University of Konstanz
Political Science & Public Administration
Julia King and Jördis Strack

Statistical Learning
Winter semester 2023/24
February 29, 2024

Classifying Breast Cancer

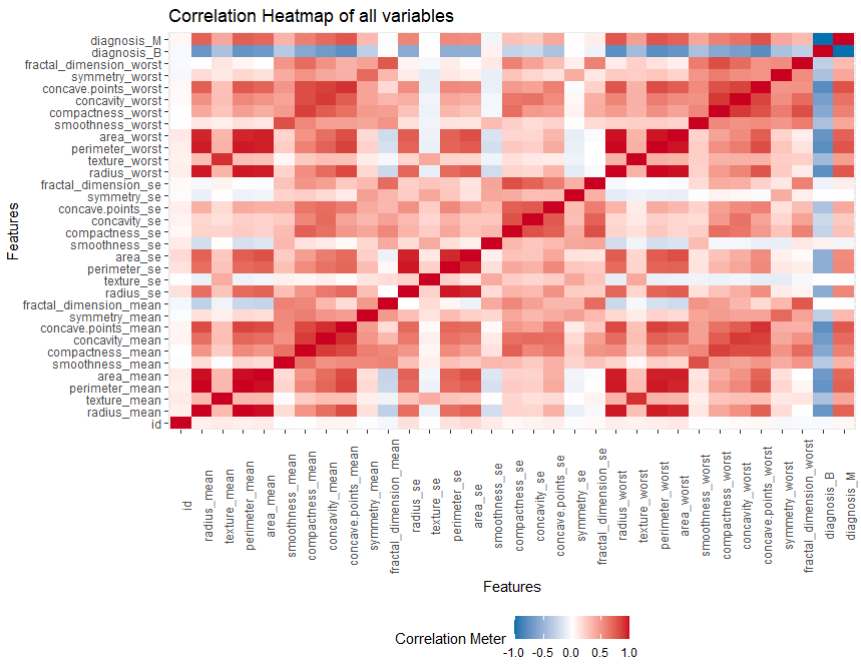
A Comparison of KNN and Random Forests in R

Classifying breast cancer presents a unique challenge, as patients with an incorrect positive test result can sigh with relief once they are found to be perfectly healthy after all, whereas a patient who was misclassified as healthy will lose valuable time to start treatment. Due to these imbalances in severity between a false positive and false negative result, any statistical models used for breast cancer classification must focus on a high recall: The proportion of cases correctly classified as positive, weighted against all true positive and false negative cases.
This project will thus focus on recall, accuracy and precision of both a K-Nearest-Neighbours (KNN) and Random Forest (RF) classifier, to identify the best statistical model to tackle the vital challenge of breast cancer detection. All code, figures and data used can be accessed on GitHub: https://github.com/julia-king-edu/statistical_learning_project

Data Exploration & Preprocessing

The patient data used for classification stem from the Breast Cancer Wisconsin (Diagnostics) Data Set, which can be accessed online via the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic). These data contain 30 features, made up of mean, standard error and largest measure of 10 biomedical patient measurements, as well as an identification and one class variable, denoting whether a patient's cancer is benign (357 cases) or malignant (212 cases) ("Breast Cancer Wisconsin (Diagnostic) Data Set", 2017). There were no missing values in either class or predictor variables and thus no imputation or removal of missing values was required, however, due to both noticeably right-skewed data and preconditions of the KNN classifier, all features were z-standardized.

Figure 1 Heatmap visualizing the three clusters of variables bio-measure, standard error, largest/"worst"



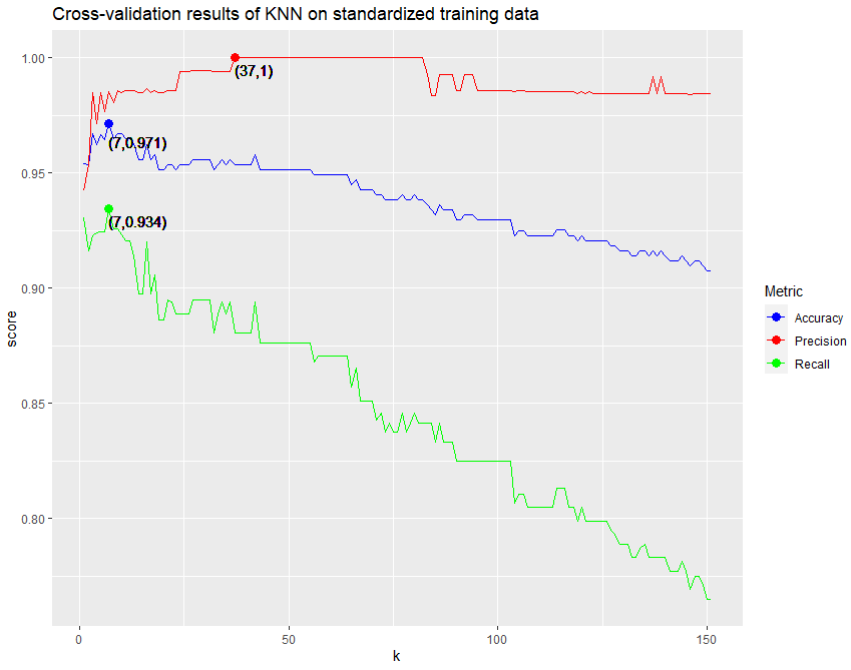
Upon further inspection, many of the features appear to be highly correlated, which is due to the structure of the dataset, see Figure 1. Still, both a Principal Component Analysis (PCA) for dimensionality reduction (for KNN) and a variable importance analysis (for RF) were carried out to confirm both the most influential predictors of breast cancer and partially remedy the high inter-feature correlations. During model comparison we present performance metrics using both the full training set as well as a training set containing the six most critical principal components for cross-validation.

Model Selection & Evaluation

K-Nearest-Neighbors and Random Forests

A sensible comparison of both KNN and RF must consider the specific advantages and drawbacks of each algorithm while maximizing the recall.
KNN models are capable of fitting any function and yield notably good results for non-linear relationships because the algorithm considers data points similar when they are clustered together nearby (James et al., 2021, p. 163). To predict classes, KNN inspects any data point's k closest neighbors directly without a training phase and then classifies it based on the share of neighbours belonging to the categories benign and malignant. The smoothness is adapted by changing k until an optimum is reached and KNN: In the case of this analysis, a value of k set to seven achieved the best recall as depicted in Figure 2.

Figure 2 Performance of KNN under different k

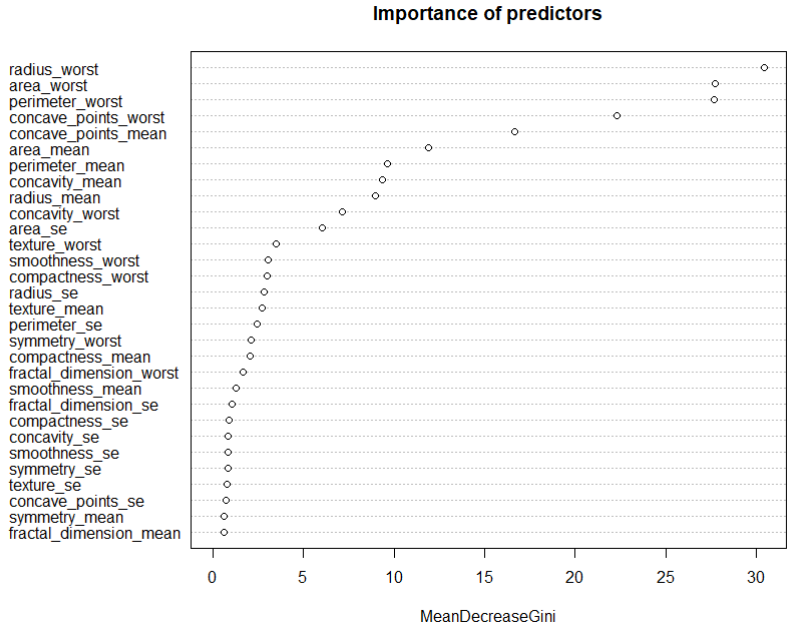


Since KNN does not create a model to be stored, however, actual predictions are computationally expensive and the understanding of the predictors and results generated is often limited, as KNN is instance based (James et al., 2021, pp. 161). Additionally, the results will be highly affected by scaling as the algorithm will compute the distance between data points, making standardization a necessity (James et al., 2021, pp. 182). Finally, while smoothness can be adapted by varying

k, this does go hand in hand with a bias-variance trade-off, causing KNN altogether to often perform worse than other (fine-tuned) classifiers (James et al., 2021, p. 163).

A different approach is that of RF, which builds and compares a forest of decision trees to identify the most suitable features for class prediction automatically without any standardization requirements and is capable of handling missing values. It further allows to perform Variable Importance Analyses to reveal which features hold the highest information or are the purest in predicting a data point's outcome class after handling noise and inter-predictor correlation (James et al., 2021, pp. 343).

Figure 3 Variable Importance after fitting the Random Forest on the regular Cross-Validation data



This makes the final RF model easier to interpret than a general decision tree, as the features most often used for splitting are revealed via the model summary. Lastly, RFs come with many different choices of parameters that can be used to fine-tune the model, which allows predictions under high performance metrics (ibid). Of course, this can be considered a drawback as well, since RFs can be easily overfitted and caution while fine-tuning is endorsed as well as the choice to create smaller trees (James et al., 2021, pp.346). Finally, RFs tend to be biased towards considering categorical variables as more important predictors, which must be considered before analysis (ibid).

KNN and RF Cross-Validation

To compare which algorithm is most suited for the task at hand, a 10-fold Cross-Validation (CV) on two types of training and pseudo-test data was performed: The first predictions were obtained for the general 10 folds of training and CV-test data for both KNN and PCA, while the second run made us of a lower-dimensional feature space obtained through PCA, which preselected the six most important predictors.
In particular, while the consequences of a false positive, including additional costs and treatments, should not be underestimated, they are far outweighed by the potentially fatal scenario

Figure 4 Comparison of performance metrics of KNN and RF classifiers

of undetected malignant cells. Thus, a model predicting breast cancer must rather display a high recall over specificity to detect any person affected by malignant cells, while at the same time yielding accurate results and a minimized misclassification rate. The results reveal, that the RF predicts with the highest recall using the general data and was thus chosen for the final predictions on the separate test data.

Fine-tuning & Final Model Selection

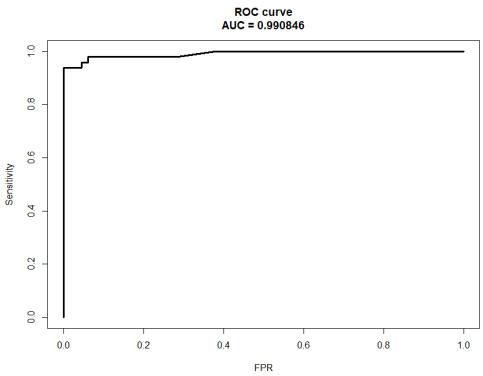
To ensure the best results possible, the RF model was then fine-tuned using Grid-Search and a grid consisting of a random sample of predictors tried at each potential splitting point and the number of trees created. To avoid model complexity and the risk of overfitting given highly correlated predictors, a smaller range of values was chosen with the randomly sampled predictors per split set to 2, 4 and 6, and the number of trees created set to 100, 200 and 300 (James et al., 2021, p.345). After fine-tuning the accuracy of the RF model had decreased slightly, while the recall approached a perfect score of 1. This indicates that the fine-tuned RF model tends to classify every breast cancer with malignant cells correctly, while incorrectly classifying patients with benign cells as having cancer at a slightly higher rate.

Metric	**Performance**
Recall	1.0
Accuracy	0.9649
F1 score	0.9706
Precision	0.9429
Specificity	0.9167

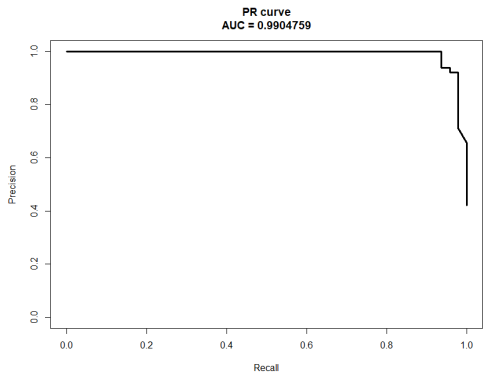
Results & Discussion

To conclude, the ROC curve (AUC = 0.99), as well as the PR curve (AUC = 0.99) of the RF classifier, are presented to indicate the high performance of the model on the breast cancer data. It is not surprising that the RF revealed itself as the most suitable algorithm in comparison to the KNN classifier, as RFs were developed to decorrelate highly correlated predictors ((James et al., 2021), pp. 344), such as bio-measures obtained from patients. While the KNN recall improved under PCA, the recall of the RF model remained undisputed and thus allowed for the most sensitive classification of patients with breast cancer.

Figure 5 Final fine-tuned Random Forest performance evaluation



(a) ROC and AUC of fine-tuned RF classifier



(b) PR curve of fine-tuned RF classifier

References

Breast Cancer Wisconsin (Diagnostic) Data Set. (2017). Retrieved February 28, 2024, from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Home - UCI Machine Learning Repository. (1995). Retrieved February 28, 2024, from https://archive.ics.uci.edu/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1