Applicability of basic machine learning algorithms for breast cancer detection

Maximilian Eineder

Abstract

In this project, a dataset containing geometric characteristics of cell nuclei found in samples of breast mass was examined. Each record of the dataset is classified as benign (sample from healthy tissue) or malignant (sample from cancerous tissue). The research questions I tried to answer using this dataset were:

1. Is it possible to accurately detect breast cancer with basic machine learning algorithms like the ones introduced in the Python for Data Science course with similar data?
2. How do these basic algorithms perform in comparison with more advanced machine learning techniques?
3. Is it possible to determine which of the features are important to the classification, and which ones have little or no contribution?

To answer the first research question, models using four basic machine learning algorithms were created, trained with 2/3 of the records in the dataset. Then, the accuracy of the models was measured using the rest of the data of the dataset. Although the results show a surprisingly high accuracy of ca. 95% for these algorithms, a real application would still require a higher accuracy.

To answer the second research question, existing notebooks on https://www.kaggle.com, in which others have described their results with the same dataset using more advanced techniques like Artificial Neural Networks, were examined and their findings were compared to the findings in this project. This comparison showed that the accuracy of such advanced techniques was not dramatically better (around 98%).

To answer the third research question, I first examined feature distribution plots and tried to predict which features should be the most important ones. Then I compared my predictions with feature importance values calculated by the models used. This showed that it is possible to identify the important features for the classification task.

Motivation

The research of this project was carried out because if even basic machine learning algorithms can produce prediction results that are close to acceptable for a real application, then it should be possible to build models that produce very accurate results. Such models could be the basis for quicker and less expensive cancer screenings, because only a small number of inconclusive results would have to be reviewed by a pathologist. As a result, the risk of developing cancer for the population could be reduced.

Dataset

The dataset used in this project is called *Breast Cancer Wisconsin (Diagnostic)* and was found at https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. It contains 569 records. Each record contains 30 attributes of cell nuclei found in samples from breast tissue and a classification as „malignant“ or „benign“.

Data Preparation and Cleaning

The dataset contains one unnamed column whose purpose was unclear. This column was removed. Apart from this, the dataset contains no null values or other noise, so no other data was removed.

Research Question(s)

In this project there are three research questions:

1. Is it possible to accurately detect breast cancer with basic machine learning algorithms like the ones introduced in the Python for Data Science course with similar data?
2. How do these basic algorithms perform in comparison with more advanced machine learning techniques?
3. Is it possible to determine which of the features are important to the classification, and which ones have little or no contribution?

Methods (I)

To answer the first research question, models using four basic machine learning algorithms were created:

- Decision Tree Classifier
- Random Forest Classifier
- Gaussian Naïve Bayes Classifier
- k-Nearest Neighbors Classifier

The models were trained with 2/3 of the records in the dataset. Then, the accuracy of the models was measured using the rest of the data of the dataset.

Methods (II)

To answer the second research question, existing notebooks on https://www.kaggle.com, in which others have described their results with the same dataset using more advanced techniques were examined and their findings compared to the findings in the project. The examined techniques were

- Artificial Neural Networks
- Gradient Boosting Machine
- Support Vector Machine

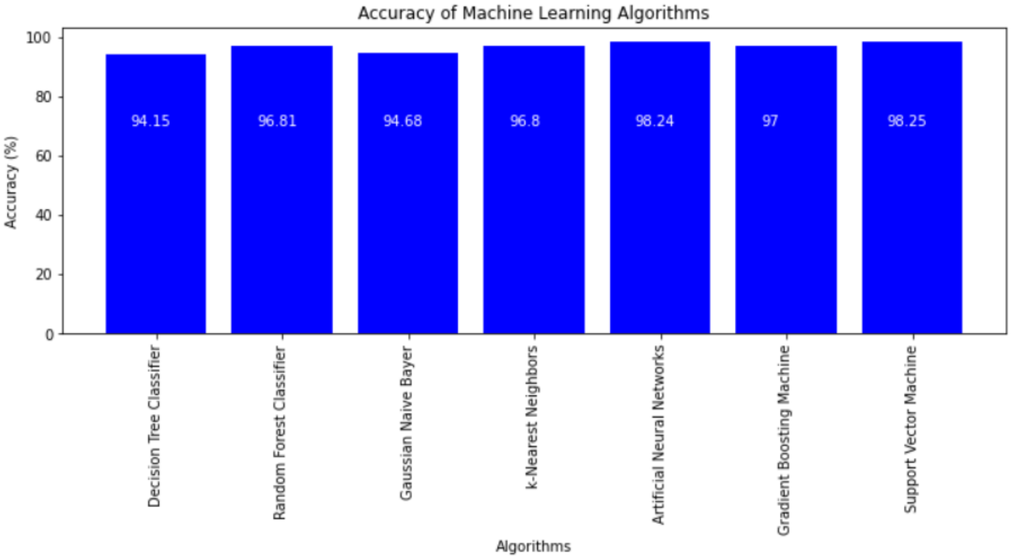
Methods (III)

To answer the third research question, I first examined feature distribution plots and tried to predict which features should be the most important ones. Then I compared my predictions with feature importance values calculated by the models used.

Findings (I)

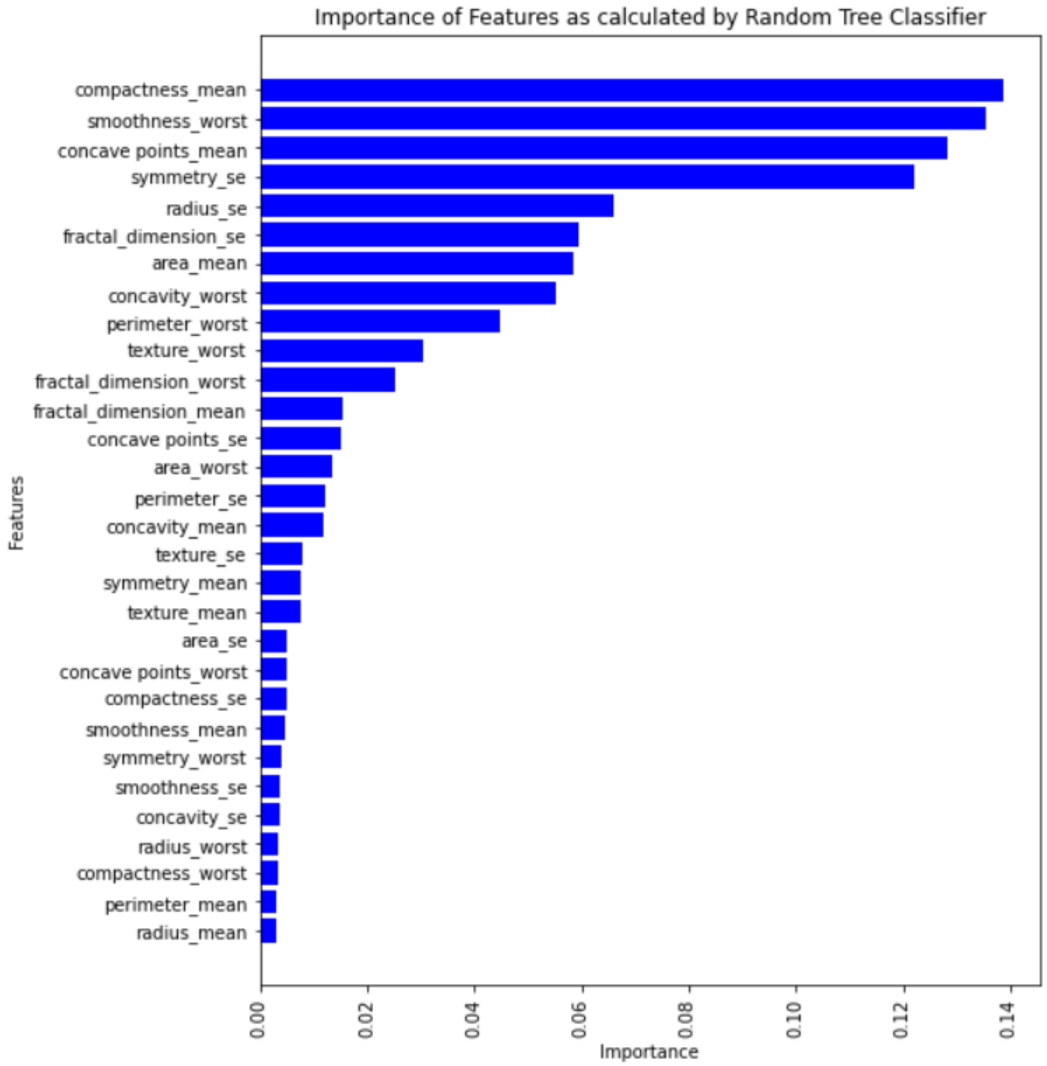
Comparison of basic and advanced machine learning techniques. The results show:

- Even basic algorithms (first four bars) have a remarkable accuracy of ca. 95% (Research question 1)
- More advanced techniques (last three bars) show a higher accuracy, but not dramatically higher (Research question 2)



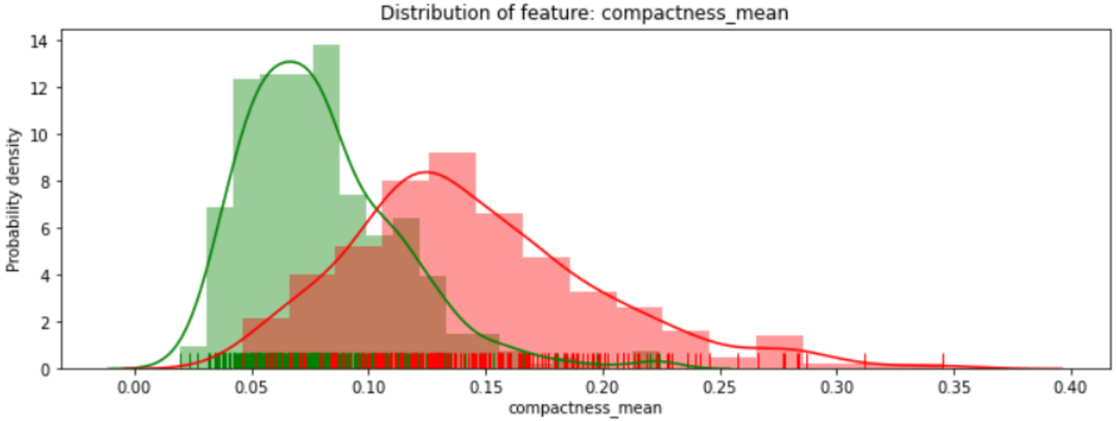
Findings (II)

It is clearly possible to identify features that highly contribute to classification and others that nearly don‘t contribute („feature importance“). As anticipated, the most important features tend to be the ones where the distribution for benign and malignant samples differed most (see next slides).



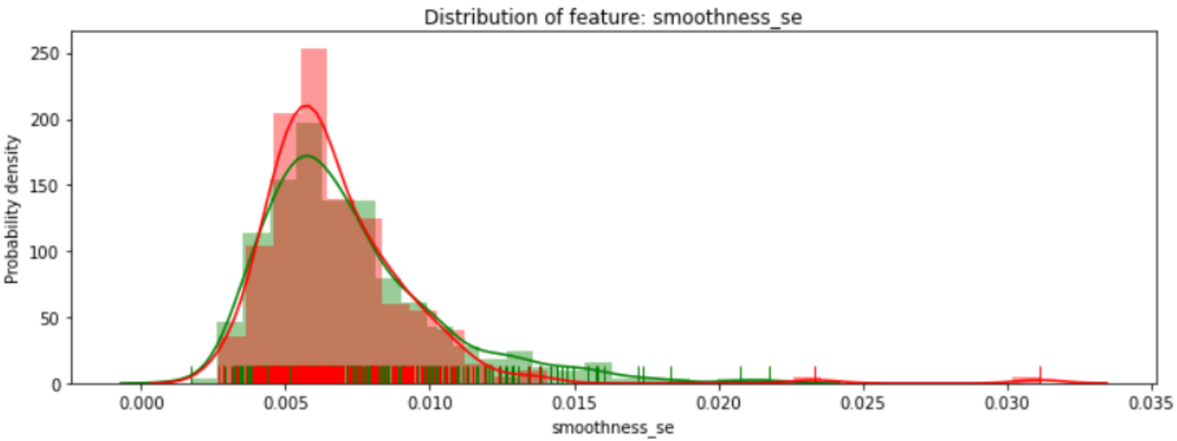
Findings (IV)

Example for a feature with high importance and significant difference in feature distribution (green: benign samples; red: malignant samples)



Findings (V)

Example for a feature with low importance and only little difference in feature distribution (green: benign samples; red: malignant samples)



Limitations

The main limition that should be mentioned is that the findings are based on a very small dataset, which contains only 569 records. To draw defenitive conclusions, a much larger dataset would be required.
In addition, the data of this dataset is of very good quality, because it contains no null values and nearly no implausible outliers. Therefore, the results might lead to more optimistic conclusions than appropriate, because in real applications one would expect data of lower quality.

Conclusions(I)

Research question one:

Is it possible to accurately detect breast cancer with basic machine learning algorithms like the ones introduced in the Python for Data Science course with similar data?

Although the results show a surprisingly high accuracy of ca. 95% for these algorithms, a real application would still require higher accuracy.

Conclusions(II)

Research question two:

How do these basic algorithms perform in comparison with more advanced machine learning techniques?

This comparison shows that the accuracy of such advanced techniques was not dramatically better (around 98% accuracy compared to around 95%).

Conclusions(III)

Research question three:
Is it possible to determine which of the features are important to the classification, and which ones have little or no contribution?
It is possible to clearly identify features that have a high contribution to classification and others that have little to no contribution.

Acknowledgements

I would like to thank https://www.kaggle.com and the authors of the dataset https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Also, I would like to thank the authors of the notebooks who examined the advanced ML techniques which I used for data comparison with my own findings.

References

- Notebook on Artificial Neural Networks (ANN): https://www.kaggle.com/thebrownviking20/intro-to-keras-with-breast-cancer-data-ann
- Notebook on Gradient Boosting Machine (GBM): https://www.kaggle.com/gpreda/breast-cancer-prediction-from-cytopathology-data
- Notebook on Support Vector Machine (SVM): https://www.kaggle.com/faressayah/support-vector-machine-pca-tutorial-for-beginner