Obesity Levels

Saketh Teki, Sai Deekshith, Revanth Kumar, Tarun Kumar Bandaru, Sameeksha Chiguru

December 12, 2022

Abstract

Obesity is strongly associated with multiple risk factors. It is significantly contributing to an increased risk of chronic disease morbidity and mortality worldwide. There are various challenges to better understanding the association between risk factors and the occurrence of obesity. The traditional regression approach limits analysis to a small number of predictors and imposes assumptions of independence and linearity. Machine Learning (ML) methods are an alternative that provides information with a unique approach to the application stage of data analysis on obesity. This study aims to assess the ability of ML methods, namely Support Vector Machine, Logistic Regression, K nearest neighbors, Decision trees, and Random forest. Identifying these risk factors could inform health authorities in designing or modifying existing policies for better controlling chronic diseases, especially in relation to risk factors associated with obesity. Moreover, applying ML methods to publicly available health data is a promising strategy to fill the gap for a more robust understanding of the associations of multiple risk factors in predicting health outcomes.

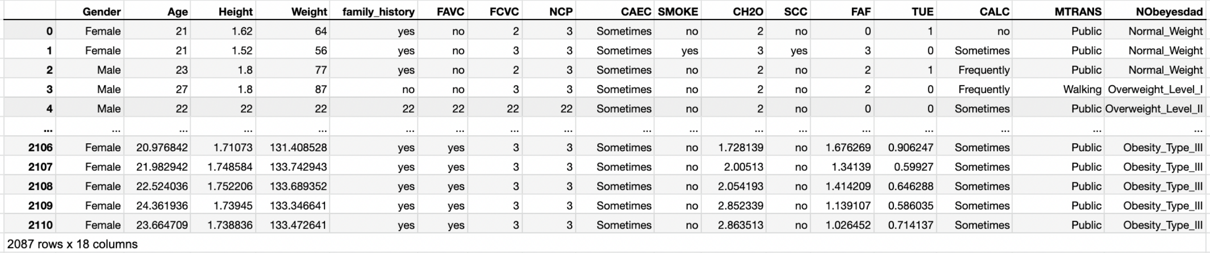
Keywords: obesity; chronic diseases;

1 Introduction

The purpose of this data is for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. This can be used to build an estimation of obesity levels based on the nutritional behavior of several regions.

2 Dataset

The dataset contains 17 attributes and 2111 records, the records are labeled with the class variable NObesity (Obesity Level), which allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.



	Gender	Age	Height	Weight	family_history	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	Female	21	1.62	64	yes	no	2	3	Sometimes	no	2	no	0	1	no	Public	Normal_Weight
1	Female	21	1.52	56	yes	no	3	3	Sometimes	yes	3	yes	3	0	Sometimes	Public	Normal_Weight
2	Male	23	1.8	77	yes	no	2	3	Sometimes	no	2	no	2	1	Frequently	Public	Normal_Weight
3	Male	27	1.8	87	no	no	3	3	Sometimes	no	2	no	2	0	Frequently	Walking	Overweight_Level_I
4	Male	22	22	22	22	22	22	22	Sometimes	no	2	no	0	0	Sometimes	Public	Overweight_Level_II
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2106	Female	20.976842	1.71073	131.408528	yes	yes	3	3	Sometimes	no	1.728139	no	1.676269	0.906247	Sometimes	Public	Obesity_Type_III
2107	Female	21.982942	1.748584	133.742943	yes	yes	3	3	Sometimes	no	2.00513	no	1.34139	0.59927	Sometimes	Public	Obesity_Type_III
2108	Female	22.524036	1.752206	133.689352	yes	yes	3	3	Sometimes	no	2.054193	no	1.414209	0.646288	Sometimes	Public	Obesity_Type_III
2109	Female	24.361936	1.73945	133.346641	yes	yes	3	3	Sometimes	no	2.852339	no	1.139107	0.586035	Sometimes	Public	Obesity_Type_III
2110	Female	23.664709	1.738836	133.472641	yes	yes	3	3	Sometimes	no	2.863513	no	1.026452	0.714137	Sometimes	Public	Obesity_Type_III

2087 rows x 18 columns

3 Exploratory Data Analysis (EDA)

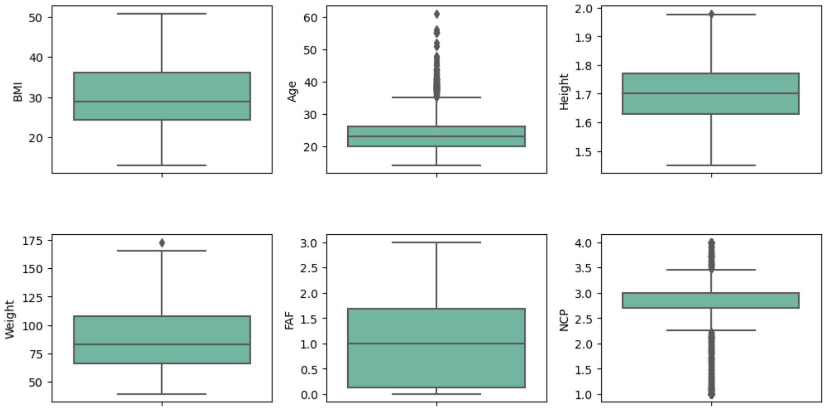
There are no missing values in the dataset. An additional feature, BMI is introduced which is created using height and weight.

3.1 Numerical Features

Numerical data are values that can be measured and organized logically. Their characteristics are numbers that describe an object's various properties.

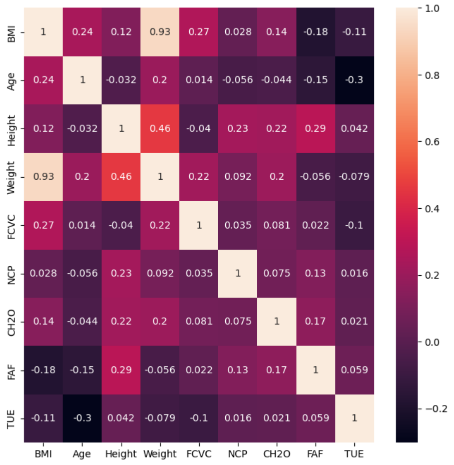
1. **Outliers and Box Plots** An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection when collecting the data.Age and NCP are observed to have

outliers.The below distribution suggests that the numerical features are at different scales. Hence, while building any model, standardizing the 2-numerical features is necessary.



Box plots of numerical features

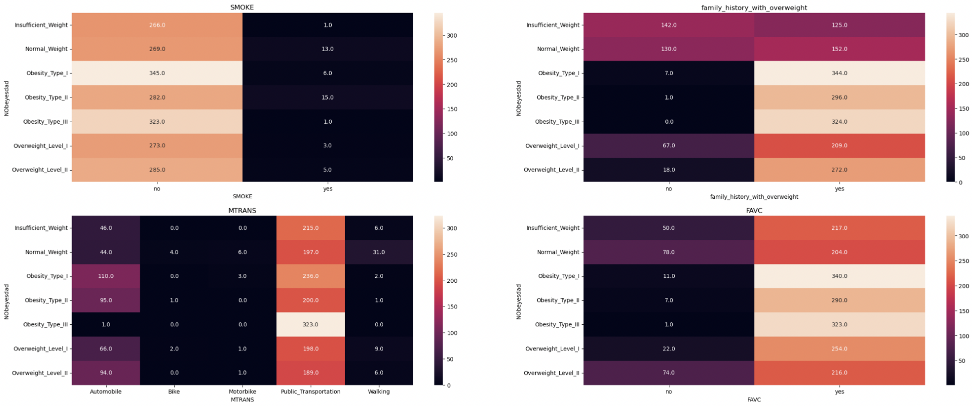
2. **Correlation** Correlation explains how one or more variables are related to each other. Pearson correlation is performed within the numerical features to check any multi-collinearity. If we observe any pair of features with high correlation, we can remove one of them from model training as it reduces the complexity of the model but retains most of the information in the dataset. As weight and BMI are highly correlated from the below chart, we can remove weight from the model-building process.



Correlation heat map of Numerical features

3.2 Categorical features

As the output is categorical, contingency tables for each of the categorical features against the output variable are created. Each table signifies the distribution of data points in the dataset in the output classes, against each category in the input feature.



Contingency table of categorical features

4 Methods (Classification)

4.1 SVM

The data is classified using separate hyperplanes using a Support Vector Machine (SVM). In SVM, data overfitting is reduced. An ideal hyperplane is created as an output of the supervised training process, which categorizes previously unknown fresh samples. The data can be separated using linear algebra transformations. If more complexity is desired, kernels can be used. SVM has a large margin intuition, which implies the best hyperplane will be as far away from the data points is possible. Even if a new example is closer to the incorrect class, it will remain on the right side of the hyperplane. SVM was implemented using SVC (Support Vector Classifier) from sklearn.svm. Data Preprocessing (Scaling) was helpful in observing fruitful results.

4.2 K-Nearest Neighbors(KNN)

K-Nearest Neighbors is a non-parametric supervised learning method, it is easy to learn and implement and it divides data into groups or classes, these groups or classes are created based on the structure of data. KNN assumes 'k' data points with similar characteristics exist closely and new data is integrated into the class that most closely resembles it.'k' is also called as number of neighbors which plays a major role. Distance is measured between data points to measure the similarities. For distance calculation we use functions like:

$$EculideanDistance = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$ManhattanDistance = \sum_{i=1}^{n} |x_i - y_i|$$

Based on k value it votes the similarities and defines the class. We used KNeighborsClassifier from sklearn.neighbors to create KNN.

4.3 Logistic Regression

The logistic Regression is one of the supervised classification models in Machine Learning developed with a probabilistic approach. Logistic Regression is extension of Linear Regression. Instead of straight line in Linear Regression, a s-shaped logistic or sigmoid curve is drawn to fit the data. Logistic regression is mainly used when the target is categorical variable. The target is predicted using the below equation:

$$y(x) = Z(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...)$$

Where,

$$x_1, x_2... = attributesofdataset$$

$$\beta_1, \beta_2... \ = weightsofattributes$$

$$\beta_0 = biasparameter$$

$$Z(a) = \frac{1}{1 + e^{-a}}$$

4.4 Decision Trees

Decision Trees are a non-parametric supervised learning method used for both classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. The input vector is passed through the tree while making a prediction, and an output class is obtained. Here, we used sklearn's DecisionTreeClassifier to implement a decision tree.

4.4.1 Variable selection criterion

where the true complexity and sophistication of decision lies. Variables are selected on a complex statistical criterion which is applied at each decision node. Now, variable selection criterion in Decision Trees can be done via two approaches:

1. **Entropy** is a measurable physical property that is most commonly associated with a state of disorder, randomness, or uncertainty.

$$Entropy = \sum_{i=1}^{n} -p(Ci).log(p(Ci))$$

2. **Gini Index** measures the probability of a random instance being misclassified when chosen randomly. The lower the Gini Index, the better the lower the likelihood of misclassification.

$$Gini = 1 - \sum_{i=1}^{n} p^2(Ci)$$

where p(ci) is the probability/percentage of class (Ci) in a node.

4.5 Random Forest Classifier

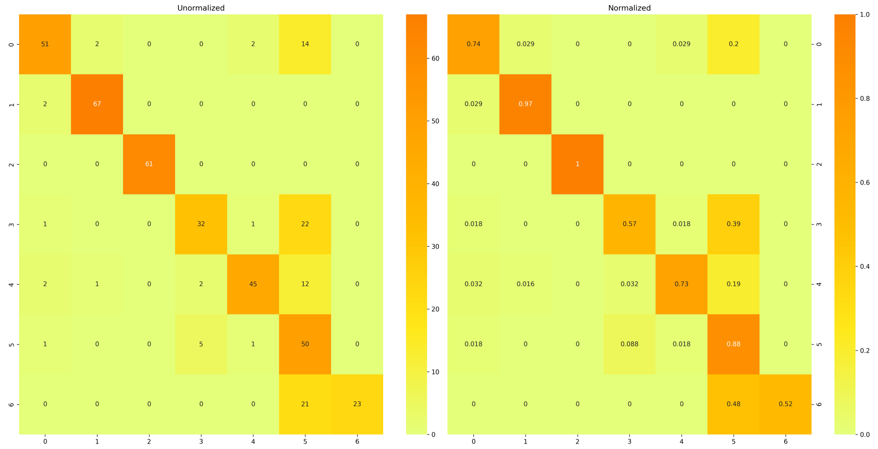
Random Forest is an ensemble algorithm in which the output is based on a collection of decision trees. Each decision is trained on a subset of data and randomly selected features available from the dataset. Under the classification setting, each decision tree outputs a class. The class which is observed to the output of majority of the decision trees is picked as the output of the random forest model. Since the output is not based on the learning of a single tree, the model is less prone to overfitting and bias.

5 Results

5.1 SVM

5.1.1 Confusion Matrix For SVM with Unscaled Data

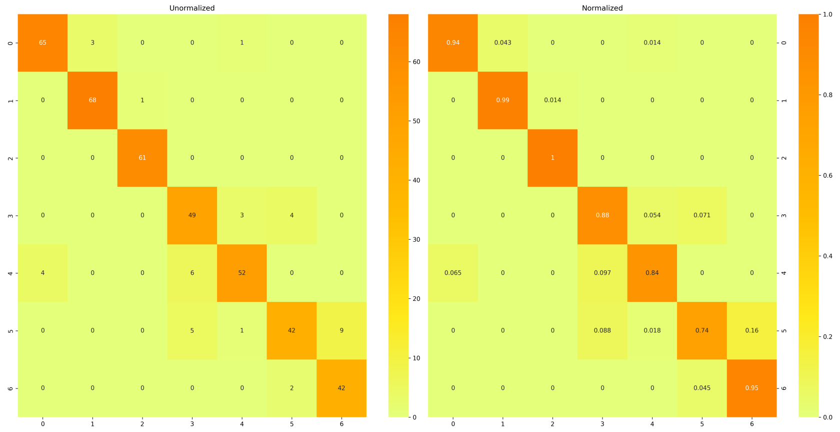
Initially, it was difficult to train an SVM model using Linear kernel. So, a Polynomial kernel with a degree 8 was used for the raw data.



Accuracy of the model with raw data : 78.70813397129187 %

5.1.2 Confusion Matrix For SVM with Min-Max Scaled Data

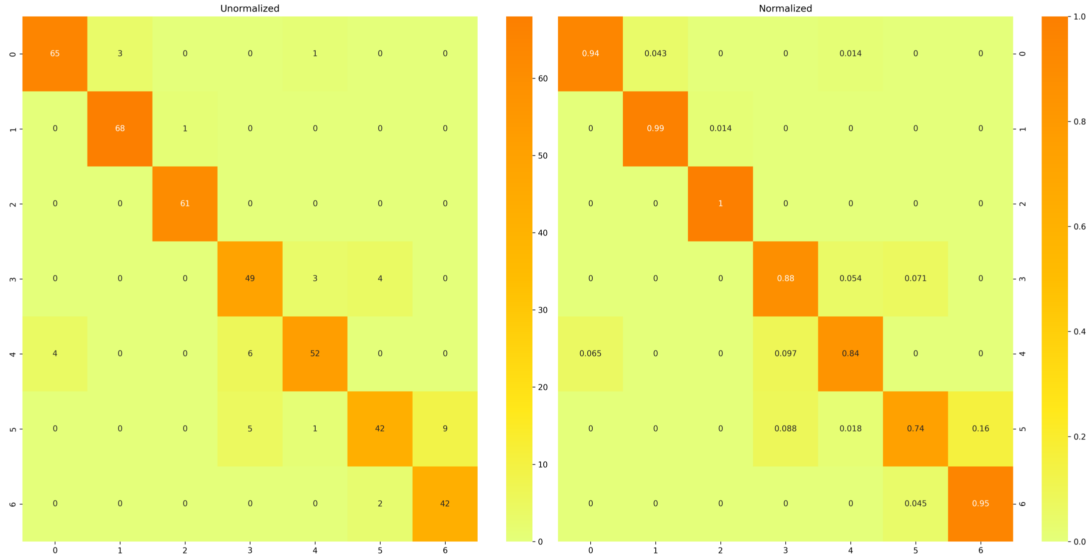
Once the scaling was performed, it was observed that model was good with linear kernel. Hence linear kernel is used. Min-Max Scaler scales all the data features in the range [0, 1] or else in the range [-1, 1] if there are negative values in the dataset. This scaling compresses all the inliers in the narrow range [0, 0.005].



Accuracy of the model with MinMax scalar: 90.66985645933015 %

5.1.3 Confusion Matrix For SVM with Robust Scaled Data

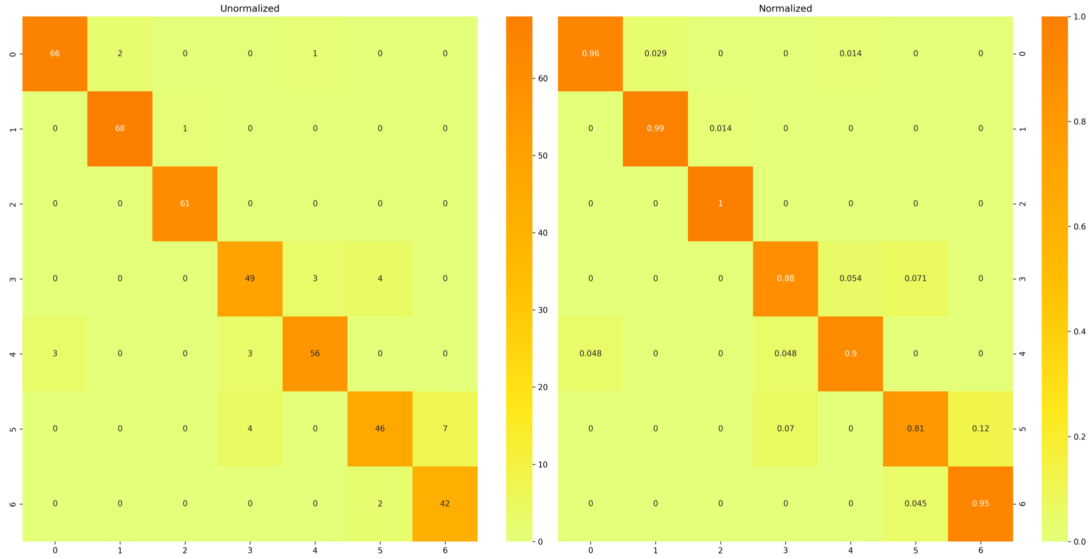
In the presence of outliers, any of the scaling techniques does not guarantee balanced feature scales, due to the influence of the outliers while computing the empirical mean and standard deviation. This leads to the shrinkage in the range of the feature values. By using RobustScaler(), we can remove the outliers.



Accuracy of the model with Robust scalar: 90.66985645933015 %

5.1.4 Confusion Matrix For SVM with Standard Scalar

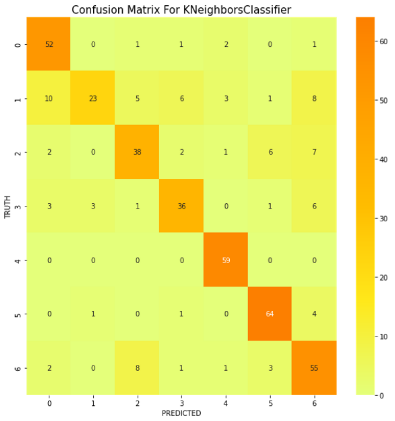
After applying the robust scalar, Standard scalar was used on the data.StandardScaler follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance.



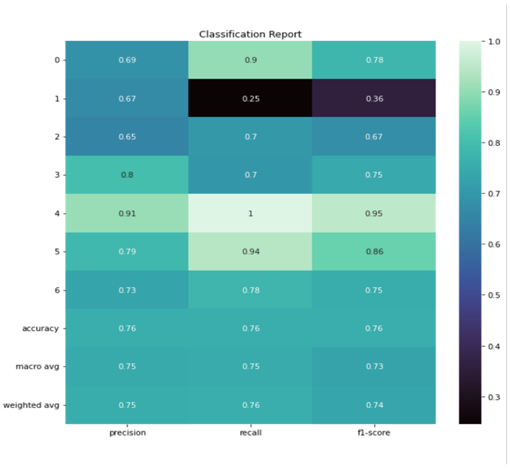
Accuracy of the model with standard scalar: 92.82296650717703 %

5.2 K-Nearest Neighbors

5.2.1 Confusion Matrix and Classification Report



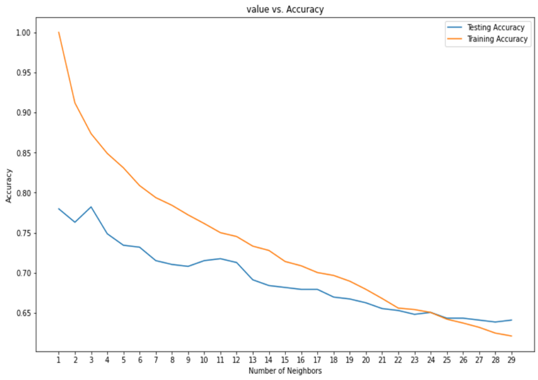
TRUTH \ PREDICTED	0	1	2	3	4	5	6
0	52	0	1	1	2	0	1
1	10	23	5	6	3	1	8
2	2	0	38	2	1	6	7
3	3	3	1	36	0	1	6
4	0	0	0	0	59	0	0
5	0	1	0	1	0	64	4
6	2	0	8	1	1	3	55

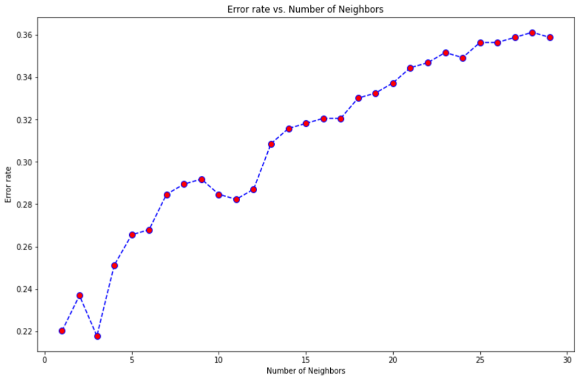


	precision	recall	f1-score
0	0.69	0.9	0.78
1	0.67	0.25	0.36
2	0.65	0.7	0.67
3	0.8	0.7	0.75
4	0.91	1	0.95
5	0.79	0.94	0.86
6	0.73	0.78	0.75
accuracy	0.76	0.76	0.76
macro avg	0.75	0.75	0.73
weighted avg	0.75	0.76	0.74

From classification report the accuracy of given data set with k as 3 is 0.76

5.2.2 Values vs Accuracy and Error rate vs Number of neighbors

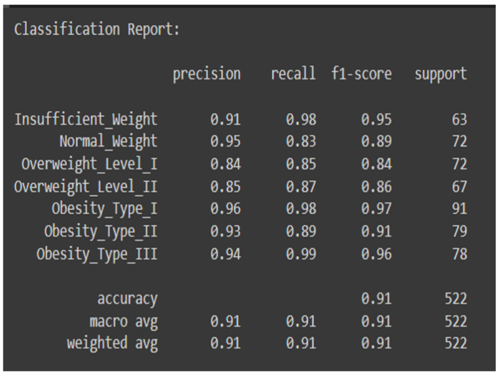




From the above graphs of values vs accuracy and Error rate vs Number of neighbors, we can decide that k=3 gives the best accuracy values for given data.

5.3 Logistic Regression

5.3.1 Classification Report and Confusion Matrix

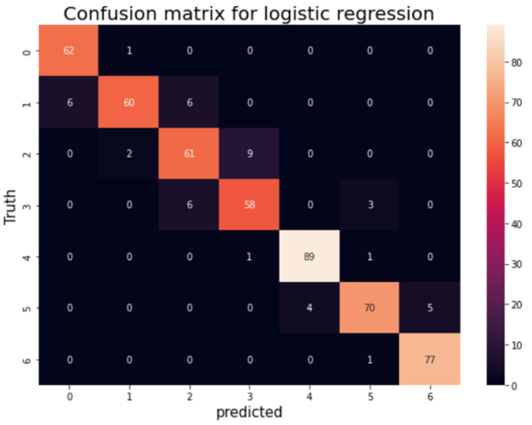


```
Classification Report:

                     precision    recall  f1-score   support

 Insufficient_Weight      0.91      0.98      0.95        63
       Normal_Weight      0.95      0.83      0.89        72
  Overweight_Level_I      0.84      0.85      0.84        72
 Overweight_Level_II      0.85      0.87      0.86        67
      Obesity_Type_I      0.96      0.98      0.97        91
     Obesity_Type_II      0.93      0.89      0.91        79
    Obesity_Type_III      0.94      0.99      0.96        78

            accuracy                          0.91       522
           macro avg      0.91      0.91      0.91       522
        weighted avg      0.91      0.91      0.91       522
```



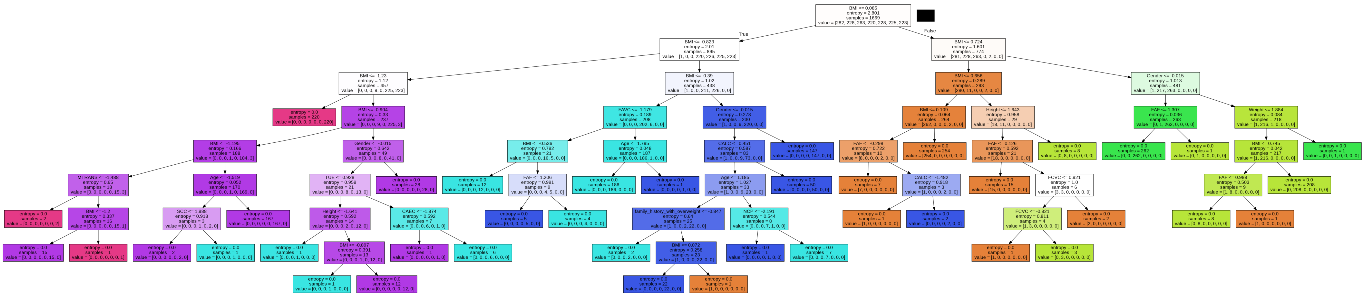
Truth \ predicted	0	1	2	3	4	5	6
0	62	1	0	0	0	0	0
1	6	60	6	0	0	0	0
2	0	2	61	9	0	0	0
3	0	0	6	58	0	3	0
4	0	0	0	1	89	1	0
5	0	0	0	0	4	70	5
6	0	0	0	0	0	1	77

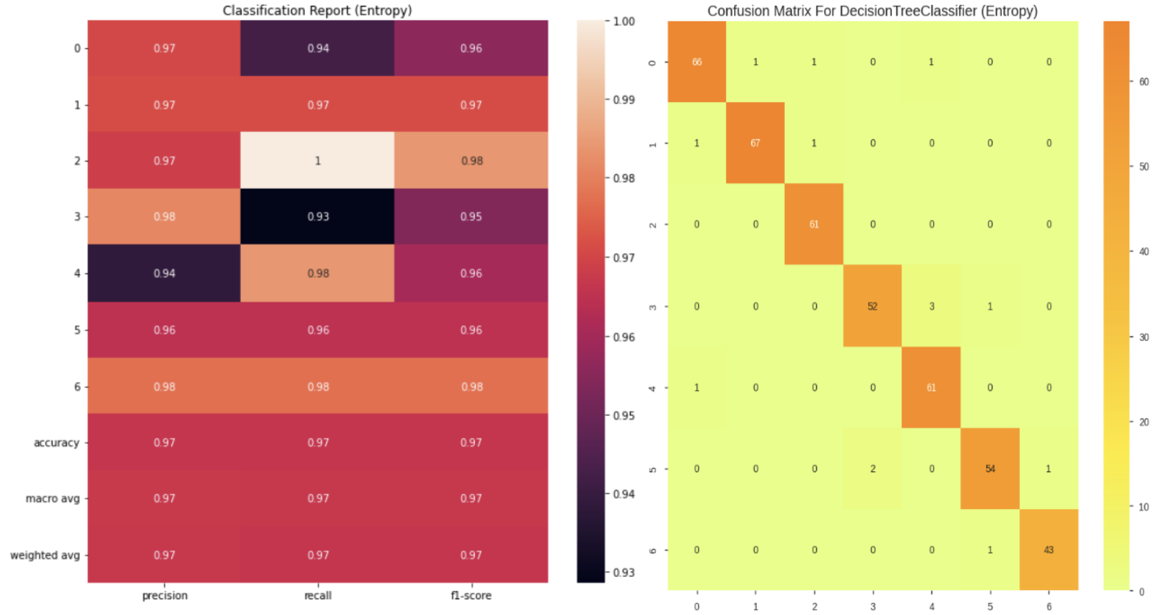
The accuracy of Logistic Regression for this data set is 91.37931034482759

5.4 Decision Trees

Here, Decision Trees are plotted using the Graphviz library which is used to visualize trees and graphs.

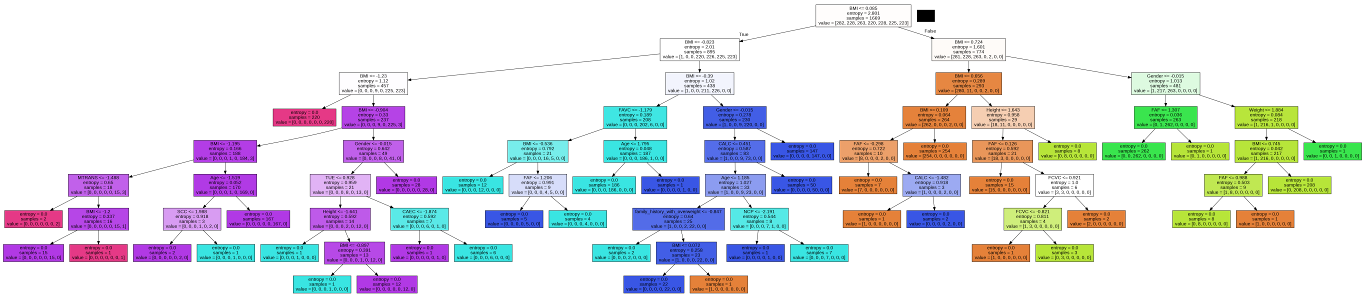
5.4.1 Using Entropy as the criterion

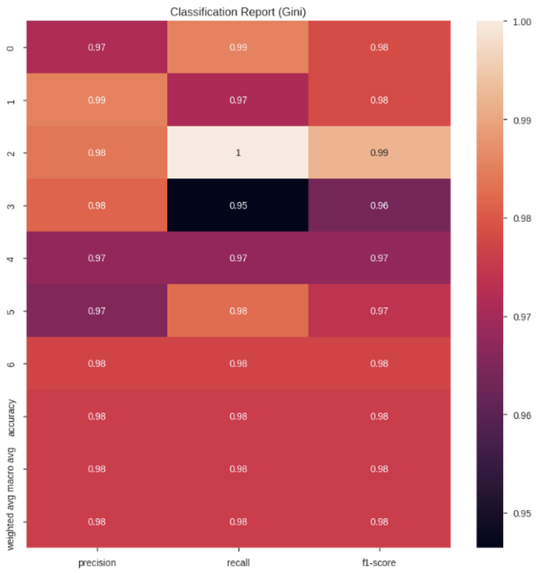


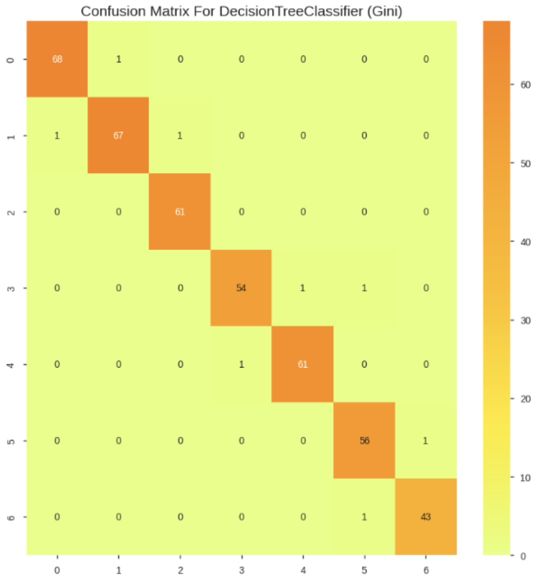


The accuracy of the Decision Tree (Entropy) for this data set is 96.1172

5.4.2 Using Gini as the criterion



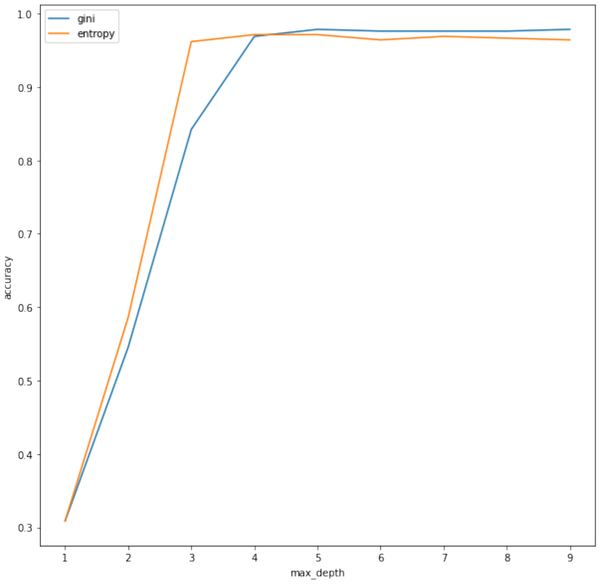




The accuracy of the Decision Tree (Gini) for this data set is 98.3253

5.4.3 Comparing Accuracy

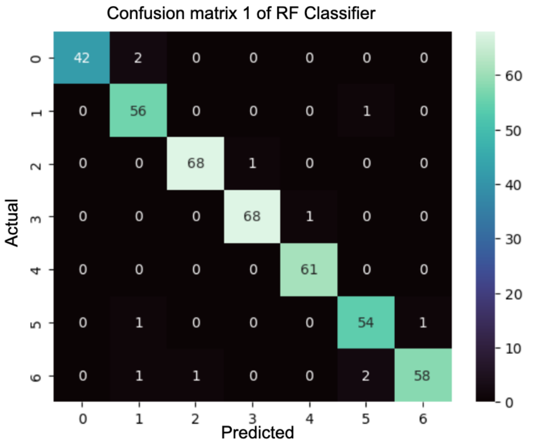
The accuracies of entropy and Gini are plotted by varying the max depth of the tree and we can conclude that entropy performs well for the lower data and as the data increases the Gini performs well as it reduces overfitting.



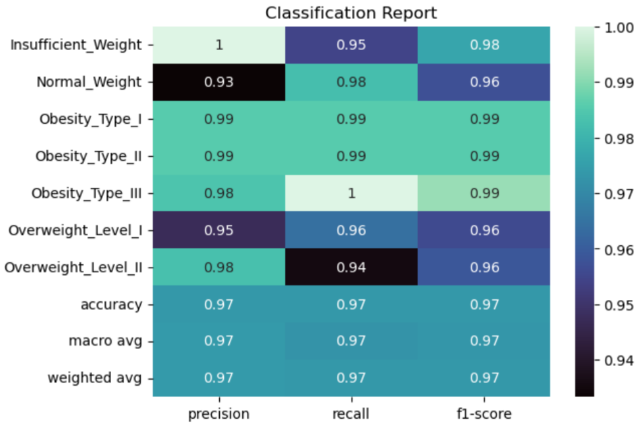
5.5 Random Forest Classifier

Based on the input features, we have trained 2 models, model 1 using all numerical features and important categorical features which are identified using contingency tables and model 2 using all numerical and categorical features.The individual decision trees tend to learn specific patterns based on the features selected and the subset of the data. Each model has confusion matric and classification report.

5.5.1 Model 1- using all numerical and important categorical features


Confusion matrix 1 of RF Classifier

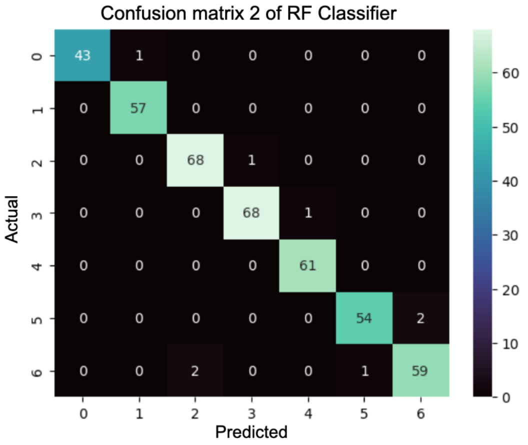
Actual \ Predicted	0	1	2	3	4	5	6
0	42	2	0	0	0	0	0
1	0	56	0	0	0	1	0
2	0	0	68	1	0	0	0
3	0	0	0	68	1	0	0
4	0	0	0	0	61	0	0
5	0	1	0	0	0	54	1
6	0	1	1	0	0	2	58


Classification Report

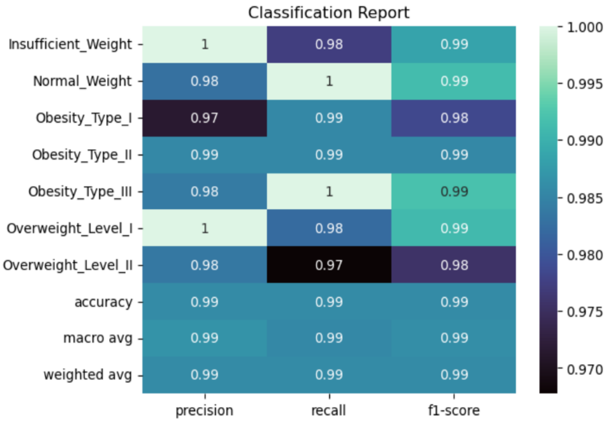
	precision	recall	f1-score
Insufficient_Weight	1	0.95	0.98
Normal_Weight	0.93	0.98	0.96
Obesity_Type_I	0.99	0.99	0.99
Obesity_Type_II	0.99	0.99	0.99
Obesity_Type_III	0.98	1	0.99
Overweight_Level_I	0.95	0.96	0.96
Overweight_Level_II	0.98	0.94	0.96
accuracy	0.97	0.97	0.97
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97

Accuracy of random forest classifier model 1 for this data set is 97.2

5.5.2 Model 2- using all numerical and all categorical features


Confusion matrix 2 of RF Classifier

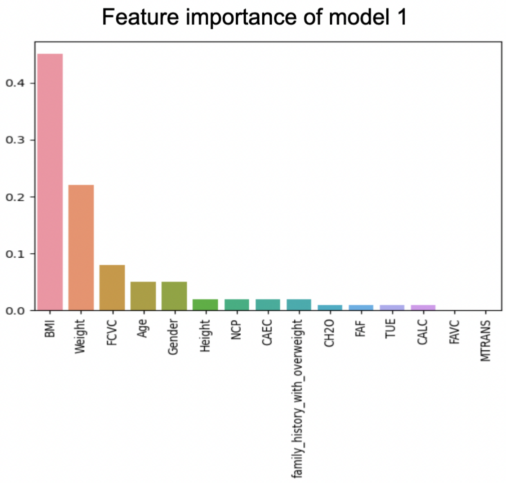
Actual \ Predicted	0	1	2	3	4	5	6
0	43	1	0	0	0	0	0
1	0	57	0	0	0	0	0
2	0	0	68	1	0	0	0
3	0	0	0	68	1	0	0
4	0	0	0	0	61	0	0
5	0	0	0	0	0	54	2
6	0	0	2	0	0	1	59

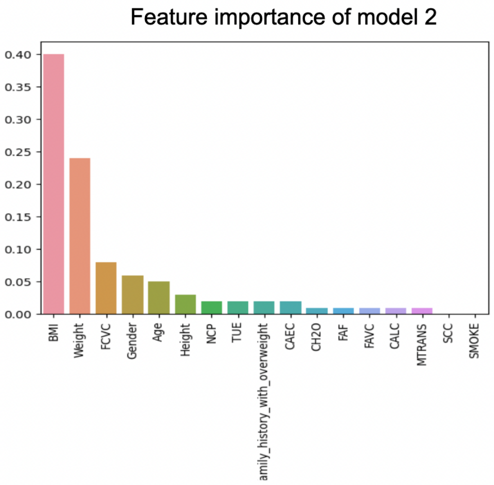

Classification Report

	precision	recall	f1-score
Insufficient_Weight	1	0.98	0.99
Normal_Weight	0.98	1	0.99
Obesity_Type_I	0.97	0.99	0.98
Obesity_Type_II	0.99	0.99	0.99
Obesity_Type_III	0.98	1	0.99
Overweight_Level_I	1	0.98	0.99
Overweight_Level_II	0.98	0.97	0.98
accuracy	0.99	0.99	0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99

Accuracy of random forest classifier model 1 for this data set is 98.5

5.5.3 Feature importance



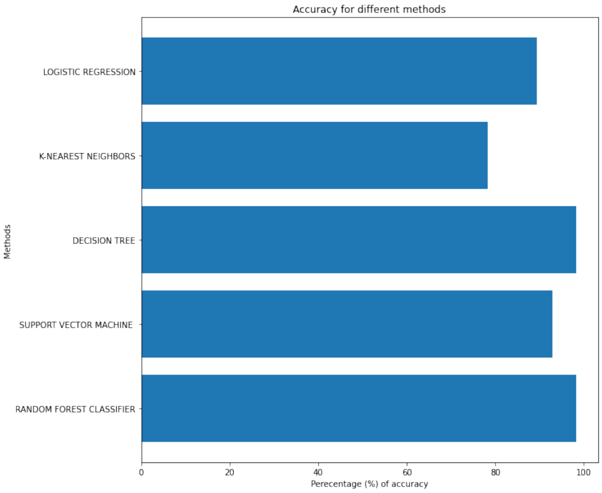


5.5.4 Analysis

As model 1 uses less features, it is less complex. Model 2 shows slightly higher accuracy as all the features are being used. It is a tradeoff between complexity and accuracy. Depending on the context of the problem statement, preference can be given to either of the models.

6 Conclusion

By comparing all the accuracies of the models that are built based on different classification algorithms for our dataset, we can say that the decision trees and the random forest classifiers perform better.



7 Future Scope

for future work, BMI is compared to the obesity index and BMI is compared to a family history with overweight and analysis can be drawn. This analysis can be used to enhance the models.

References

- https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition
- https:https://stacyy.medium.com/itp-449-exploratory-data-analysis-project-obesity-levels-based-on-eating-habits-and-physical-82fa10775c2e
- https://www.neuraldesigner.com/learning/examples/obesity-level
- https://www.chisite.org/events-all/exploring-obesity-disparities-in-marginalized-groups

Contribution

- SVM - Saketh Teki
- KNN- Sai Deekshith Chintalwar
- Logistic regression – Revanth Kumar Chinthagumpala
- Decision trees – Tarun Kumar Bandaru
- Random Forest Classifier– Sameeksha Chiguru