CSE4088 Introduction to Machine Learning

Diagnosis Prediction and Classification of Breast Cancer

Team Members:

Yunus Ahmed Stahlschmidt	150119814
Sameeh N O Kunbargi	150119693
Muhammed Fatih Öztel	150119907

Abstract

Breast cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.

This analysis aims to observe which features are most helpful in predicting malignant and benign breast cancer. Furthermore, our in-depth analysis will allow us to observe general trends which may aid us in the model and hyperparameter selection. The goal is to classify whether the cancer is benign or malignant. To achieve this we will use different machine learning classification methods to fit a function to our data that can predict the class of new input.

Overview

We observed which features are most helpful in predicting malignant and benign breast cancer. We also observed general trends which help in model and hyperparameter selection. Finally, we implemented 7 different models and compared them. We split up the required work relatively equally, which was as follows:

- Analysis of the dataset - joint task
- PLA & Logistic Regression - Muhammed Öztel
- SVM with Linear and RBF Kernel - Sameeh Kunbargi
- Decision Tree and Random Forest Classifier - Yunus Stahlschmidt
- Neural Network - joint task

Project Accomplishment

1. Dataset

The machine learning methodology has long been used in medical diagnosis. The Wisconsin Breast Cancer dataset [4], created by Dr. William Wolberg, has been widely used in research experiments. The dataset consists of 569 instances, which were collected over a range of 4 years, where each instance has a total of 32 attributes. Of these 32 attributes, only 30 are relevant for training since every instance has an id and the actual diagnosis results, which we will need in order to evaluate the scores of the different models we will use. The biggest problem we faced with this dataset is that it only has a very small amount of instances, which is one of the main problems of medically related datasets. Of these instances, we used 80% for training and 20% for testing.

2. Attributes Correlation In The Dataset

Down below in Figure 1, you can see a heatmap about the correlation of the attributes of the dataset. As you can see there are a number of attributes that are highly correlated. These were obtained by calculating the mean of the attributes such as the radius or the perimeter of the tissue suspected to be cancerous. These metrics affect the machine learning models quite a bit, which we will show later in detail since they affect the linear separability of the data.

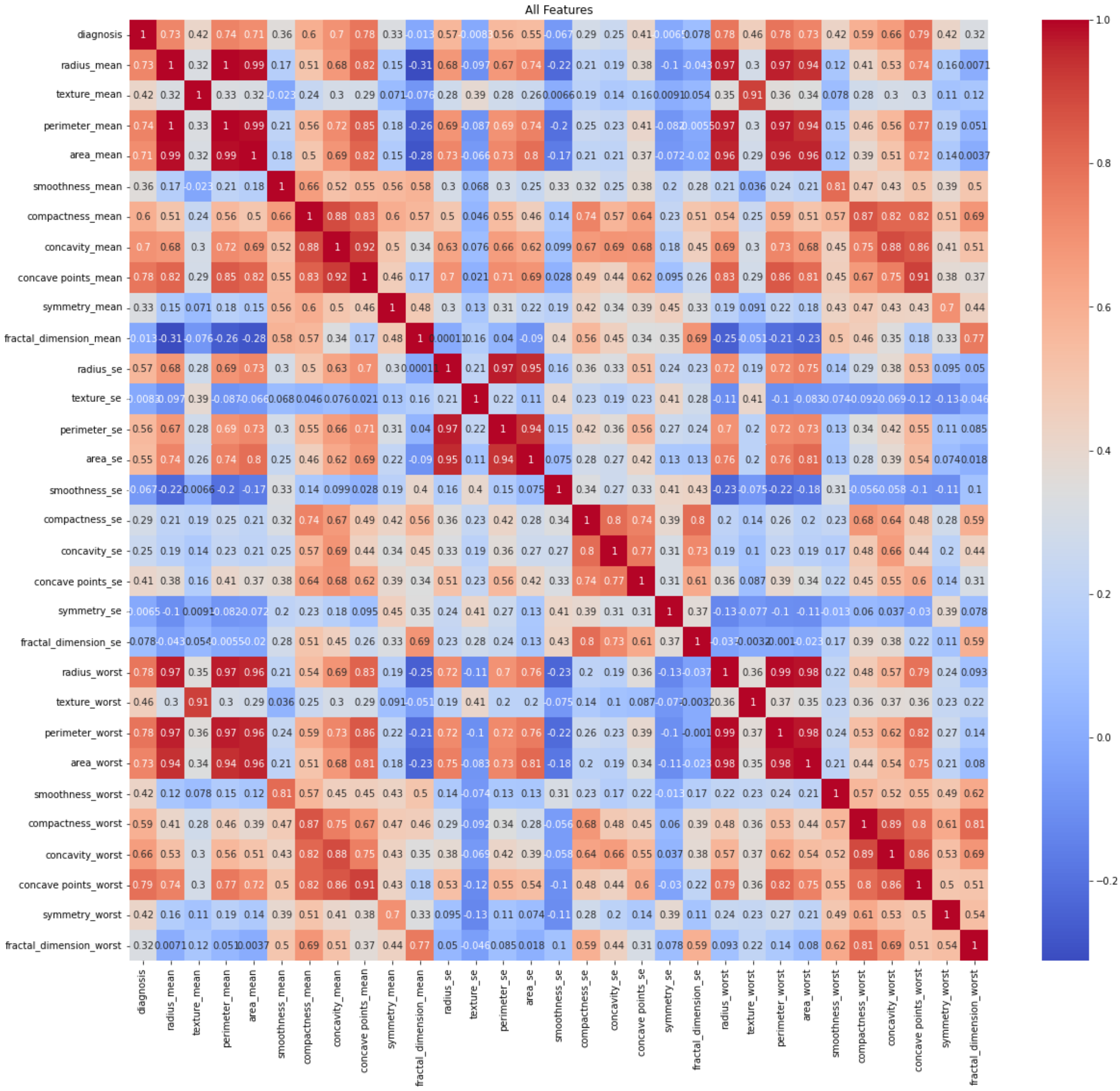

Figure 1: Correlation heatmap of the features in the dataset

As you can tell from this second heatmap in Figure 2, once we remove some of the features from our dataset, there are fewer critical points which makes it easier to apply a non-linear transformation in order to make the data linearly separable. This will help in the training step of some of the machine learning models we used as we'll see later.

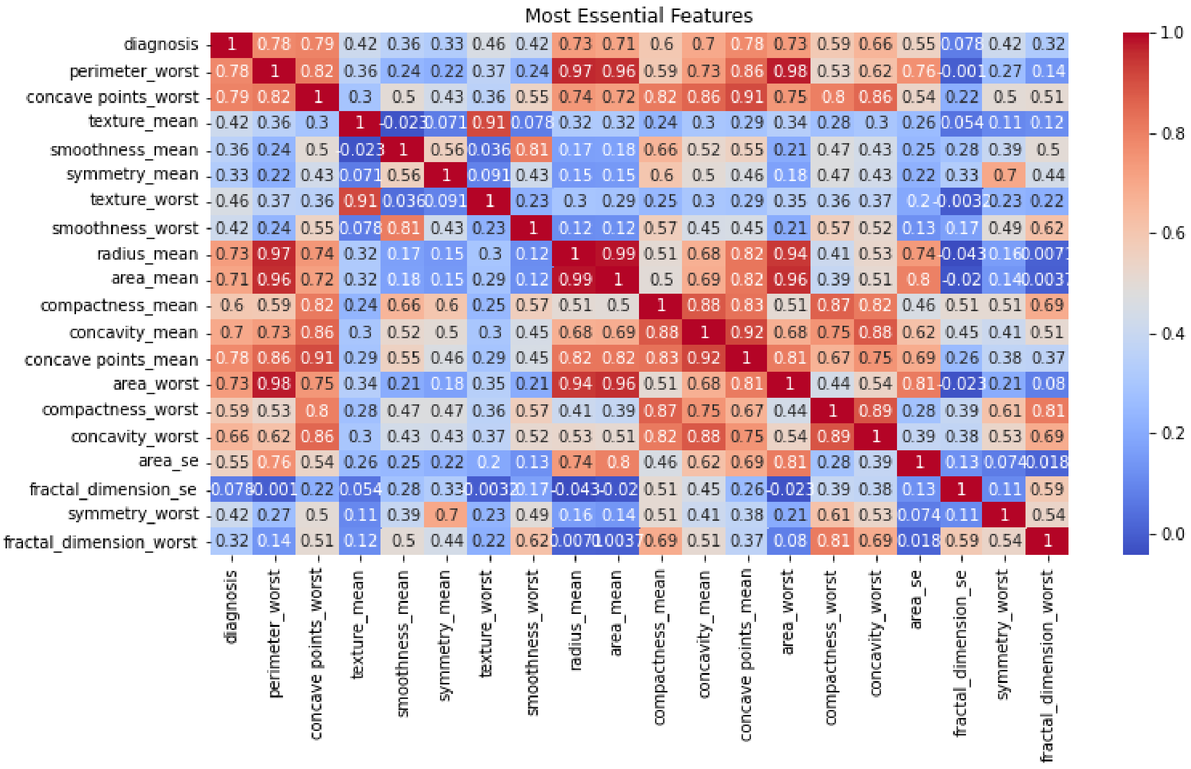

Figure 2: Correlation heatmap of the most essential features

3. Feature Selection Process

We used KDE hist plot to observe the data distribution, in order to get a better understanding of how the data affect the result of the diagnosis.

We used this knowledge to select the most skewed columns which have the biggest effect on the prediction. Some of the features distribution graphs can be seen in Figure 3.

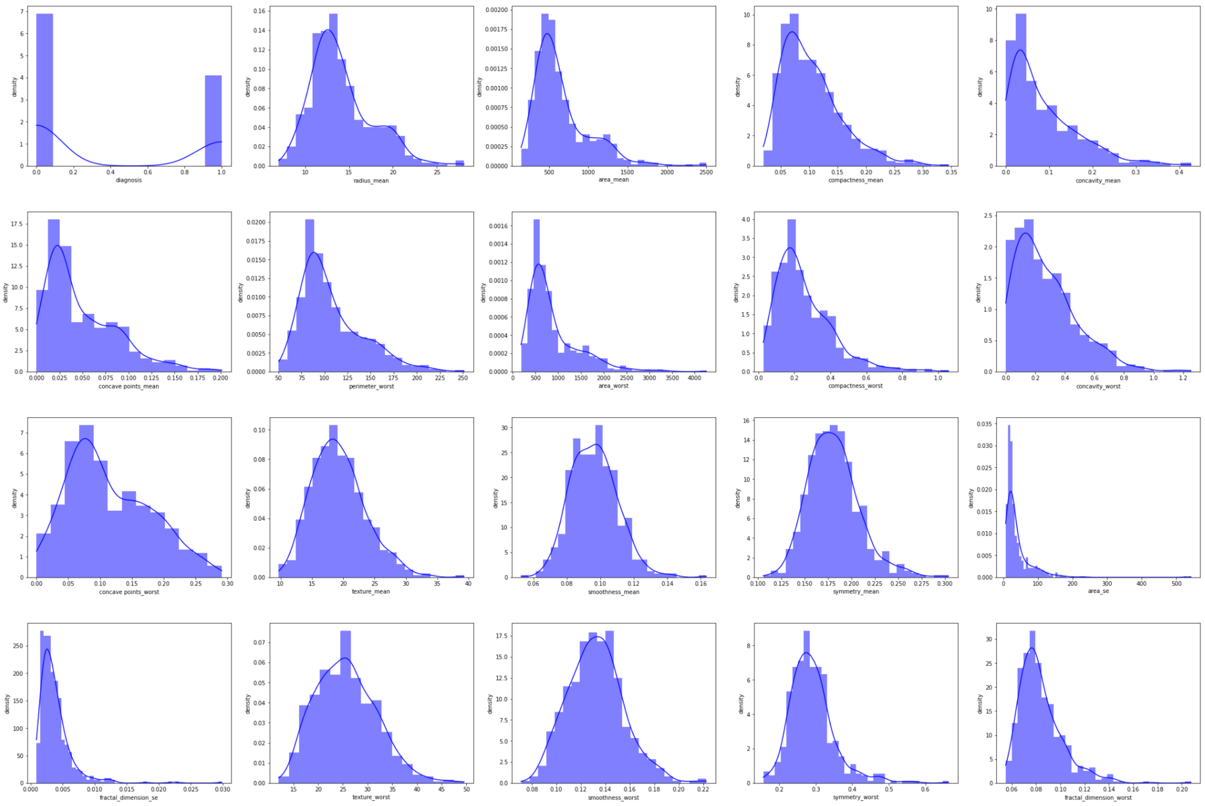

Figure 3: KDE histplot of the features

4. Algorithms

a. PLA & Logistic Regression

The first two algorithms we implemented were PLA and Logistic Regression. As we can see from the Confusion matrices in Figure 4, the PLA algorithm has a relatively good score, as does Logistic Regression.

b. SVM - Linear & RBF Kernel

For Support Vector Machines, we tried both Linear and RBF kernels. From the confusion matrices, we see that they are very close where the RBF kernel is a bit better than the linear kernel.

c. Decision Tree & Random Forest

The final algorithms we used from Scikit-Learn are the Decision Tree and Random Forest classifiers. The Random Forest Classifier is essentially an ensemble of decision trees that run in parallel. Once all of them have produced a result, the algorithm concludes a majority voting, which will produce a more accurate prediction. This is also represented in the confusion matrices in Figure 4.

d. Neural Network

We also implemented a Neural Network with 4 hidden layers and a total of 182K trainable parameters. We optimized our model so we only need to train it for 16 epochs in order to get a really high score. We had dropouts and batch normalization for some of the layers as we can see from the model visualization. We have used the ReLu activation function in the first three layers, with the last layer being a Sigmoid activation function. Also, we did some fine-tuning in order to get better testing accuracy.

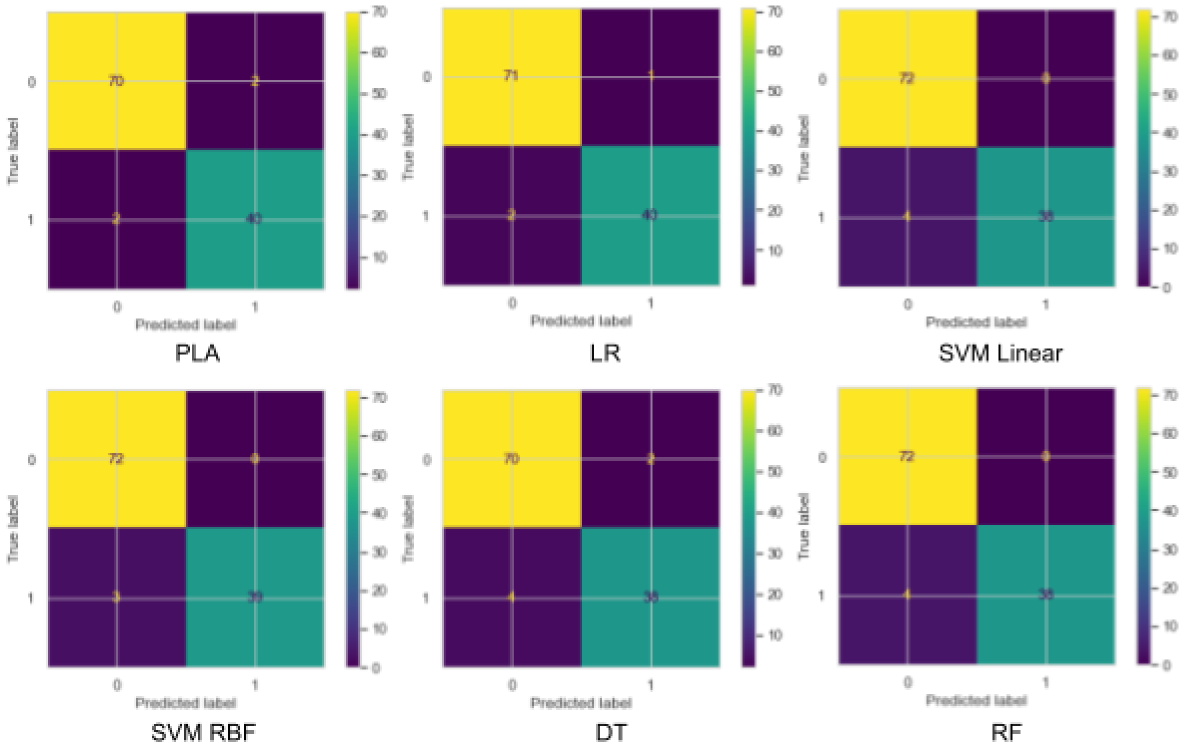

Figure 4: Confusion Matrices for 6 classifier algorithms

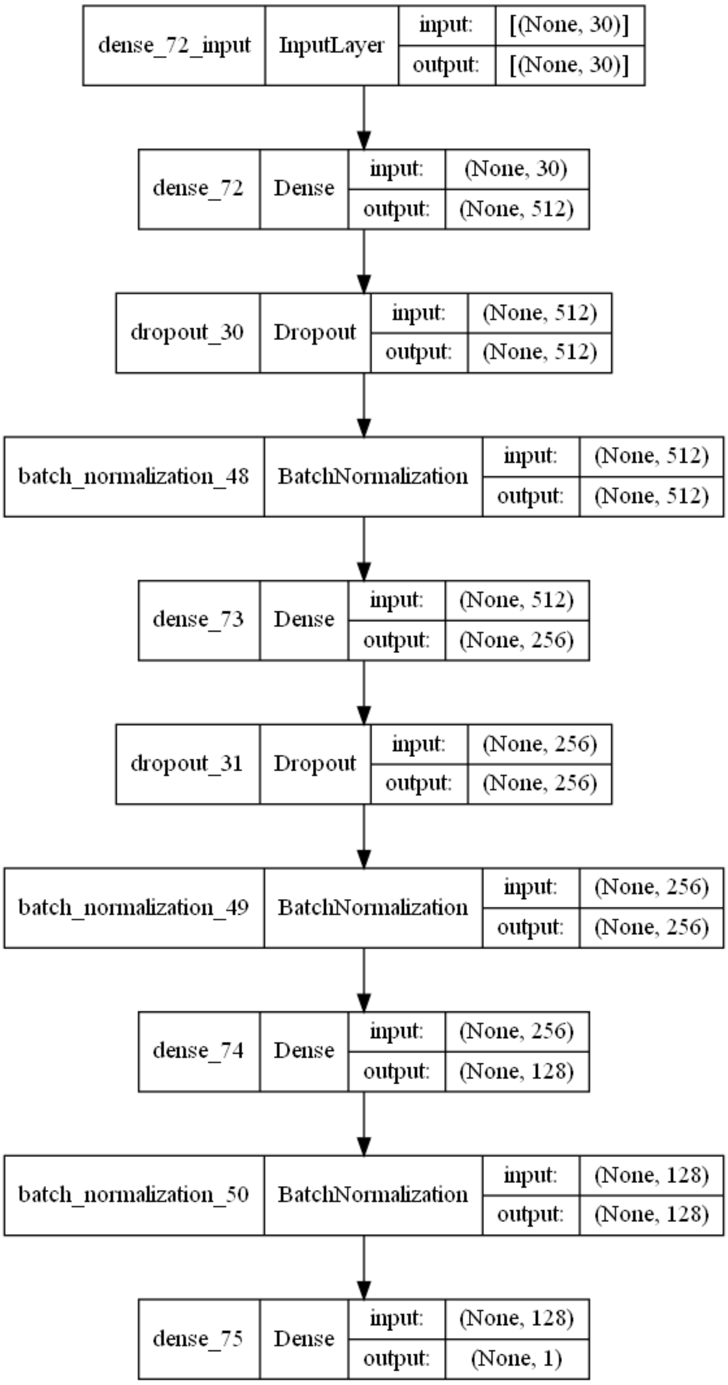
dense_72_input | InputLayer | input: [(None, 30)] | output: [(None, 30)]
dense_72 | Dense | input: (None, 30) | output: (None, 512)
dropout_30 | Dropout | input: (None, 512) | output: (None, 512)
batch_normalization_48 | BatchNormalization | input: (None, 512) | output: (None, 512)
dense_73 | Dense | input: (None, 512) | output: (None, 256)
dropout_31 | Dropout | input: (None, 256) | output: (None, 256)
batch_normalization_49 | BatchNormalization | input: (None, 256) | output: (None, 256)
dense_74 | Dense | input: (None, 256) | output: (None, 128)
batch_normalization_50 | BatchNormalization | input: (None, 128) | output: (None, 128)
dense_75 | Dense | input: (None, 128) | output: (None, 1)

Figure 5: Model visualization of Neural Network

5. Results

Before implementing the neural network, we compared the 6 algorithms we used from Scikit-Learn with different feature extractions. On the left side of Figure 6, we have 20 features that are the most essential features for breast cancer detection, we selected these based on the previously described feature analysis. This gives a relatively good result compared to the full dataset, whose scores we will cover in the next slide. The graph on the right in Figure 6 shows the scores once we took out the 5 highest correlated features, which we derived from the heatmap analysis as previously shown. From these two graphs, we can clearly see the impact of these 5 features on the linear separability of the dataset, the accuracy of the PLA algorithm dropped by over 30 percent while other algorithms showed a smaller but still significant decrease in their scores.

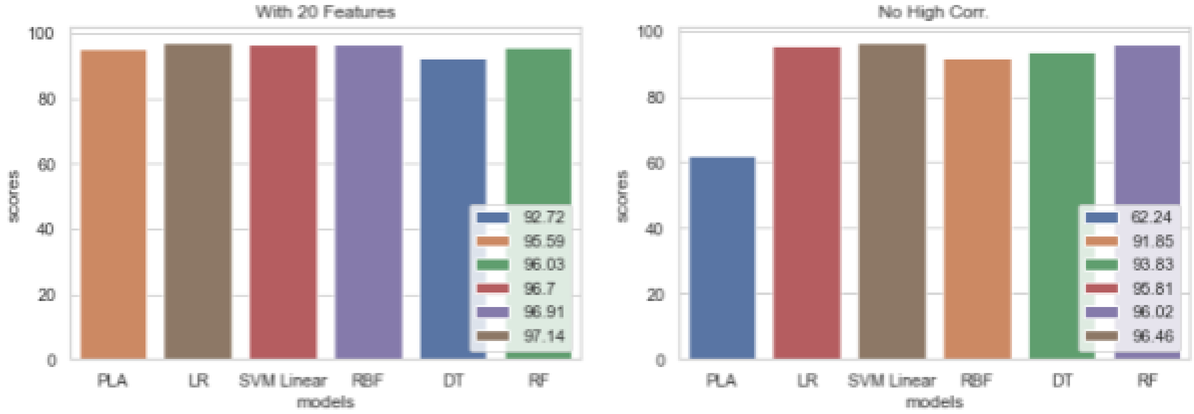

Figure 6: Test results for feature extracted sets

And finally, we got the results for all our models, once for all of the features and once for a revised version of the extracted features with a total of 19 features. In Figure 7, we can see that the Neural network is the best-resulting model for both cases where its f1 score is over 99 percent. We can also see that the 10 features we left out in the analysis of the previous slides make only a small impact for most algorithms, the biggest increase was in the Decision Tree algorithm with an increase of just over 2%. On the graph to the right, we can see the best overall results for all of the models. We achieved this by going over the features once again and discarding one feature which actually impacted the score in a negative way. Here the gain is only very minimal in most cases, but we can still see the positive impact of effective data preprocessing.

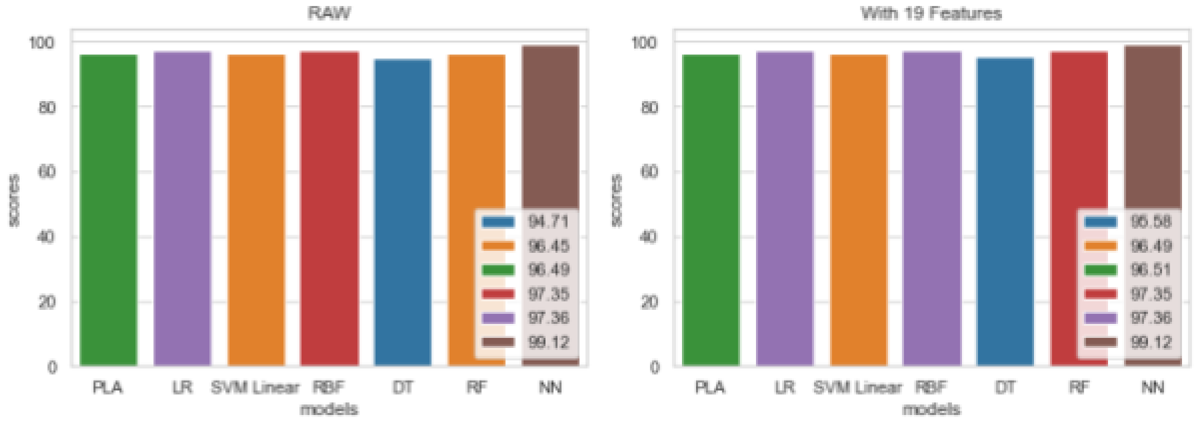

Figure 7: Test result for all classifier models with 2 sets

Summary

To sum up what we did in this project:

- Initial literature research and topic selection
- Research on algorithms used to solve this problem
- Analyze the dataset
- Extracted important & non-important features
- Applied six different Scikit-Learn classifier algorithms
- Developed a Neural Network, trained and tested it

List of references:

[1] Harinishree, M. S., Aditya, C. R., & Sachin, D. N. (2021, April). Detection of Breast Cancer using Machine Learning Algorithms–A Survey. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1598-1601). IEEE.

[2] SAYGILI, A. (2018). Classification and diagnostic prediction of breast cancers via different classifiers. *International Scientific and Vocational Studies Journal*, *2*(2), 48-56.

[3] Telsang, V. A., & Hegde, K. (2020, December). Breast Cancer Prediction Analysis using Machine Learning Algorithms. In *2020 International Conference on Communication, Computing and Industry 4.0 (C2I4)* (pp. 1-5). IEEE.

[4] Dataset
http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29

[5]
https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

[6]
https://towardsdatascience.com/deep-learning-in-winonsin-breast-cancer-diagnosis-6bab13838abd

[7]
https://www.kaggle.com/subhankar007/breast-cancer-accuracy-98-3#RANDOM-FOREST