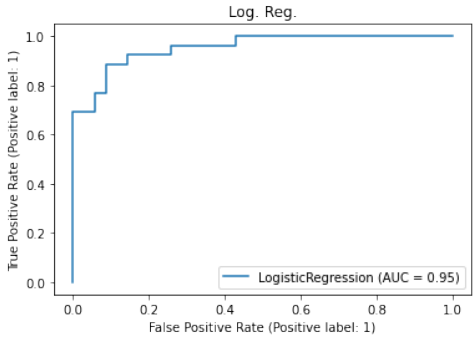

Figure 6 ROC Curve from the Logistic Regression Classifier

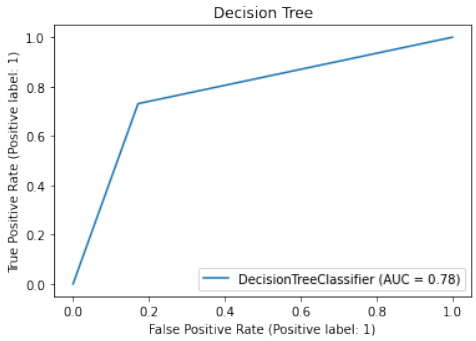

Figure 7 ROC Curve from the Decision Tree Classifier

6 Discussion

After comparing the results of the classifiers in phase 1 and 2 of the project, it was evident to us how the balance of the data is affecting the performance of the prediction models. When training with balanced data, the classifiers could better predict the presence of heart disease. The best prediction model applied to this dataset was the Logistic Regression Classifier, which showed an accuracy of about 87%.
During the exploratory analysis, we could detect some problems with the data, such as missing values and imbalanced data. The first was solved by imputation with the respective mean, a method introduced in the lectures. The problem with imbalanced data was overcome by merging the minority classes into one class, which converted the target group into a binary class. In the implementation of the classifiers, the Decision Tree based prediction model was overtrained. The solution to this problem was not explored in practice due to time constraints, but two different techniques could be implemented: Pre-pruning and Post-pruning. Pre-pruning is stopping the growth of the decision tree earlier, while post-pruning is generating a complete tree and later removing some of the branches. These two techniques would prevent overtraining when implementing a decision tree classifier.
This project provided us with good knowledge of a typical project for a data scientist. In particular, in dealing with data preprocessing and problems such as unbalanced data and missing values, typical issues in data science were encountered. In addition, by participating in this project, we were able to develop practical skills with regard to machine learning and the implementation of classifiers.

7 Appendix

Workload distribution:

Dominik Bannwitz (Bioinformatics) - contributed 16 hours
Florian Herzler (Bioinformatics) - contributed 16 hours
Maximilian Otto (Bioinformatics) - contributed 16 hours
Mariana Steffens (Data Science) - contributed 6 hours

References

1. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, 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**, 304–310 (1989)
2. 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)