Estimation of obesity levels based on eating habits and physical condition Data Set

Project Specification

The goal of this project is to estimate the obesity levels of individuals based on various factors such as their age, height, weight and their habits in general, physical habits, eating/drinking habits and smoking habits.

Only 23% of the values in the dataset provided for this project are genuine, it means that 77% of the data was generated and probably needs cleaning.

The purpose of this project is to gain knowledge and experience in handling errors in the data, commonly known as "outliers“, and applying supervised learning algorithms that were covered in our theoretical class.

Cleaning Process (Removing Outliers)

Like we said in the project description, the dataset needed some cleaning and for that we used the Body Mass Index, BMI, to classify our data.
BMI = Weight / (Height * Height)
Then we compared the BMI classification with the one already presented and removed the data wrongly classified. This induced a decrease of 170 rows in our data, 2111 to 1941.

Features used and why

For the algorithms implemented we used different inputs with different features.

- all features (17 features)
- all features with MBI and without Height and Weight (15 features)
- less features with MBI and without Height and Weight (12 features)
- less features without MBI (13 features)
- all features without MBI (16 features)

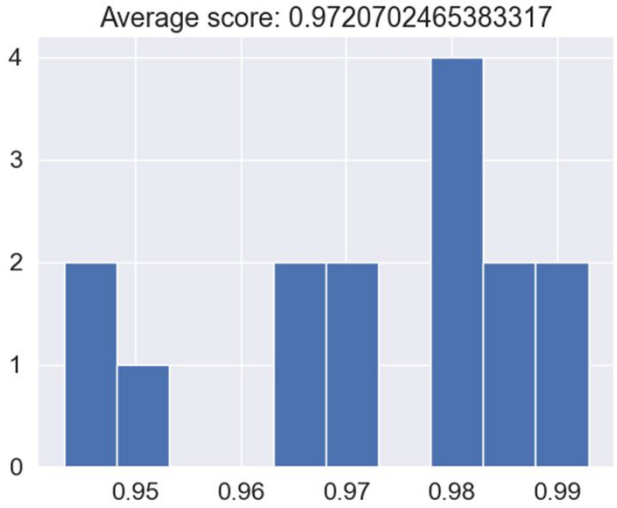
We decided to add the MBI feature and use it in the algorithms. In this cases we could not clean the data as we talked before, because this would result in ***Overfitting***.

Algorithms Implemented and Theirs Best Results

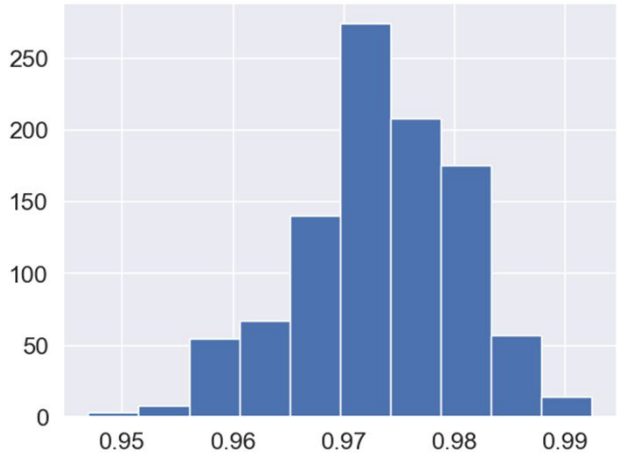
Decision Tree Classifier:

- Best input -> all features (17 features)

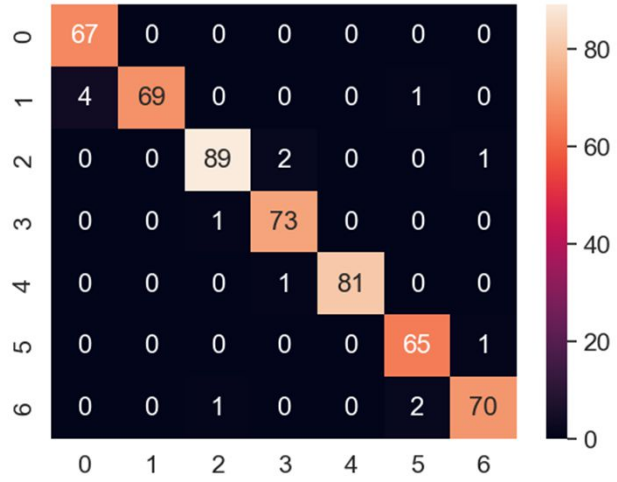
Cross Validation Histogram



1000 iterations Histogram



Confusion Matrix

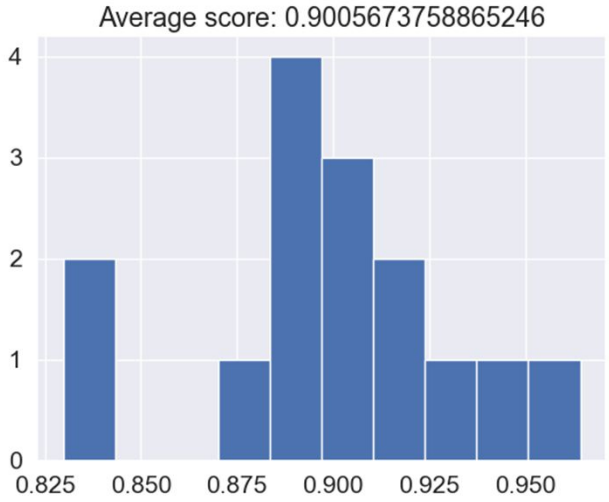


Algorithms Implemented and Theirs Best Results

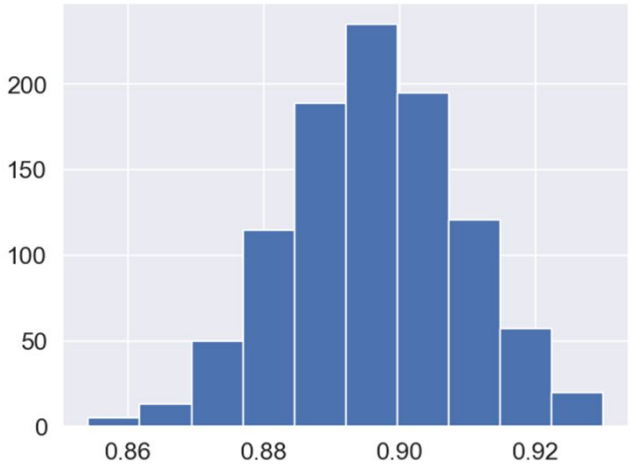
Support Vector Machine (SVM):

- Best input -> less features with MBI and without Height and Weight (12 features)

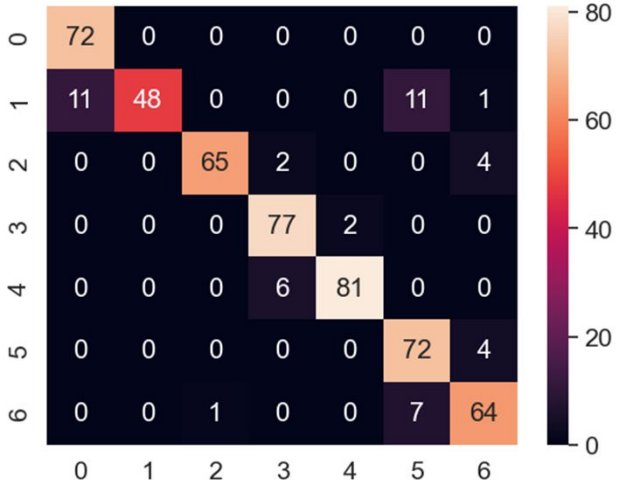
Cross Validation Histogram



1000 iterations Histogram



Confusion Matrix

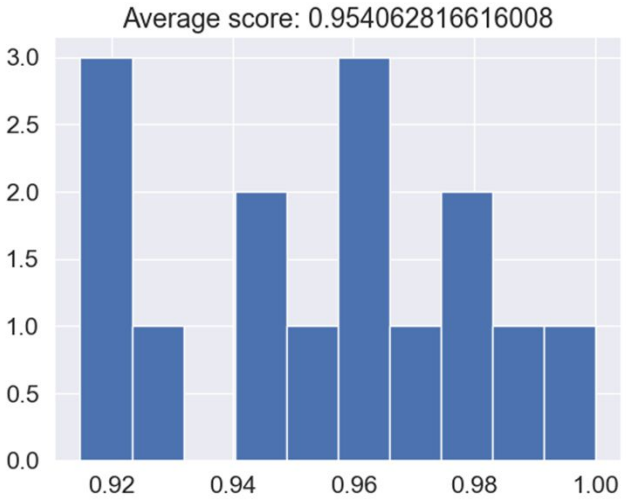


Algorithms Implemented and Theirs Best Results

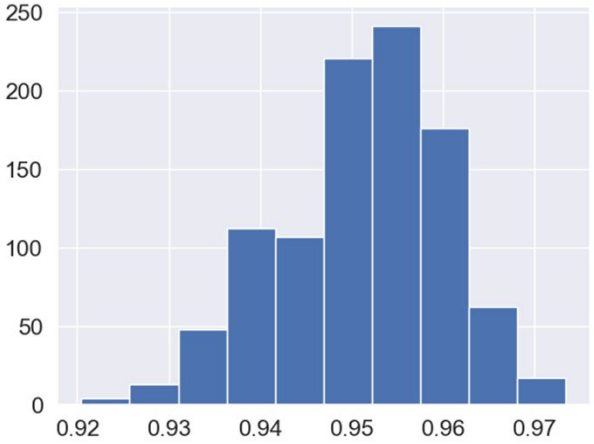
K Neighbors Classifier:

- Best input -> all features (17 features)

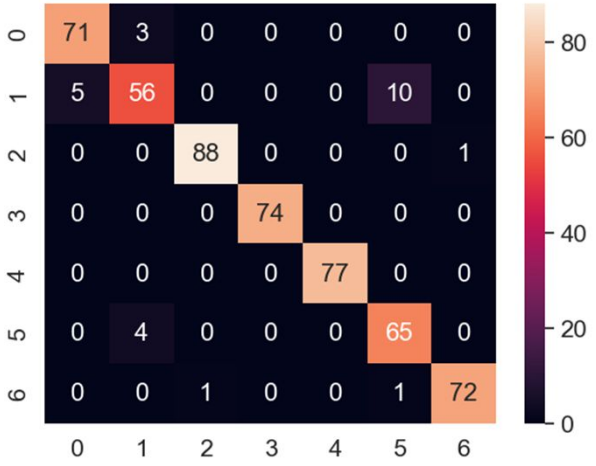
Cross Validation Histogram



1000 iterations Histogram



Confusion Matrix



Algorithms Implemented and Theirs Best Results

Grid Search:

- Decision Tree Classifier with 17 features

```
Best score: 0.9758628841607566
Best parameters: {'max_depth': 40, 'max_features': 17}
```

- Support Vector Machine with 12 features

```
Best score: 0.973498817966903
Best parameters: {'C': 100, 'gamma': 0.01, 'kernel': 'rbf'}
```

- K Neighbors Classifier with 17 features

```
Best score: 0.9668152651131375
Best parameters: {'n_neighbors': 3, 'p': 1, 'weights': 'distance'}
```

Conclusions

In conclusion, the Decision Tree Classifier demonstrated superior performance in comparison to other classifiers for the given dataset. The inclusion of the BMI feature in the input variables led to the best results, even considering the presence of outliers in the dataset. By analyzing the Confusion Matrix, we gained valuable insights into the specific errors made by the classifiers. Furthermore, the Cross Validation analysis revealed that the best performance was achieved with a value of 15.

Overall, the Decision Tree Classifier's ability to handle outliers and leverage the BMI feature contributed to its outstanding performance, making it a favorable choice for classification tasks in this context.