The Prediction of Cognitive Decline in Patients with Parkinson’s Disease Using a Linear Regression Model and a Random Forest Algorithm

Pisai Daria-Ioana

Grupa 333AA

1. **Introduction**

This essay aims to provide insight into the applicability of machine learning algorithms in one of the most important research fields: medicine.

The scope of this research paper is to analyze and predict the cognitive decline of people affected by Parkinson’s disease.

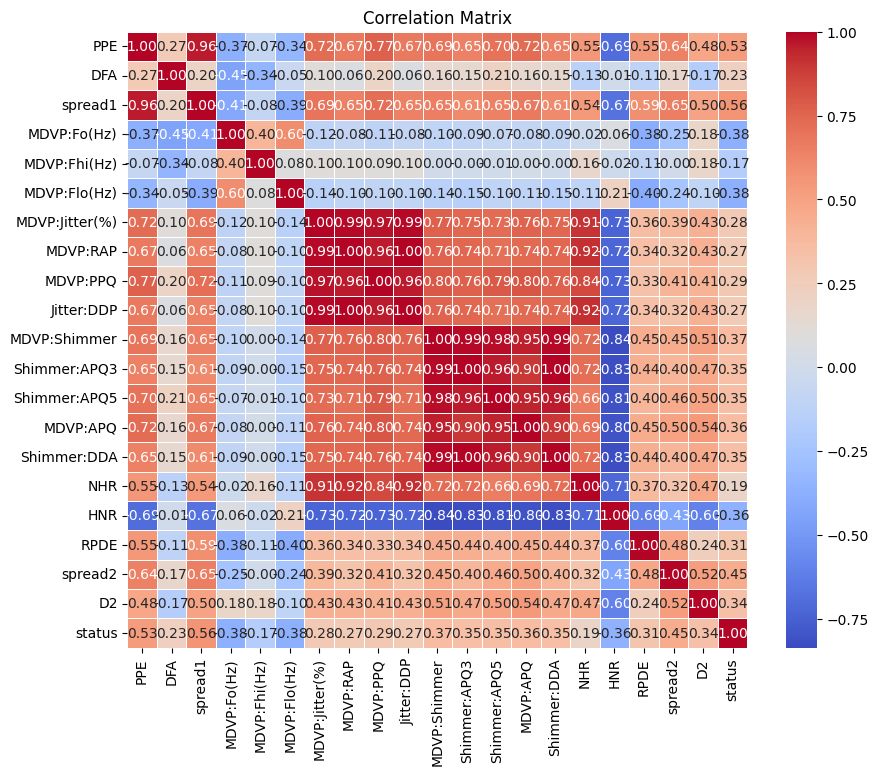
The dataset is composed of 197 patients, and it contains the relevant biomarkers for the study of cognitive decline. The target variable in the model I created is PPE, which stands for pitch period entropy. This variable shows exactly the neural state of the patient, and as the disease advances, the speech becomes impaired and so the PPE indicator grows. Therefore, I studied the other biomarkers’ influence on PPE, as this can lead to an accurate prediction of the level of cognitive decline.

The Linear Regression algorithm was chosen for the analysis of this study because it can model relationships between a target and multiple variables. Moreover, by analyzing the coefficients of the regression model, it is possible to identify which factors have greater influence on PPE and it allows for a clearer understanding of the progression of the Parkinson’s disease.

The Random Forest algorithm is usually utilized for complex relationships and for non-linear data, but it can also provide an easy way to measure the importance of different features in predicting the target. This algorithm has a high prediction accuracy, which makes it perfect for the prediction of medical outcomes, such as cognitive decline. In this study, Random Forest and Linear Regression are used in tandem, to observe which one performs better and which one is a better fit for this specific research case.

1. **Methodology**

Firstly, I started by studying the correlation between each of the biomarkers and PPE, to see which of them could have a larger effect on this variable. I created a correlation matrix and afterwards I collected most of the biomarkers with absolute values close to 1, as those are the relevant variables that should be used for the prediction.

****

To find the biomarkers that are statistically relevant in predicting PPE I also calculated the p-value, which represents an important aspect of hypothesis testing. It’s used to assess whether the target variable is consistent with the null hypothesis (no relationship between variables). Therefore, a small p-value (p ≤ 0.05) would conclude that a variable has a significant effect on the observed variable.

After I calculated the p-value for every biomarker and after taking into consideration the results of the correlation matrix, I concluded that the variables affecting cognitive decline are MDVP:PPQ, MDVP: Jitter(%), MDVP: Shimmer, RPDE, DFA, spread1, spread2 and D2.

For the development of the code, I used multiple libraries, such as Pandas, Scikit-learn, Seaborn and Matplotlib. I utilized Pandas for dataset manipulation, Scikit-learn for splitting data and for training both the linear regression model and the random forest regression model and for the calculation of metrics such as the mean squared error and the R² score. Seaborn and Matplotlib were used to enable the visualization of data.

After deciding on the set of relevant features, these were selected for training the models. Most of the features include audio and physiological measurements. Because the Linear Regression algorithm usually assumes normally distributed data, I used StandardScaler, that would standardize the features that were stored in X.

The data was split into training, validation and testing sets using the function train\_test\_split(), which ensures that 60% of the data will be used for training (117 cases) and 20% of the data for testing (39) and 20% of the data used for validating (39).

Both the Linear Regression algorithm and the Random Forest algorithm were trained using the training set, which contains X\_train (the features) and y\_train (the observed variable).

The models were then applied to calculate the prediction of the target variable (y\_temp) using the validation data (X\_temp). Afterwards, the models were applied on the test set (X\_test), for a more accurate prediction of the final target variable(y\_test).

1. **Results**

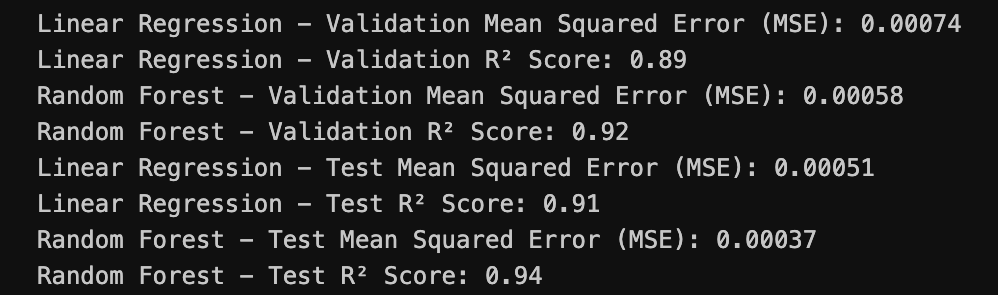
The performance of both models is evaluated using two metrics: Mean Squared Error and the R-squared score. The MSE measures the average of the squared differences between the actual and the predicted values. Ideally, if the models fit the prediction right, MSE should have a low value. The R-squared score is used to calculate the variance in the dependent variable from the independent variables. The models will have a high performance if the R-squared is high.

In the case of Linear Regression, the MSE value is 0.00091 for the validation set, which means that the model predicts with great accuracy the target variable. There is a small difference between the MSE calculated by the Linear Regression algorithm and the one calculated by the Random Forest for the validation case (0.00082), which indicates that the Random Forest Algorithm is slightly more accurate in this context.

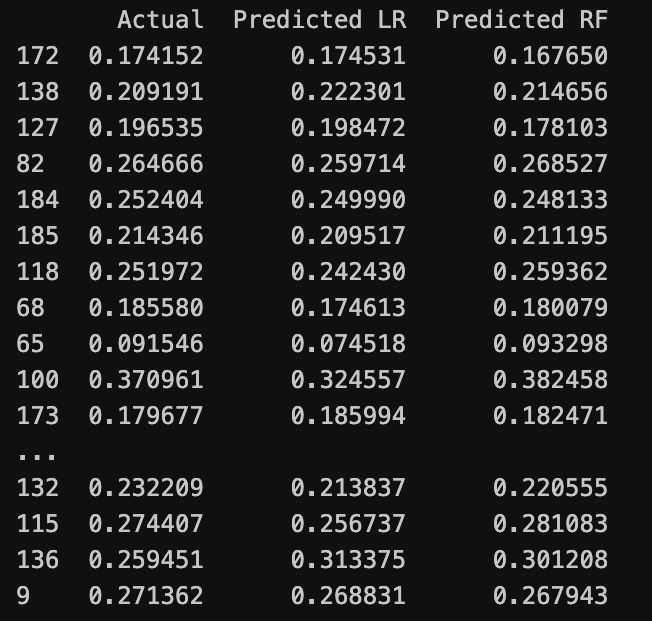
Nevertheless, for the test set the prediction was improved, and so the MSE for Random Forest was 0.00037 and for Linear Regression, it was 0.00051.

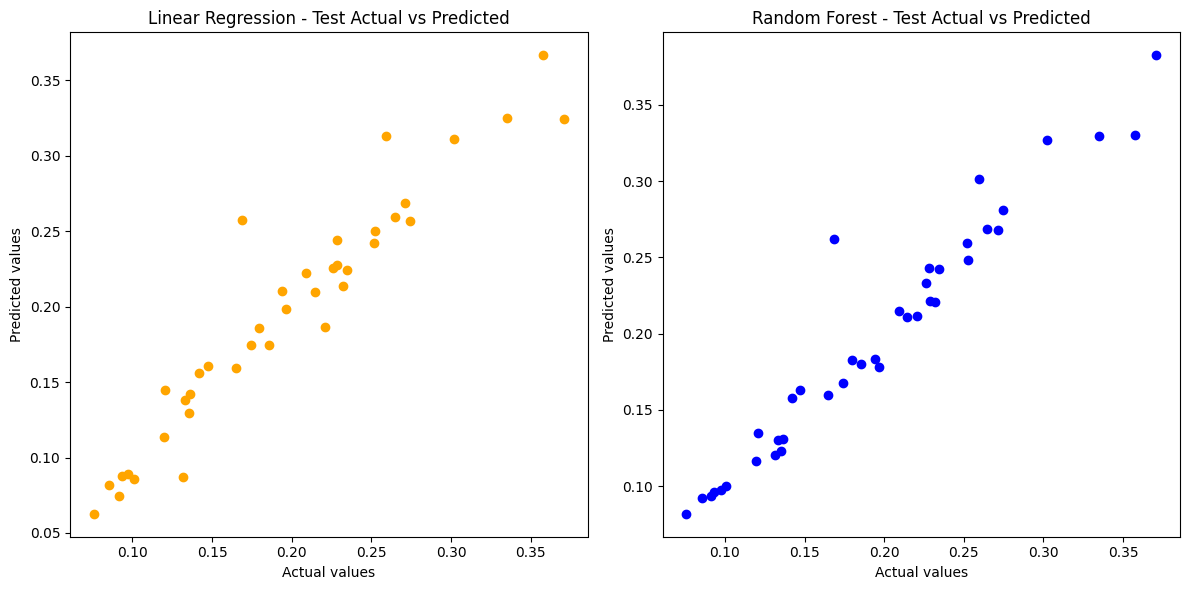
The R-squared score in the case of the validation set shows that both models can be used for predicting cognitive decline, as it is relatively high (0.89 for Linear Regression and 0.92 for Random Forest). For the test set, the results improved and R-squared was 0.91 for the Linear Regression model and 0.94 for the Random Forest Algorithm.

However, it is observable that for both metrics the Random Forest Algorithm marginally outperformed Linear Regression and can therefore be perceived as the more robust model for this research case.

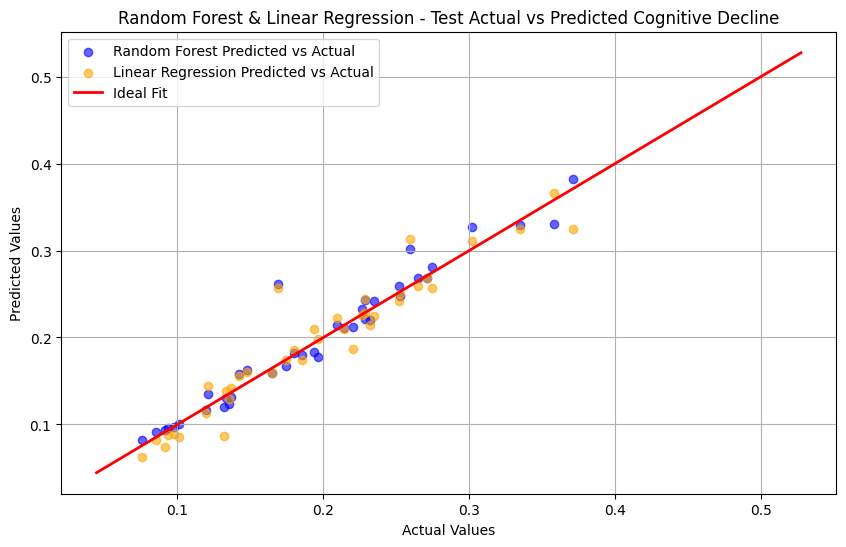


To compare the actual values and the predicted values from both models I created a DataFrame, which helps with the visualization of both models. This way, the data can be numerically compared in an easy manner, by viewing the results from the three columns. The actual data represents the true value of the target variable (PPE), predicted LR contains the predicted values of PPE made by the Linear Regression model and predicted RF contains the predicted values of PPE made by the Random Forest model.



In the end, I plotted three different graphs. The first graph is a scatter plot that shows the values for PPE predicted by the Linear Regression model in comparison with the actual values. The second graph is a scatter plot that shows the values for PPE predicted by the Random Forest model in comparison with the actual values.

The third graph shows the predicted data in comparison with the actual data from both models, while also drawing a red line that represents the ideal case where predicted values match the actual values perfectly.



1. **Conclusions**

Firstly, the study demonstrates the applicability and usability of machine learning algorithms in the field of medicine. The predicting powers of technology could help in various ways and some day, even in treating incurable diseases.

Secondly, the results of the study indicate that the featured biomarkers MDVP:PPQ, MDVP: Jitter(%), MDVP: Shimmer, RPDE, DFA, spread1, spread2 and D2 significantly influence the PPE biomarker, which is the target variable of this study and it can be used as an indicator for cognitive decline in patients affected by Parkinson’s disease. The featured biomarkers are related to physiological measurements, but Parkinson’s disease firstly affects motor and speech functions and afterwards the cognitive ones. Therefore, by accurately predicting the decline in motor functions, we can predict the decline in cognitive functions as well, because the first affects the latter.

Additionally, we used two algorithms to compare their predicting performance and accuracy. In this case, the conclusion was that Random Forest predicted better the target variable, which was observed using R-squared score and MSE. Random Forest can be used for much more complex data, so the model could also be applied to a bigger dataset with a high performance, which makes it more flexible and preferable to the Linear Regression model.

In conclusion, after plotting the graphs it can be observed that both models accurately predict cognitive decline in Parkinson’s disease patients, as the scatter plots are all clustered around the ideal case, with very few outliers.