Prediction of the Body Mass Index Category

Serhat Habil Çelik
Statistics
Middle East Technical University
Ankara, Turkey
serhat.celik_01@metu.edu.tr

***Abstract*—In this report, Body Mass Index Category is analysed via R Studio. By looking at heart rate or stress level, peoples' BMI Category is determined. Before going modelling, data was cleaned to work on data comfortably by strgingr library in R Studio 4.4.0. Then, scaling is done to the data. After that, feature engineering and regularization is used to remove overfitting if it exists. After all these steps, six different models are done in due order. Our aim here is to find the best model comparing their accuracy and F1 Score at the end.**

***Keywords*—LASSO, Random Forest, Neural Network, Naïve Bayes, Logistic Regression, SVM, XGBoost**

I. Introduction

Body mass index (BMI) is a tool which doctors and dietitian use to estimate the amount of body fat by using height and weight measurements. There are some subgroups in BMI such as underweight, healthy weight, overweight and obese. However, normal, obese and overweight categories will be used for prediction in this project. They are adjusted for gender and age. BMI category can have different values according to gender and their age. In the given data, there are several variables helping to find BMI category of people. In this research, three different categories are predicted by the different modelling techniques such as Random Forest and Logistic Regression.

II. METHODOLOGY

A. Dataset

The dataset is taken from https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset.The data have 374 observations and 13 variables itself. Four of the variables are categoric and the rest are numeric values, except Blood Pressure. Dependent variable is selected as BMI Category in this research. The dataset does not contain missing values. It will be created with R Studio. Description of variables is given in Table 1.

Variable	Type of Variable	Description
Person ID	Numeric	An identifier
Gender	Categoric	The gender of the person (Male, Female).
Age	Numeric	The age of the person in years.
Occupation	Categoric	The occupation or profession of the person.
Sleep Duration	Numeric	The number of hours the person sleeps per day.
Quality of Sleep	Numeric	The rating of the quality of sleep, ranging from 1 to 10.
Physical Activity Level	Numeric	The number of minutes the person engages in physical activity daily.
Stress Level	Numeric	The rating of the stress level experienced by the person, ranging from 1 to 10.
BMI Category	Categoric	The BMI category of the person (Underweight, Normal, Overweight).
Blood Pressure	Character	The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.
Heart Rate	Numeric	The resting heart rate of the person in beats per minute.
Daily Steps	Numeric	The number of steps the person takes per day.
Sleep Disorder	Categoric	The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

Table 1: Description of Variables

B. Data Cleaning

At first, cleaning and tidying of data should be done by R Studio 4.4.0. The first step in data cleaning is deleting the dot between names of column in the original data. Not applicable to the analysis, person id is removed from the original data. After that, we move the second step which is to remove the blood pressure column from the original data due to the fact that its original format is impractical. Blood pressure is separated into two new columns called systolic pressure and diastolic pressure by using stringr library. Being numeric, they could be used easily for future steps. Finally, converting gender, occupation, bmicategory and sleep disorder from numeric to factor is done. In this step, we try to avoid any inconsistency in the data for future analysis.

C. *Descriptive Statistics*

Descriptive statistics epitomises traits of a data set. It is done with continuous variables here. Systolic pressure and diastolic pressure are added here. Summary of nine continuous variables is given in Table 3 and Table 4.

	Age	Sleep Duration	Sleep Quality	Activity Level	Stress Level
Min.	27.00	5.800	4.000	30.00	3.000
1st Qu.	36.00	6.400	6.000	45.00	4.000
Median	43.00	7.200	7.000	60.00	5.000
Mean	42.31	7.111	7.319	59.03	5.385
3rd Qu.	50.00	7.800	8.000	75.00	7.000
Max.	59.00	8.500	9.000	90.00	8.000
NA's	16	27	13	22	16

Table 2: Descriptive Statistics of Continuous Variables

	Heartrate	Daily Steps	Systolic Pressure	Diastolic Pressure
Min.	65.00	3000	115.0	75.00
1st Qu.	68.00	5600	125.0	80.00
Median	70.00	7000	130.0	85.00
Mean	70.16	6824	128.6	84.63
3rd Qu.	72.00	8000	135.0	90.00
Max.	86.00	10000	142.0	95.00
NA's	20	17	20	19

Table 3: Descriptive Statistics of Continuous Variables

There are nine numeric values in the cleaned data. A consideration of the above table shows that the first quantile (Q1) of heartrate is 68, which means that 25% of the data fall below 68. Also, the third quantile (Q3) of heartrate is 72, and it means that 25% of the data fall above 72. It has 20 NA (not available). Additionally, while the minimum value of daily steps is 3000, the maximum value of them is 10000. Its median is 7000, a value which means that half of the data fall below 7000 and the rest fall above 7000. It also has 17 NA. Its mean value is smaller than median, which can indicate that it may be right-skewed. There are four categoric variables. However, summary of two of them will be discussed in this report. One of them is gender and the second is BMI Category.

Frequency of Genders	Male	Female
	172	180

Table 4: Frequency of Genders

Frequency of BMI Category	Normal	Obese	Overweight
	209	9	141

Table 5: Frequency of BMI Category

When we look at the Table 4, the number of females appears to be bigger than that of males. It could be concluded that there is not a big difference between them. The Table 5 shows that there are 204 Normal, 9 Obese and 143 Overweight people according to the BMI Category. It is an abbreviation of Body Mass Index defined as the body mass divided by the square of the body height. It is obviously seen that the number of normal people is dominant compared to obese and overweight people. However, obese and overweight will be combined to prevent imbalanced classification problem in the future analysis.

D. *Exploratory and Confirmatory Data Analysis*

In this part, three different questions have been answered and analysed using R Studio. Chi – Square and ANOVA tests have been used for confirmatory data analysis. For the plots, data with missing values were used. Plots used in EDA were done with missing values; however, CDA was done with imputed data here.

Question I - Is there any relationship between heart rate and daily steps?

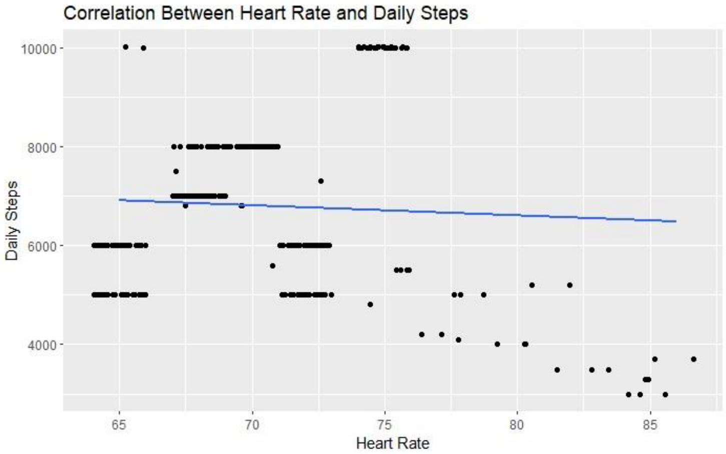

Figure 1 Scatter Plot of Heart Rate and Daily Steps

There exists negative significant connection between heart rate and daily steps looking at the plot. When heart rate goes up, daily steps decrease sharply as seen. Jittering was used for Figure 1.

Question II - Do different BMI Categories shows significantly different sleep duration?

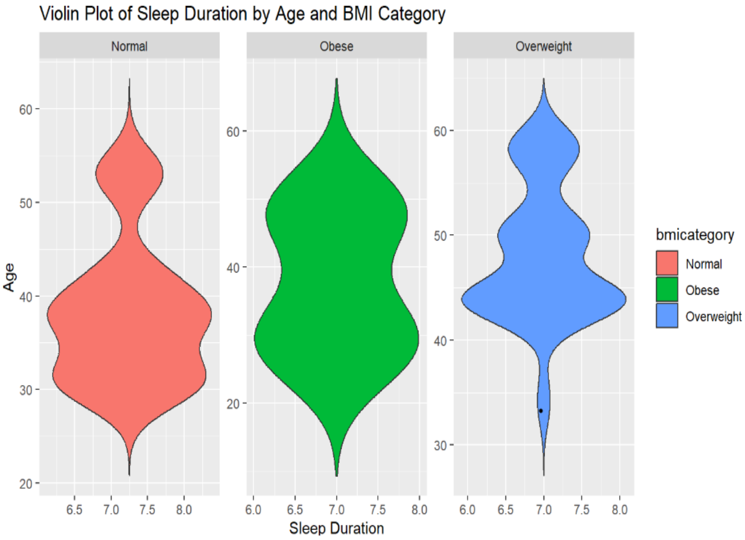

Figure 2 Violin Plots of Sleep Duration by Age and BMI Category

Kruskal – Wallis rank sum test was applied for this question. This is because the data are not normally distributed despite all efforts to make it normal as shown in Table 6. According to the Kruska – Wallis rank sum test, there are significant differences between the treatment groups because p – value is less than 2.2e-16. The output is given below in Table 6.

Best Normalize	Box – Cox	Scaling
1.121 e-11	< 2.2e-16	< 2.2e-16

Table 6: P – value of Transformations

Kruskal – Wallis rank sum test		
data: sleepduration by bmicategory		
Kruskal – Wallis chi – squared = 222.61	df=2	p-value < 2.2e-16

Table 7: Kruskal – Wallis Rank Sum Test

Question III - How do gender frequencies vary across different BMI Category?

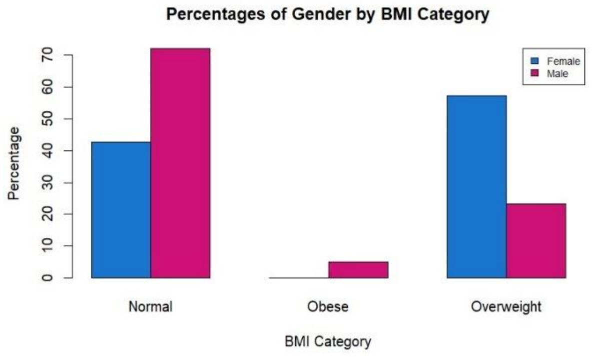

Figure 3 Bar Plot of BMI Category in Terms of Gender

	Normal	Obese	Overweight
Female	529.6043	22.55722	364.8385
Male	550.3957	23.44278	379.1615

Table 8: Expected Value of Cells

One of the assumptions of Chi – Square is that the expected value of cells should be 5 or greater in at least 80% of cells. As it is seen in Table 7, it is satisfied after finding expected value of cells. The results of Chi Square Test shows that p value is lower than 2.2e-16 so there is a strong significant relationship between bmicategory and gender. Details exists in R Markdown.

E. Missingnes

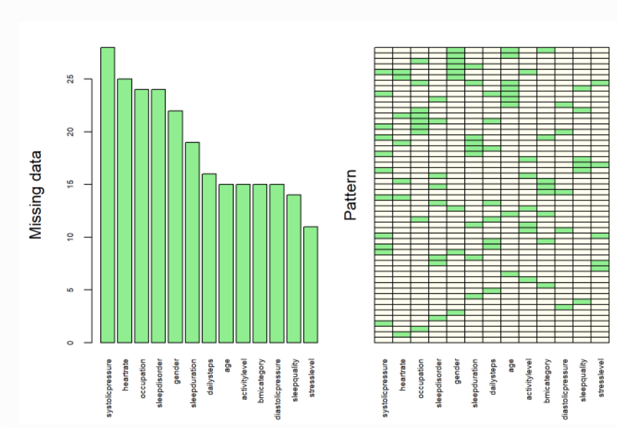

Figure 4 Missingness Pattern

It is of importance to be aware of the missing mechanism of the data that will be analysed in a correct way. In our data, there is no relationship between the probability of being missing and the values of any of the variables in the data in terms of dependency. This is why missingness mechanism could be defined as Missing Completely at Random (MCAR). Additionally, according to the mcar_test function in R, it fails to reject H_0, meaning that the data are MCAR. We have used mice package as described in Van Buuren and Groothuis-Oudshoorn to impute missing values in the data. The method is selected as pmm, an abbreviation of predictive mean matching. The number of NA's is 243, while it is 0 after imputation. The distribution is checked to see whether it has changed or not after imputing in Figure 5 below. According to the Figure 5, imputed data fit very well to the model. In this step, scaling was applied to imputed variables. The rest will continue with scaling values except categoric ones.

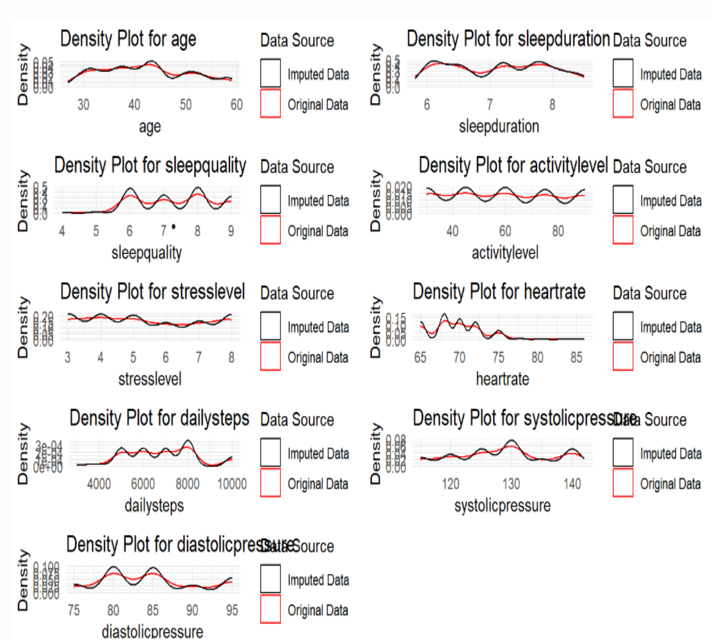

Figure 5 Distribution Before and After Imputation

III. PREPARATION FOR MODELING

As we all know, modelling is a substantial part of data analysis. Statisticians can make predictions regarding future. Firstly, obese and overweight are combined to remove imbalanced classification problem. After doing this, regularization was done for data. Regularization was followed by feature engineering and cross validation.

Frequency of BMI Category	Normal	Overweight
	209	150

Table 9: Frequency of BMI Category after Combining

A. Regularization

Regularization is a technique used to calibrate models to minimize the adjusted loss function and inhibit overfitting or underfitting. Ridge Regression, LASSO and Elastic Net have been used to reduce the effect of multicollinearity. Their RMSE values are shown in Table 10 below. According to these results, Elastic Net is the best regularization method because it has lower RMSE compared to Ridge and LASSO.

Ridge	Lasso	Elastic Net
0.1388098	0.1387886	0.1387874

Table 10: RMSE of Regularization Techniques

B. Feature Engineering

Feature Engineering can be defined as the process of building new features or transforming existing features in order to develop the performance of a machine-learning model. The aim is to improve model accuracy by providing more meaningful and relevant information. In this context, LASSO and BORUTA feature selection methods were applied to the data. According to the results of LASSO and BORUTA, all variables are important. Removing any variable from the model is not necessary according to the results. RMSE of BORUTA is lower than LASSO. Its value is 0.1343921. Detail explanation is given in R Markdown. Encoding values are used here.

C. Cross – Validation

Cross validation makes it possible for us to have cases in our testing set that are different from the ones in our training set, thus innately offering protection against overfitting. We will use four different cross validation techniques in our project. Table 10 provides RMSE values of each technique separately. K – fold cross validation is the best choice for cross validation. This is because its RMSE is the lowest compared to others. In the statistical modelling part, we will use K – Fold Cross Validation as the train and test data. It was the last step before proceeding the statistical modelling.

	Validation Set	LOOCV	K – Fold Cross Validation	Repeated K – Fold Cross Validation
RMSE	0.13894	0.137411	0.1365815	0.1366733

Table 11: RMSE of Cross Validation Techniques

IV. MODELING

Before modelling, regularization is conducted to check whether overfitting exists or not in the model. Ridge, LASSO and Elastic Net were evaluated with different lambda values. Train function in caret package selects best lambda and regularization type. According to the results, elastic net is the best model with lambda = 0.001. Its accuracy is 0.989, while its F1 score is 0.990. Hence, accuracy and F1 score are too high, which means that there is no overfitting. The model has efficaciously learned the patterns in the data. We can go on with logistic regression.

A. Logistic Regression

Bmicategory is taken as response here. We first check VIF values of independent variables. VIF of sleep duration, sleep quality, activity level, stress level, heart rate, systolic pressure and diastolic pressure are higher than 10, which is a threshold for VIF. K-Fold Cross Validation is used in logistic regression. Accuracy and F1 Score are shown in Table 12.

Accuracy	F1 Score
0.975	0.973

Table 12: Accuracy and F1 Score of Logistic Regression

B. Random Forest

Random Forest is a supervised learning technique which can be considered as a modification of bagging. The best mtry parameter is found 12 by caret package in R. Accuracy and F1 Score are shown in Table 13.

Accuracy	F1 Score
1	1

Table 13: Accuracy and F1 Score of Random Forest

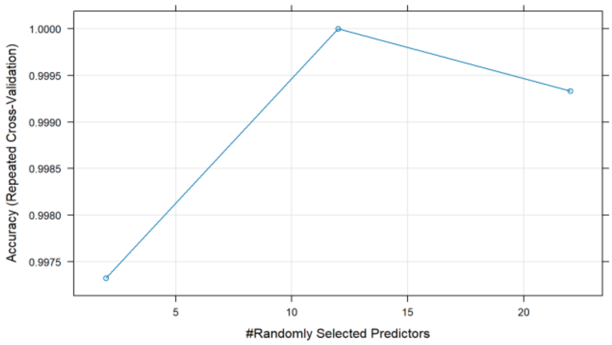

Figure 6 The Best Mtry value of Random Forest

C. Neural Network

Artificial Neural Network is used to learn and improve the accuracy of training data in time. The best Tunes for the Neural Network are size 5 and decay 0.1 found by train function embedded in R. BMI Category is taken as response with two levels named Normal and Overweight. Accuracy and F1 Score are shown in Table 14.

Accuracy	F1 Score
0.991	0.992

Table 14: Accuracy and F1 Score of Neural Network

D. Support Vector Machines

Support Vector Machine (SVM) is a powerful and versatile machine learning algorithm primarily used for classification in modelling. SVM aims to find the best decision boundary that could maximize the margin between classes, thereby achieving a good generalization performance. Accuracy and F1 Score are shown in Table 15.

Accuracy	F1 Score
0.9973	0.9978

Table 15: Accuracy and F1 Score of SVM

E. Naïve Bayes

Naive Bayes algorithm is primarily based on Bayes theorem found by Thomas Bayes. K- Fold Cross Validation is as a cross-validation technique here. Accuracy and F1 Score are shown in Table 16.

Accuracy	F1 Score
0.978	0.982

Table 16: Accuracy and F1 Score of Naive Bayes

F. XGBoost

XGBoost only works with numeric vectors. Therefore, we need to convert all other forms of data into numeric vectors. After converting categoric variables into numeric vectors, max depth and nround are determined to be used in the XGBoost model. K- Fold Cross Validation is used in XGBoost. Here, we cannot achieve MSE scores because it works with numeric values. Its MSE score is equal to 0.3937599.

nround	max_depth	eta	gamma
150	3	0.4	o

Table 17: Best Tune for XGBoost

V. PERFORMANCE COMPARISON

When comparing performance of models, we will use Accuracy and F1 score for each model. The higher accuracy the better model. Accuracy and F1 Score is achieved from confusion matrix of each model created by R separately.

VI. RESULTS

There are 6 models benefited in this project. These are Logistic Regression, Random Forest, Neural Network, SVM, Naïve Bayes and XGBoost, respectively. Accuracy and F1 Score of all are so close to each other. However, random forest is the best model compared to other models because its F1 Score is equal to 1, which means perfect precision and recall. In general, it can be said that the data are very good at predicting BMI. There is no problem with the data.

VII. CONCLUSION

This work started with brief introduction of Body Mass Index at first. Then, it continued with a data description of the data. After the data description, data cleaning was done properly to get more logical results. In addition, we used regularization, feature engineering and cross validation in our project. These steps are so crucial because removing multicollinearity or checking overfitting is done here. Lastly, six different classification models were used to assess model performance. F1 score was a bit useful for the data. Therefore, it is preferred when analysing results achieved from models.

VIII.REFERENCES

[1] Hastie, T., Martin, R., Hastie, W., Tibshirani, & Wainwright. (n.d.). Monographs on Statistics and Applied Probability 143 143 Statistics Statistical Learning with Sparsity copy to come from copywriter for review The Lasso and Generalizations Statistical Learning with Sparsity. https://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS_corrected_1.4.16.pdf

[2] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

[3] IBM. (2023). *What Are Neural Networks?* Www.ibm.com; IBM. https://www.ibm.com/topics/neural-networks