**Supplementary Material**

**Glossary**

**Abbreviations:**

ML: Machine Learning DT: Decision Trees

RF: Random Forest BMI: Body Mass Index

**Terms:**

UCI (University of California Irvine): Is a collection of databases, data generators, and domain theories that researchers, educators, and students use to analyze machine learning algorithms

SMOTE (Synthetic Minority Oversampling Technique): Is a statistical method that increases the number of cases in a dataset to balance the class distribution:

Weka tool (Waikato Environment for Knowledge Analysis Tool): is a collection of machine learning and data analysis free software licensed under the GNU General Public License.

**Intermediate Results and Graphs**

|  |  |  |
| --- | --- | --- |
| BMI Ranges for Each Category: | Min\_BMI | Max\_BMI |
| Insufficient\_Weight | 12.999 | 19.082 |
| Normal\_Weight | 18.491 | 24.913 |
| Overweight\_Level\_I | 22.827 | 28.770 |
| Overweight\_Level\_II | 25.714 | 30.363 |
| Obesity\_Type\_I | 29.912 | 35.171 |
| Obesity\_Type\_II | 34.049 | 39.787 |
| Obesity\_Type\_III | 36.774 | 50.812 |

**BMI Column:**

First, we derived a new BMI column in attempt to make it our new target column, giving much more depth and complexity to our model, while making it more suitable for Logistic Regression.

Then we associated the target columns to the corresponding bmi ranges as shown in the table below.

As we can see the bmi ranges for each category overlap with each other making it a challenge to assign the overlapping values to a certain category.

Initial thought was to use GMM (Gaussian Mixture Model) along with k-means for initializing the clustering to give a probability percentage for each case. For instance, a person with bmi value of 18.6 that falls between 2 categories would have results like Insufficient\_Weight: 0.3 and Normal\_Weight: 0.7, assigning the case to Normal\_Weight in the end.

Below is a graph showing the distribution of bmi for each category, giving insights about the degree of overlapping as well.

A graph of different colored columns

Description automatically generated

With exception of the final two categories the overlapping is very minor and can be dealt easily with the use of the proposed clustering methods.

Since preprocessing is not the main focus of this coursework and it would be very time consuming changing and mapping the target variable, we decide to ditch this approach and mark it as future work. So using Decision Trees instead of Logistic Regression would be more reasonable choice, considering the balanced and categorical target variable.

**Target Distribution:**

Below we see alternative target distribution graph, showcasing the balance in classification classes.

We decided to ditch this over the pie chart mainly because it was very basic and not so aesthetic for the poster.

A graph of a bar chart

Description automatically generated with medium confidence

**Grid Search/Hyperparameter Tuning:**

|  |  |
| --- | --- |
| **Intermediate results from grid search:** | **Best values:** |
| **RF:** |  |
| **1st run:** |  |
| nTrees = [50, 100, 250, 500] | nTrees=100 |
| minLeafSizes = [1, 3, 5, 8, 10] | minLeafSizes=1 |
| maxSplits = [25, 50, 100, 150, 200] | maxSplits=200 |
| **2nd run:** |  |
| nTrees = [100, 200, 300, 400, 500] | nTrees= 300 |
| minLeafSizes = [1, 2, 3] | minLeafSizes=1 |
| maxSplits = [50, 100, 200, 300, 400] | maxSplits=400 |
| **3nd run:** |  |
| nTrees = [100, 200, 300, 400, 500] | nTrees= 300 |
| minLeafSizes = [1, 2, 3] | minLeafSizes=1 |
| maxSplits = [200, 300, 400,500] | maxSplits=400 |
| **DT:** |  |
| **1st run:** |  |
| minLeafSizes = [1, 3, 5, 8, 10] | minLeafSizes=3 |
| maxSplits = [25, 50, 100, 150, 200] | maxSplits=100 |

We started with a wide range of values for both methods in search of finding the best values in between the ranges not on the edges as this would imply underfitting or overfitting.

RF took us 3 runs to finally find the best parameters inside the ranges and not the edges, while with DT from the first run we were able to specify the range we wanted to apply a more detailed grid search.

**Metrics Representation:**

We also had this graph for visualizing and comparing the evaluation metrics of each model. This graph was not on the poster because there was simply no space left and while we can write the metrics in a line or two, same thing could not happen with confusion matrices and roc curves.

A screenshot of a graph

Description automatically generated