Accuracy of Machine Learning Models in Predicting Obesity Types

Ruchit Patel, Dhyey Patel, Honghao Zhu, *Pace University*

Abstract – Three classification models are compared to find which model will be the most accurate in predicting the Obesity Levels of individual based on 17 different features. The dataset includes information on eating habits and physical characteristics that can be used to estimate the prevalence of obesity in people from Mexico, Peru, and Colombia. After utilizing the K-Nearest Neighbors classifier, Naïve Bayes Gaussian classification, and Light Gradiant Boosting classifier (LGB) in our project, we concluded that the LGB classifier, with its 98 percent accuracy rating, was the best model.

I. INTRODUCTION

Based on their eating patterns and physical conditions, people from Mexico, Peru, and Colombia can have their obesity levels estimated using the data in this dataset. The data is classified using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. The data comprises 2111 records and 17 attributes. The records are labeled with the class variable NObesity (Obesity Level). 23% of the data was gathered directly from users via a web platform, while the remaining 77% of the data was artificially created using the Weka tool and the SMOTE filter. The following classifiers were employed in this project:

i. K-Nearest Neighbors (KNN)

The basic idea of similarity is the basis for how the KNN classifier functions. It makes the assumption that similar objects are found nearby, as in the saying "birds of a feather flock together." An object in KNN is categorized by the majority vote of its neighbors. The object is categorized into the class that its 'k' closest neighbors (where 'k' is a positive integer, usually small) share the most of. The object is just put into the class of its closest neighbor if k=1.The method used to measure the distance between data points is the essential part of KNN. Manhattan, Hamming, and Euclidean distances are examples of common measurements. For classification and regression prediction issues, KNN is frequently utilized. For large datasets, it is computationally costly, sensitive to the size of the data, and contains extraneous features.

ii. Naïve Bayes classifiers

naïve Bayes classifiers is based on using the Bayes theorem with strong (naïve) independence assumptions between the features.
The continuous values connected to each class are thought to follow a Gaussian (normal) distribution in the Gaussian model.
The variance and mean of the features in the training set define the Gaussian distribution for a given class. The classifier determines which class has the highest probability when classifying a new data point by calculating the probability that the data point belongs to each class.

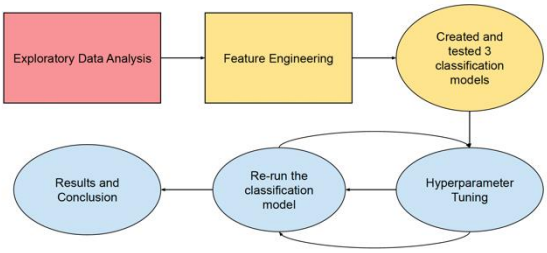
iii. Light Gradient Boosting Classifier (LGB)

A gradient boosting framework called LightGBM makes use of tree-based learning algorithms. With its distributed and efficient design, it offers the following benefits: better accuracy, reduced memory usage, quicker training speed, and support for GPU and parallel learning.
LightGBM grows trees in a leaf-wise (vertical) manner as opposed to other boosting techniques, which grow level-wise (horizontally). The leaf with the maximum delta loss will be selected for growth. When compared to a level-wise algorithm, this can reduce loss more.
It requires less memory to operate and can handle large amounts of data. Focusing on result accuracy is one of LightGBM's other key advantages.
Additionally, the model directly supports categorical features.

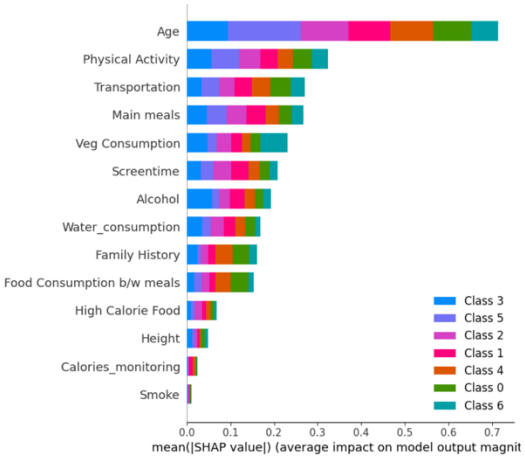
II. RELATED WORK

In their paper on early education intellectual development Mo and LI [1] build a classification model of children's language ability development based on curriculum characteristics which was created using the LightGBM method. The outcome demonstrated that the kids' intellectual growth was consistent with the teachers' experiences. They arrived to the conclustion that the LightGBM methond is feasible to build a classification model due to the efficiency being 89.1%. Another research by authors Martono, Kuramaru, Igarashi, Yokobori and Ohwada on blood alcohol concentration screening used multiple classification models including K-nearest neighbor and LightGBM. In which LightGBM model had accuracy rate being 90.8%. In addition, the study makes use of SHAP (Shapley Additive Explanations) to determine the significance of each feature and give the model interpretability. According to this analysis, certain variables have a high feature importance. These variables include calcium ions, glucose, sodium ions (anion gap), carboxyhemoglobin concentrations, lactate, and pH. The physiological and biochemical elements pertinent to blood alcohol concentration prediction are clarified by these findings.

III. Methodology



We took inspiration for the research paper and decided to use SHAP(Shapley Additive Explanation) in order to visualize our date in an interesting data which helps us understand the relationship between our features and the target variable. Our dataset had a mixture of both categorical and numerical data, so in order to get our data ready for classification models some feature engineering was involved. Converting our categorical values to numerical we were able to train our model.

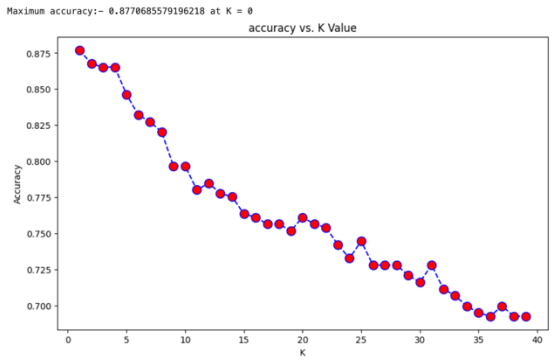


The most impactful feature which had a high impact to obesity level was weight. It was left out of the list, the list above display the next 14 variables which have an impact on obesity level.

- KNN Classification

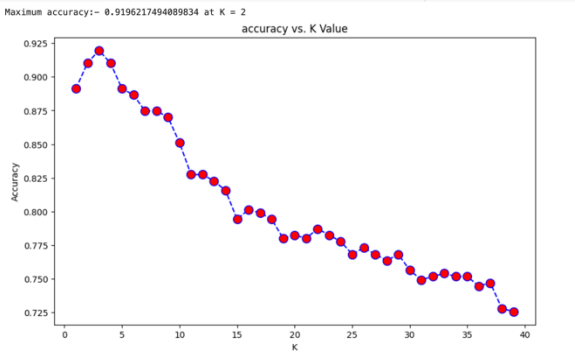
Euclidean KNN Classification

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$



Manhattan KNN Classification

$| x_2 - x_1| + |y_2 = y_1|$



We found the Manhattan distinct to work best in our case. Although this might be the case all the time. K value of 2 will give us the best accuracy. We also tuned the hyperparameters of the LightGBM model by creating a grid search object which goes through many combination of parameters to find the best combination of learning_rate, n_estimators, num_leaves and max_depth to get the most accurate predictions.

IV. Results

1) KNN Classifier was divided into two section, Manhattan and Euclidean. The major difference being how the distance between two point is being calculates accuracy scores produced by these models have a drastic difference.

- Euclidean distance

	precision	recall	f1-score	support
0	0.42	0.44	0.43	54
1	0.64	0.52	0.58	65
2	0.51	0.43	0.47	49
3	0.71	0.65	0.68	65
4	0.76	0.92	0.83	60
5	0.77	0.87	0.82	54
6	0.88	0.92	0.90	76
accuracy			0.69	423
macro avg	0.67	0.68	0.67	423
weighted avg	0.68	0.69	0.69	423

This is the results before finding the right K value and running fit model. After the scored improved when we set the n_neighbors=1 parameter to be 1 as per the k value indicated.

```
Accuracy: 0.8770685579196218
Precision: 0.875258000209026
F1 score: 0.8794935296224927
```

This make the score improve by 18.7%.

- Manhattan distance

	precision	recall	f1-score	support
0	0.50	0.46	0.48	54
1	0.66	0.57	0.61	65
2	0.61	0.47	0.53	49
3	0.71	0.62	0.66	65
4	0.77	0.98	0.86	60
5	0.80	0.96	0.87	54
6	0.88	0.93	0.90	76
accuracy			0.73	423
macro avg	0.70	0.71	0.70	423
weighted avg	0.71	0.73	0.71	423

The scored improved when we found the lowest error rate at k = 2. Setting the n_neighbors=2 improved the accuracy score by 18 %.

```
Accuracy: 0.9101654846335697
Precision: 0.9109105684504621
F1 score: 0.9102321234492966
```

2) Naive Bayes Gaussian Classification

	precision	recall	f1-score	support
0	0.44	0.28	0.34	54
1	0.47	0.29	0.36	65
2	0.70	0.29	0.41	49
3	0.31	0.40	0.35	65
4	0.65	0.82	0.73	60
5	0.57	0.93	0.70	54
6	0.93	0.99	0.96	76
accuracy			0.59	423
macro avg	0.58	0.57	0.55	423
weighted avg	0.59	0.59	0.56	423

The over accuracy score was really disspointing in the Naïve Bayes Gaussian Classification, to improve the accuracy score, we found and tuned hyperparameters. We used cross-validation using KFold = 5. We set up grid search parameters and a parameter var_smoothing = 0.000811. Now running the Naïve Bayes model again with the new parameter improved the score by 3%

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
Best Parameters: {'var_smoothing': 0.0008111308307896872}
Test Accuracy: 0.624113475177305
```

3) Light Gradian Boosting classifier (LGB)

By far the highest improvement is seen in the Light Gradian Boosting classifier (LGB) model. First time running the model gave us the accuracy score of 25.77% which we were disappointed by.
Again the use of Grid Search Cv was seen extremely useful as it helped in identifying important hyperparameters:

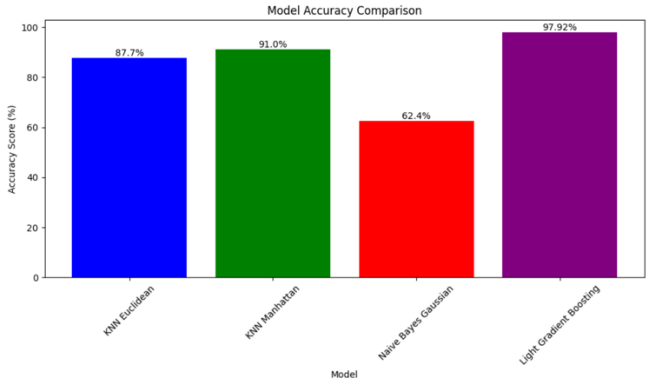
learning_rate = 0.1
n_estimators = 200
num_leaves = 20
max_depth = 5

```
Accuracy score:  0.9792636033220375
```

4) Overview

Models	Accuracy Score
KNN Euclidean	87.7%

KNN Manhattan	91%
Naive Bayes Gaussian Classification	62.4%
Light Gradian Boosting classifier (LGB)	97.92%



V. Conclusion

The purpose of this project was to build different classification models that can classify various levels of obesity using 17 independent features and compare them based on 3 metrics i.e. Accuracy, Precision, and F1 score.

For the first part of the project, data cleaning, and EDA were performed to clean the data of any missing or duplicated values and find the relation among various features to better understand the data.

And, for the classification modeling, KNN Classification (Euclidean and Manhattan distance), Naïve Bayes Gaussian, and Light Gradient Boosting(LGB) models were developed to test the initial performance of the models. The result being KNN model using Manhattan distance performed the best as accuracy came out to be around 91%, followed by KNN(Euclidean distance), Naïve Bayes and LGB model performing poorly with accuracy of around 89%, 57% and 25% respectively. Since, the performance of Naïve Bayes and Gaussian models were not satisfactory, it prompted us to perform hyperparameter tuning using grid search cross-validation to check if it improved any metrics. As a result, the metrics exponentially increased for the LGB model and a small improvement in the Naïve Bayes. To conclude the project, LGB model(with hyperparameter tuning) performed the best with accuracy, precision and f1 score of 98%, 96%, and 96%, respectively.

VI. References

Y. Mo and X. Li, "Research on Classification Modeling Algorithm of Language Development Curriculum Based on Iterative LightGBM Method," 2022 IEEE 4th Eurasia Conference on IOT, Communication and Engineering (ECICE), Yunlin, Taiwan, 2022, pp. 373-377, doi: 10.1109/ECICE55674.2022.10042882.

N. P. Martono, S. Kuramaru, Y. Igarashi, S. Yokobori and H. Ohwada, "Blood Alcohol Concentration Screening at Emergency Room: Designing a Classification Model Using Machine Learning," 2023 14th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 2023, pp. 255-260, doi: 10.1109/ICTS58770.2023.10330879.