CALORIES BURNED: Is it optimum of less?- A CASE STUDY

MARY JOHN

2148140

INTRODUCTION

Some people blame their weight on how their body breaks down food into energy, also known as metabolism. It's true that the rate at which the body breaks down food is linked to weight. But a slow metabolism isn't usually the cause of weight gain. Metabolism does help decide how much energy a body needs. But weight depends on how much a person eats and drinks combined with physical activity.

You can't easily control the speed of your basal metabolic rate, but you can control how many calories you burn through physical activity. The more active you are, the more calories you burn.

Technology has seen a fast pace revolution with watches being produced that keep in check with fitness of a person. Current fitness activity trackers can account for steps, calories burned, heart rate, and distance travelled.

ABOUT THE DATA

This dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 04 December 2016 – 01 July 2019. Thirty eligible female Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID or timestamp. Variation between output represents use of different types of Fitbit trackers and individual tracking behaviours / preferences. The data is public data from FitBit Fitness Tracker Data. It's a dataset from thirty female FitBit users that includes minute-level output for physical activity, heart rate, and sleep monitoring. It's a good database segmented in several tables with different aspects of the data of the device with lots of details about the user behaviour. There are total of 1000 observations were present.

PROBLEM STATEMENT

The maintenance the wellness or health of a person based on his or her weight that is calories is growing need in today's world. With smartwatches always in play, tracking our every movement, we can now determine how many calories we have lost. But how many calories lost is good enough for our health or to reduce our weight.

DATA PREPROCESSING

We cleaned and refined our data by replacing 0 in the null values indicating that there wasn't any activity in the FitBit user's app from the watch. We also converted the Activity Date into Date time format from object format and created extra columns as months, years and days from it. We also created extra columns based on the time taken to sleep once in bed and also based on the calories burned.

Time taken to sleep:

1- 180 minutes or more
2- Between 100 to 180 minutes
0- Less than 100 minutes

Calories Burned:

1: High calories burned (>=2000)

0: Less calories burned (>2000)

METHODOLOGIES USED

- LOGISTIC REGRESSION

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

Properties of Logistic Regression:

1. The dependent variable in logistic regression follows Bernoulli Distribution.
2. Estimation is done through maximum likelihood.
3. No R Square, Model fitness is calculated through Concordance, KS-Statistics.

- K NEAREST NEIGHBOURS CLASSIFICATION

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

- GAUSSIAN NAIVES BAYES

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. We have explored the idea behind Gaussian Naive Bayes along with an example.

Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality. They find use when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naive Bayes Classifier.

- DECISION TREE

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

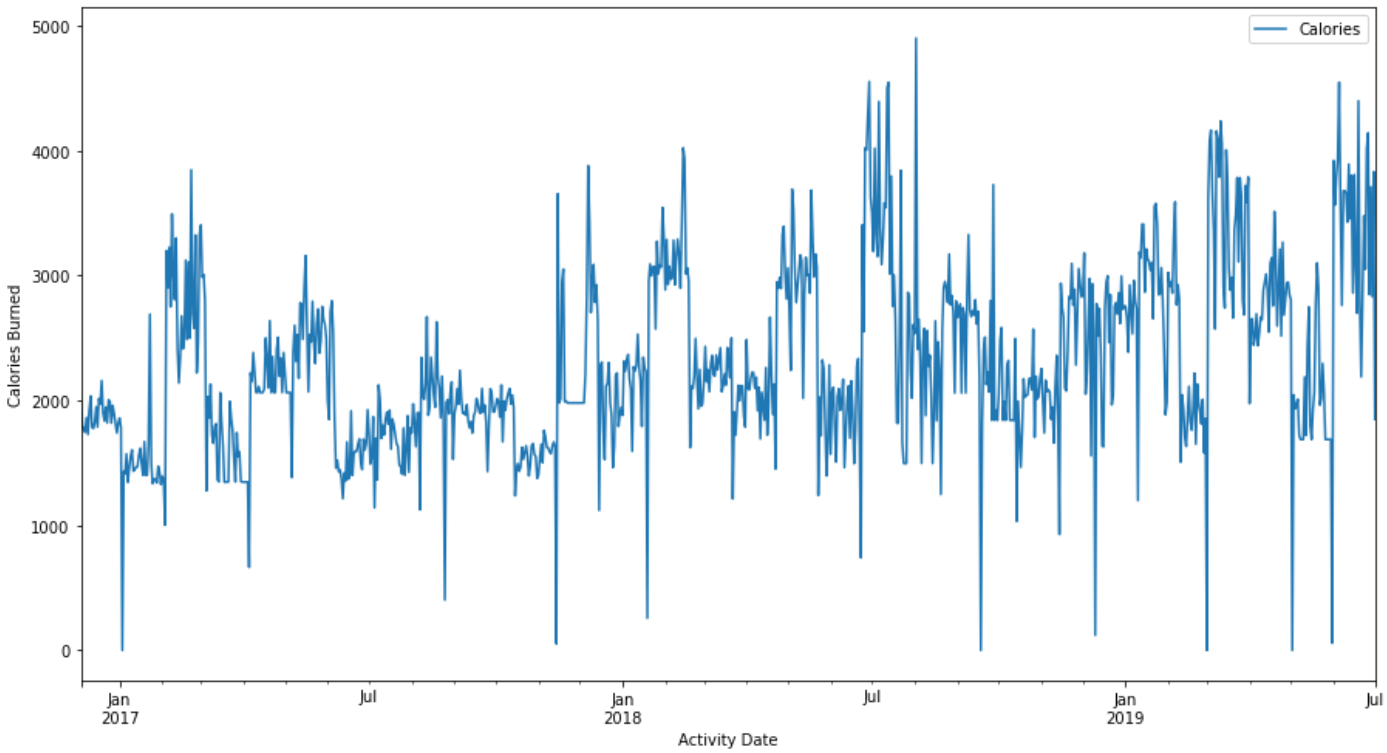
- RANDOM FOREST

The Random Forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees.

The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then. It collects the votes from different decision trees to decide the final prediction.

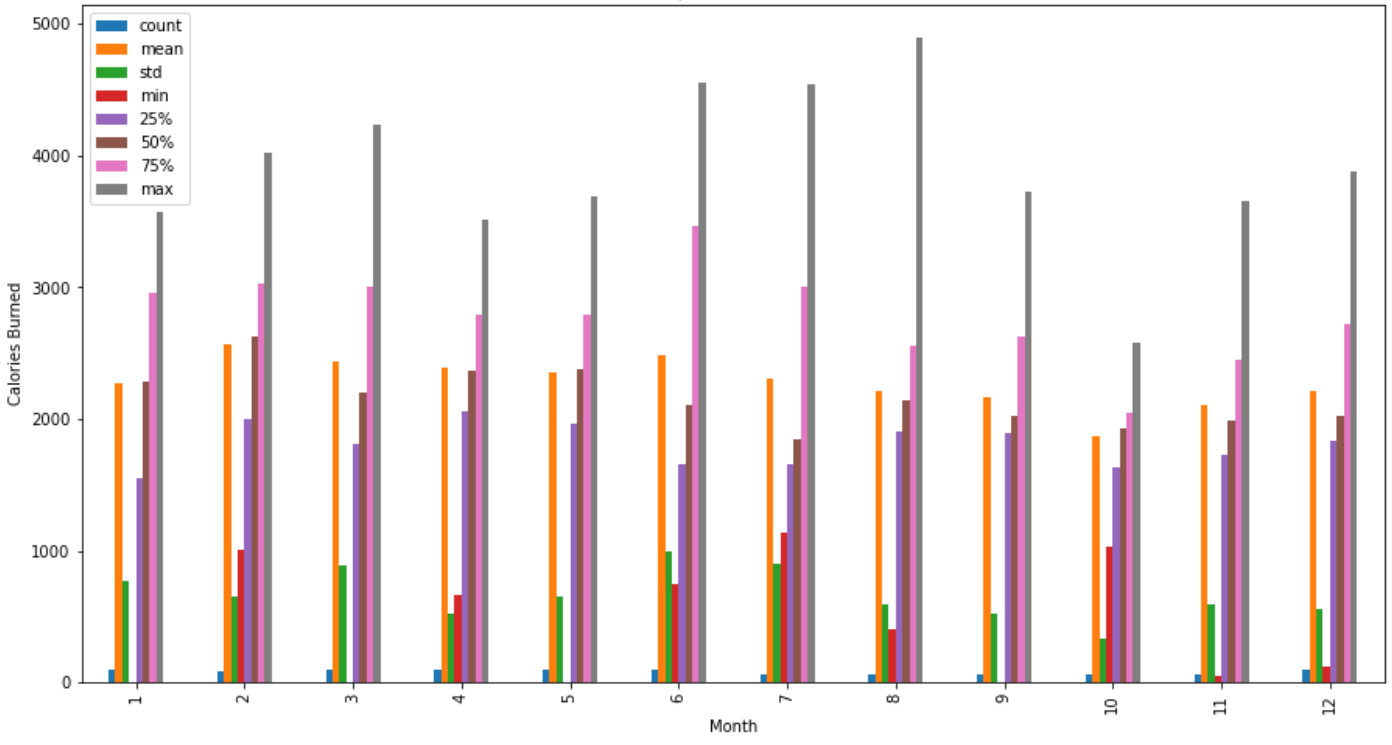
EXPLORATORY ANALYSIS

FIGURE 1: Calories Burned from 2016 to 2019

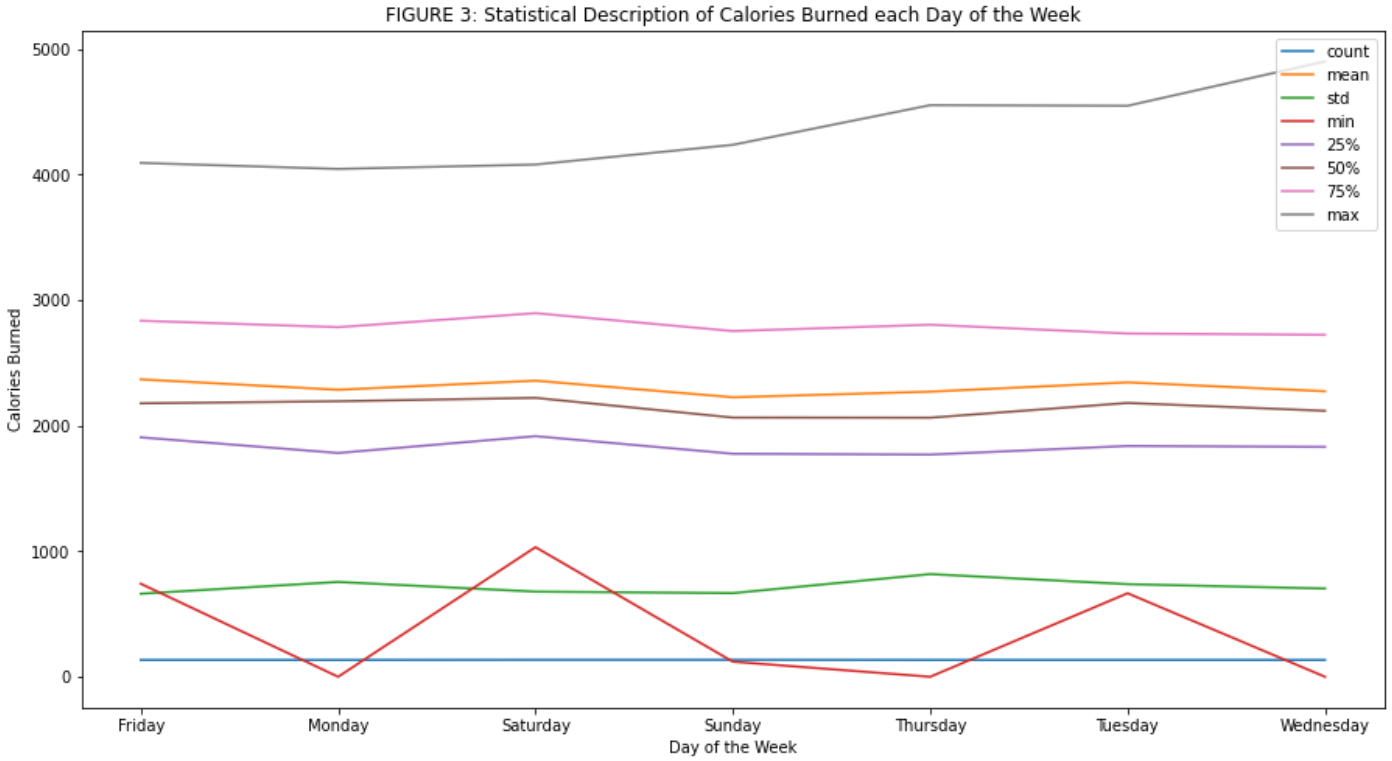


From Figure1 we see that the July and August month has the frequency of most calories burned. This could be due to people getting more time to workout or go to gym or walk due to it being summer vacation time in India.

FIGURE 2: Statistical Description of Calories Burned in each month



From figure 2, we can infer that the maximum calories burned by a person in during the August month and the least is during the month of January and September. The mean calories burned by a person is the highest in the month of February while the least is during the month of October. These variations purely depend on the person's interest to workout.



From Figure 3 we can infer that the maximum calories burned in on Wednesday while the least calories burned that is 0 calories burned was on Mondays, Thursdays and Wednesdays. These variations purely depend on the person's interest to workout.

The Figure 4 shows that, most of the time a good amount of calories were burned and the calories burned has a linear positive relationship with the total number of steps taken and the total distance covered by a person. This is

true practically since the more steps one covers and the more distance a person travels burns more calories than being sedentary.

Any analysis requires us to answer vital questions about trend and relationship found in the data, which would in turn help in building a business model or give a policy suggestion or change algorithm for better marketing. Here we study mainly the following things:

1. The days of the week the users were found to the most and the least active based on the number of steps taken.

2. Days when they use the devices the most often.

3. How much low, moderate, high device usage a person does

FIGURE 5: Mean Total Steps taken each Day of the Week

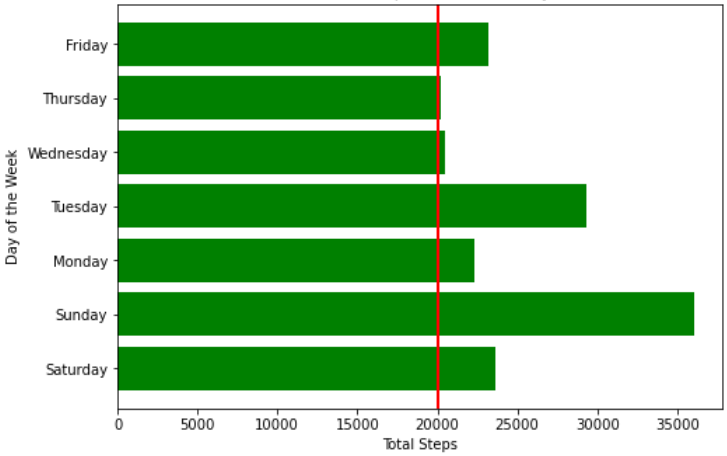
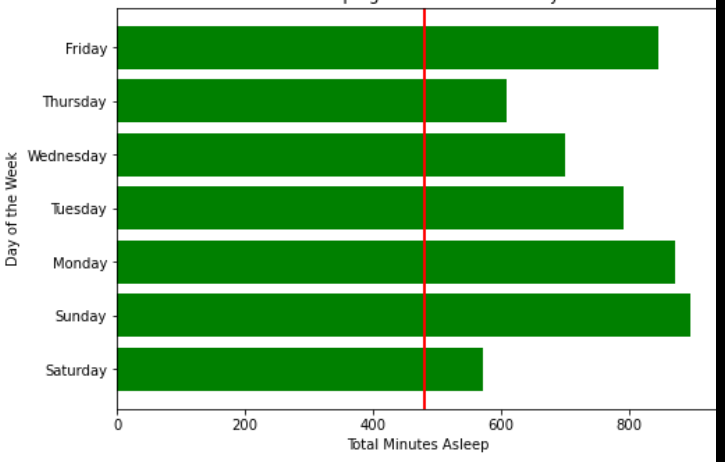


FIGURE 6: Mean sleeping time taken each Day of the Week



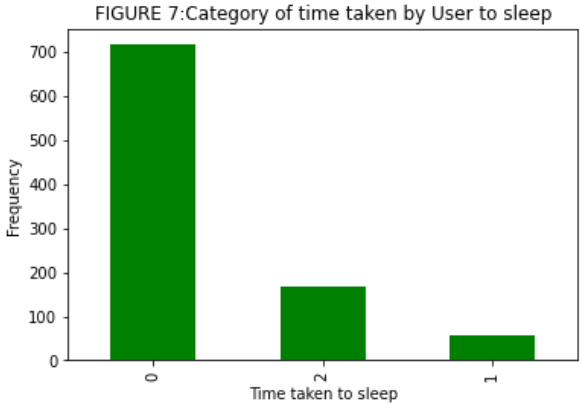
From Figure 5 we can infer that:

1. Users are most active on Sundays while least active on Thursday
2. A study conducted in 2011 revealed that healthy adults can easily take between 4,000-18,000 steps per day. It is recommended that most adults should aim for 20,000 steps per day. In this the user takes more than the recommended steps every day.

From Figure 6 we can infer that:

1. Users have slept most on Sundays followed by Mondays.
2. It is recommended that most adults should aim for 7-8 hours' sleep per day. In this the user takes more than the recommended steps every day.

FIGURE 7:Category of time taken by User to sleep



From Figure 7 we can infer that:

1. Quite often, the person takes a lot of time to fall asleep. This means that the time taken to fall asleep being greater than 3 hours once in the bed has the highest frequency compared to moderate amount that is between around 2-3 hours and for it's the least for less time taken to fall asleep that is falling asleep within 1 and ½ hours.

2. This could be due to a lot of reasons. In this technology age we are quite dependent on social media and our mobile phones. We spend hours on mobile before falling asleep either watching movies, or being in social media or reading e-books. Such reasons have lead us to increase the time taken to fall asleep once in bed.

FIGURE 8:No. of times users logged in app across the week

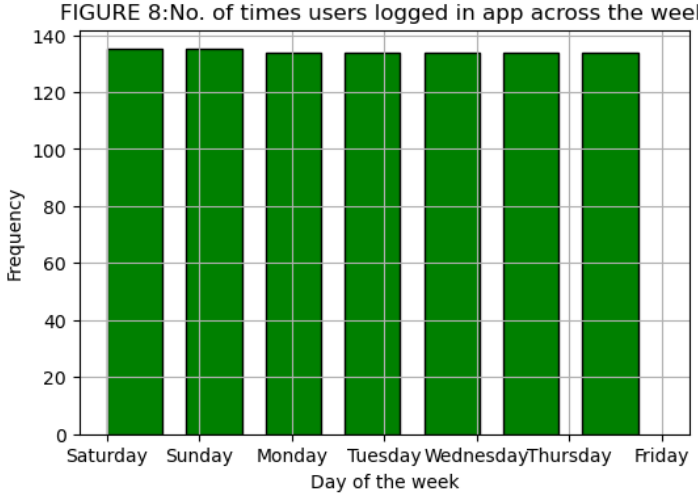
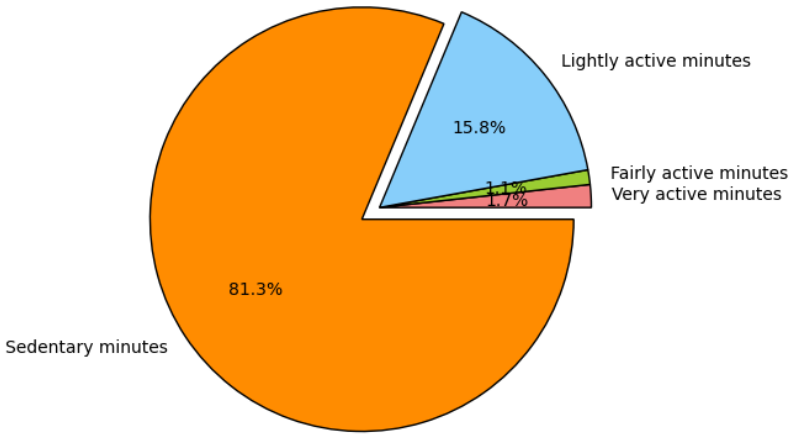


Figure 8 shows that all the days of the week shows an equal interest by the FitBit user to log onto the fitness app connected to the smartwatch. This could be due to keeping a regular track of one's fitness and health.

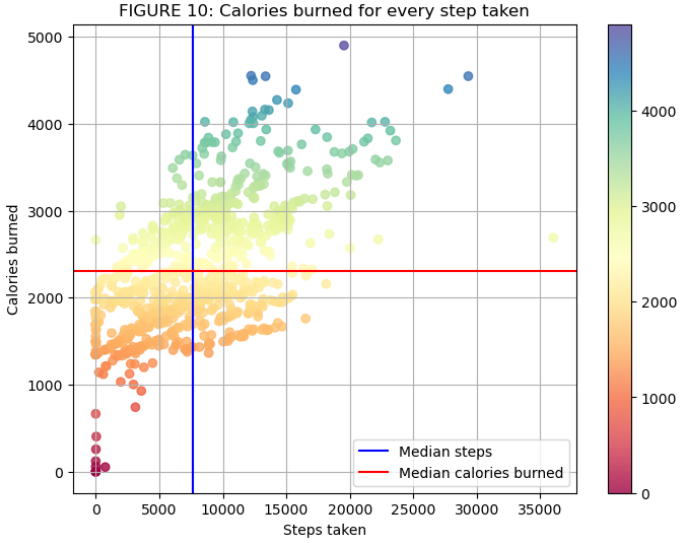
FIGURE 9: Percentage of Activity in Minutes



As seen from the pie chart in Figure 9 we can infer that:

1. Sedentary minutes takes the biggest slice at 81.3%. This indicates that users are using the FitBit app to log daily activities such as daily commute, inactive movements (moving from one spot to another) or running errands.
2. App is rarely being used to track fitness (ie. running) as per the minor percentage of fairly active activity (1.1%) and very active activity (1.7%). This is highly discouraging as FitBit app was developed to encourage fitness.

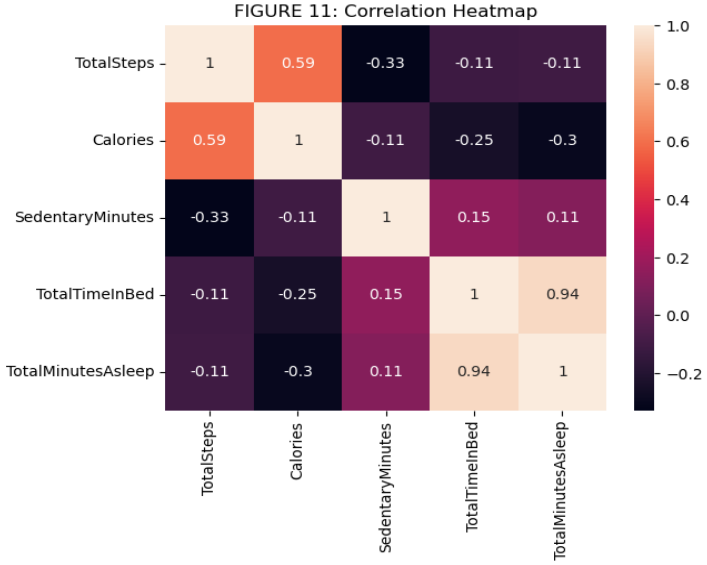
FIGURE 10: Calories burned for every step taken



From the scatter plot in Figure 10, we discovered that:

1. There is a positive correlation between calories burned and steps taken.
2. We observed that intensity of calories burned increase when users are at the range of > 0 to 15,000 steps with calories burn rate cooling down from 15,000 steps onwards.
3. Noted a few outliers:
   Zero steps with zero to minimal calories burned.
   1 observation of > 35,000 steps with < 3,000 calories burned.
   Deduced that outliers could be due to natural variation of data, change in user's usage or errors in data collection (ie. miscalculations, data contamination or human error).

FIGURE 11: Correlation Heatmap

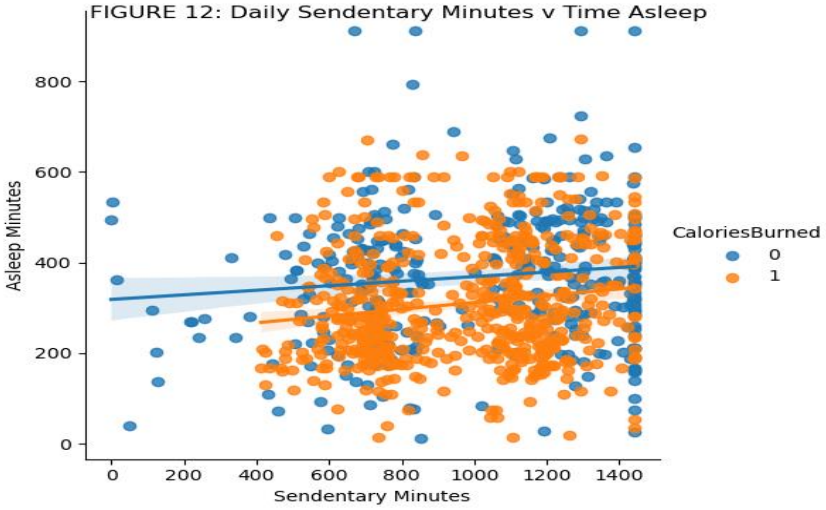


	TotalSteps	Calories	SedentaryMinutes	TotalTimeInBed	TotalMinutesAsleep
TotalSteps	1.000000	0.591568	-0.327484	-0.113992	-0.109648
Calories	0.591568	1.000000	-0.106973	-0.245805	-0.299968
SedentaryMinutes	-0.327484	-0.106973	1.000000	0.151073	0.113810
TotalTimeInBed	-0.113992	-0.245805	0.151073	1.000000	0.935367
TotalMinutesAsleep	-0.109648	-0.299968	0.113810	0.935367	1.000000

The above Figure shows the positive and the negative correlation between the variables. We can see a high positive correlation between the number of steps taken and the calories that were burned. We also see that there's a negative correlation between the Sedentary minutes that is the minutes spend without doing anything, and with the total time in bed and the total minutes asleep.

We thus can say, have a sedentary lifestyle without doing any activities will lead to an unhealthy lifestyle. Instead of losing calories we will gain them more (depends on metabolism) and our weight will increase. With more physical activity we burn more calories and this is proven here.

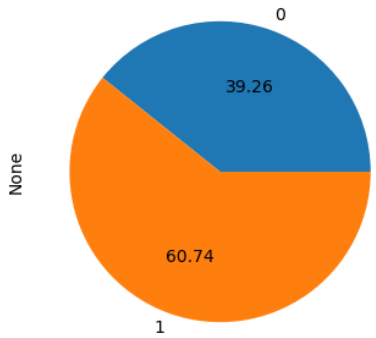
FIGURE 12: Daily Sendentary Minutes v Time Asleep



From this figure 12, we can see that it is a scattered from the line implying that there is a high level of heteroscedasticity among the sedentary minutes and sleeping time. This in a way suggests that the sedentary minutes a person spends is not the same as the sleeping time of the person, they are mutually exclusive events. In 24 hours, cycle, separate time is there were there isn't any physical activity and a separate time is there for the sleeping time.

This in turn can tell us about the person's daily life. These sedentary minutes mainly is due to the office hours spend by the person and the person has a role that makes them sit at one place without any movement (jobs involving information technologies which has 90% of desk jobs which doesn't lead to a person moving around). Hence, we can see a clear distinction.

FIGURE 13: Percentage of Calories Burned above and below Threshhold



From this pie chart in Figure 13, we can say that there is a (almost) 3:2 ratio between high amount of calories burned and negligible amount of calories burned respectively. Hence, SMOTE analysis isn't performed in this case.

MODELLING

Here we are modelling for the category of calories burned that is whether there is a good amount of workout by the person if the calories burned are moderately higher compared to the calories burned by a person without workout which is lower than a threshold of 2000 calories which we loose daily even without workout.

The variables that we considered for this are:

Target Variable (y): Calories Burned-Categorical variable

1: High calories burned (>=2000) 0: Less calories burned(>2000)

Response Variable (X): All are continuous variables

1. TotalSteps
2. TotalDistance
3. TrackerDistance

4. LoggedActivitiesDistance
5. VeryActiveDistance
6. ModeratelyActiveDistance
7. LightActiveDistance
8. SedentaryActiveDistance
9. VeryActiveMinutes
10. FairlyActiveMinutes
11. LightlyActiveMinutes
12. SedentaryMinutes
13. runDistance
14. TotalMinutesAsleep
15. TotalTimeInBed
16. TimeTakenToSleep

The classification models we considered based on the target and response variables taken into consideration are:

1. Logistic Regression
2. K Nearest Neighbours (KNN)
3. Gaussian Naives Bayes
4. Decision Tree
5. Random Forest

We first split the dataset into train and test set with test dataset size being 20% of the whole data. We performed grid search cross validation in 4 of the models and for KNN we performed K-Fold cross validation separately to find the optimum parameters. The following table was obtained:

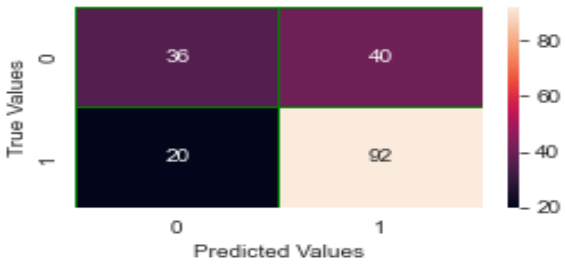
	Model	Best_Score	Best_Params
0	decision_tree	0.700269	{'criterion': 'gini', 'max_depth': 10}
1	random_forest	0.715084	{'max_depth': 1, 'n_estimators': 50}
2	naive_bayes	0.650301	{}
3	logistic_regression	0.680851	{'C': 10}

A. LOGISTIC REGRESSION

We performed the Logistic Regression using the sklearn package in python. The following were the results obtained:

```
Logistic Regression Classification Test Accuracy 0.6808510638297872
RMSE: 0.565

array([[36, 40],
       [20, 92]], dtype=int64)
```

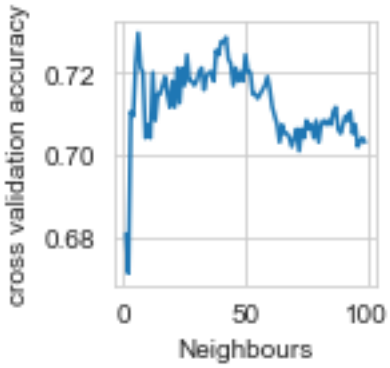


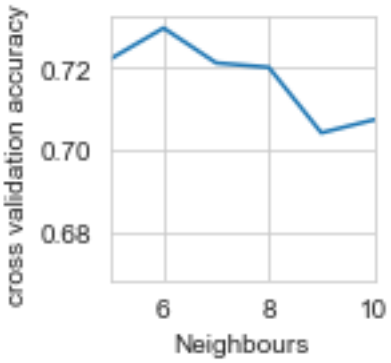
The accuracy of the model obtained is 68.08%. This means that the 68.08% variation in determining the category of calories burned is explained by the 16 response variables chosen. The Root mean square error is 0.565 which is low.

Also, 128 correct predictions are made compared to 60 wrong predictions, showing the chances of correct predictions are more than that of wrong predictions.

B. K NEAREST NEIGHBOURS

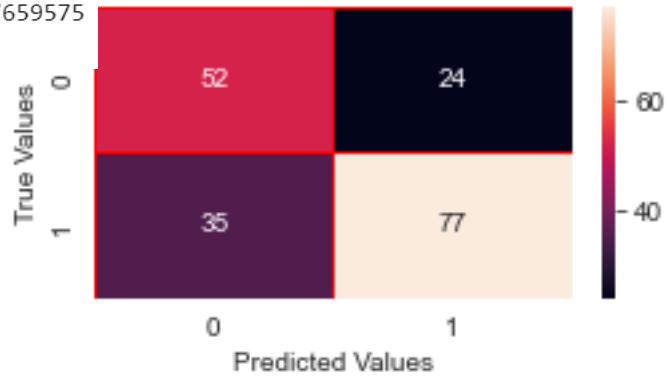
We first performed K-Fold cross validation to find the optimum number of nearest neighbours that should be taken to form the model.





Based in this figure, we can choose the number of nearest neighbours to be considered as 6 since the accuracy is the highest there. Thus, we used sklearn package to form and fit the model. The following were the results obtained:

```
Score for Number of Neighbors = 6: 0.6861702127659575
RMSE: 0.56
```

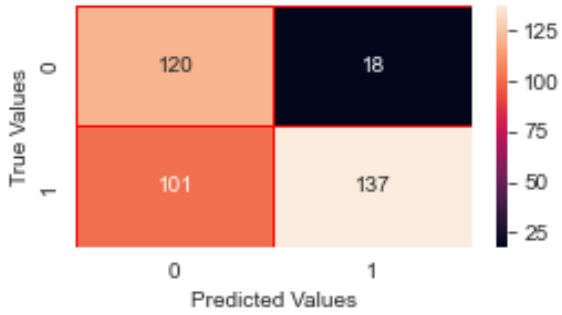


The accuracy of the model obtained is 68.61%, slightly higher than logistic regression. This means that the 68.61% variation in determining the category of calories burned is explained by the 16 response variables chosen. The Root mean square error is 0.560 which is low. Also, 129 correct predictions are made compared to 59 wrong predictions, showing the chances of correct predictions are more than that of wrong predictions.

C. GAUSSIAN NAIVES BAYES CLASSIFICATION

We performed the Naives Bayes Classification using the sklearn package in python. We chose the Gaussian Naives Bayes since all the response variables are of continuous nature. The following were the results obtained:

```
Naive Bayes Classification Score: 0.6835106382978723
RMSE: 0.563
```



The accuracy of the model obtained is 68.35%, slightly higher than logistic regression but lesser than KNN. This means that the 68.35% variation in determining the category of calories burned is explained by the 16 response variables chosen. The Root mean square error is 0.563 which is low. Also, 257 correct predictions are made

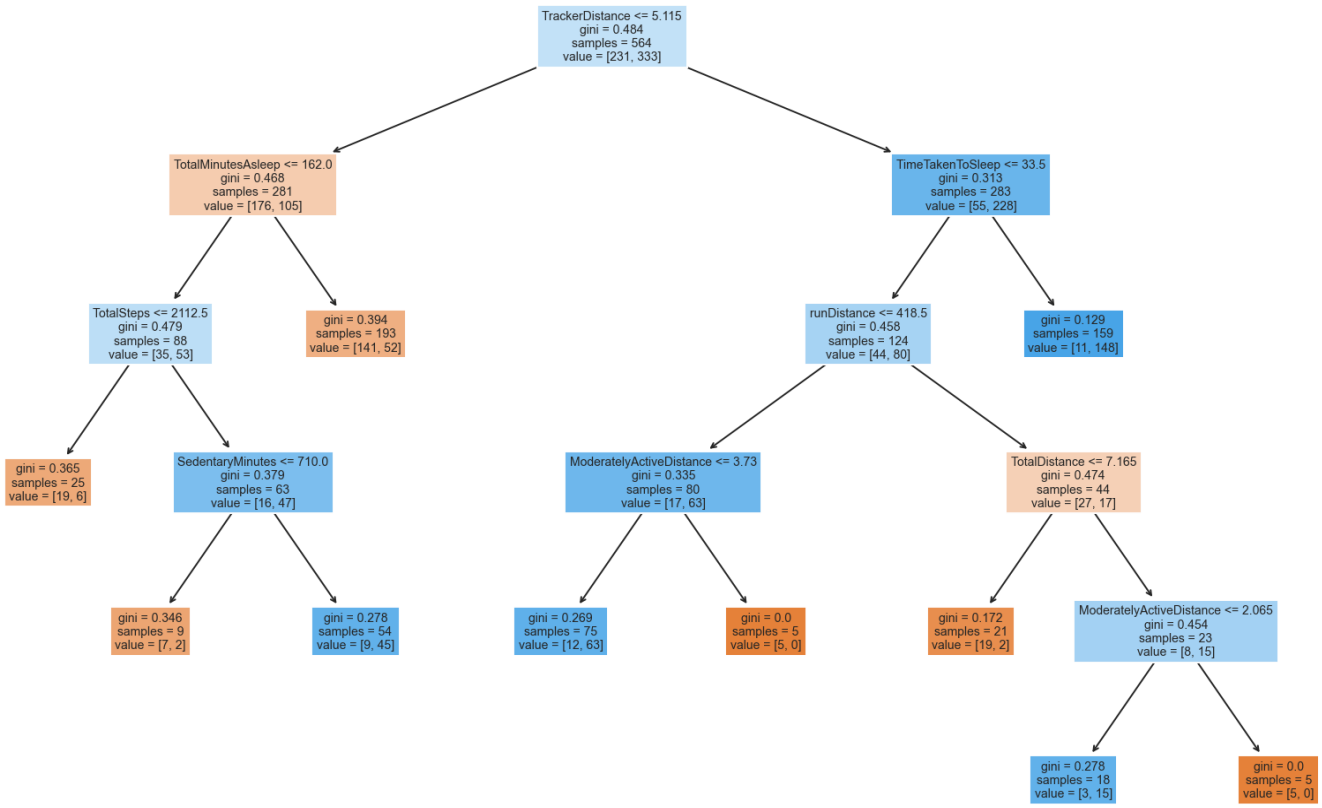
compared to 119 wrong predictions, showing the chances of correct predictions are more than that of wrong predictions.

D. DECISION TREE CLASSIFIER

We got the maximum depth of the tree from Grid search cross validation as 10. But we again ran a loop to get the maximum number of leaf nodes that can be used along with the maximum depth of the tree to get the highest accuracy for the model.

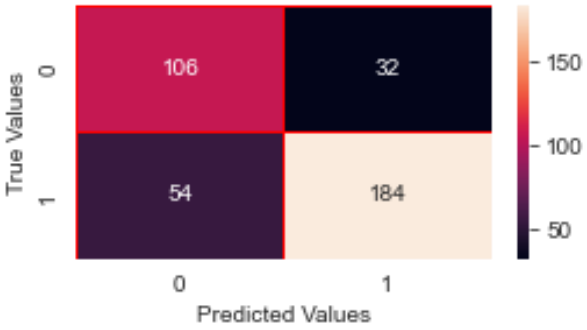
	Model_Accuracy	MAX_DEPTH	MAX LEAF NODES
5	**0.77128**	**10**	**10**
9	0.77128	50	10
13	0.77128	100	10
6	0.75266	10	50
10	0.73936	50	50
14	0.73936	100	50
11	0.73404	50	100
15	0.73404	100	100
0	0.72606	1	2
1	0.72606	1	10
2	0.72606	1	50
3	0.72606	1	100
4	0.72606	10	2
7	0.72606	10	100
8	0.72606	50	2
12	0.72606	100	2

Based on this we created the Decision tree classifier and obtained the following tree and results:



The tree is generated using the Gini Coefficient and has a maximum depth of 10 and maximum number of leaf nodes as 10.

```
Decision Tree Classification Score:  0.7712765957446809
RMSE: 0.478
```

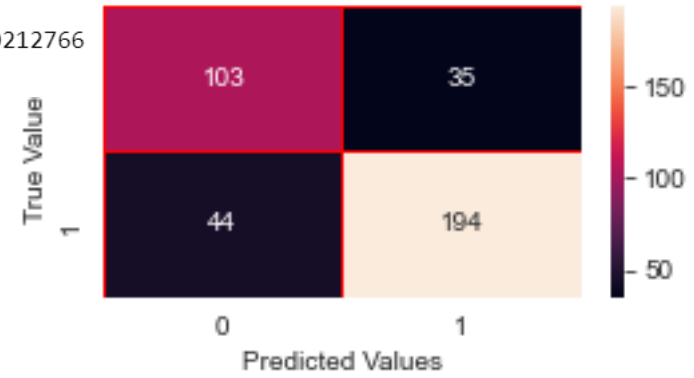


The accuracy of the model obtained is 77.12%, higher than all the above models. This means that the 77.12% variation in determining the category of calories burned is explained by the 16 response variables chosen. The Root mean square error is 0.563 which is low. Also, 290 correct predictions are made compared to 96 wrong predictions, showing the chances of correct predictions are more than that of wrong predictions.

E. RANDOM FOREST CLASSIFIER

We used Random Forest classifier in the sklearn package to perform the modelling. As detected in the grid search cv, we got the optimum number of estimators to be 50 to get the best accuracy for the model. After fitting the model, the following were the results obtained:

```
Random Forest Classification Score:  0.7898936170212766
RMSE: 0.458
```



The accuracy of the model obtained is 78.98%, higher than all the above models. This means that the 78.98% variation in determining the category of calories burned is explained by the 16 response variables chosen. The Root mean square error is 0.563 which is low. Also, 297 correct predictions are made compared to 79 wrong predictions, showing the chances of correct predictions are more than that of wrong predictions.

MODEL COMPARISON

We compared the model based on the accuracy score, the root means square error and the Receiver Operating Characteristic (ROC) Curve. We plotted it for a better analysis.

	Models_Used	Accuracy_Score	RMSE
0	Logistic Reg.	0.680851	0.565000
1	KNN	0.686170	0.560000
2	Naive Bayes	0.683511	0.563000
3	Decision Tree	0.771277	0.478000
4	Random Forest	0.789894	0.458000

FIGURE 14: Comaprison of Accuracy between Models

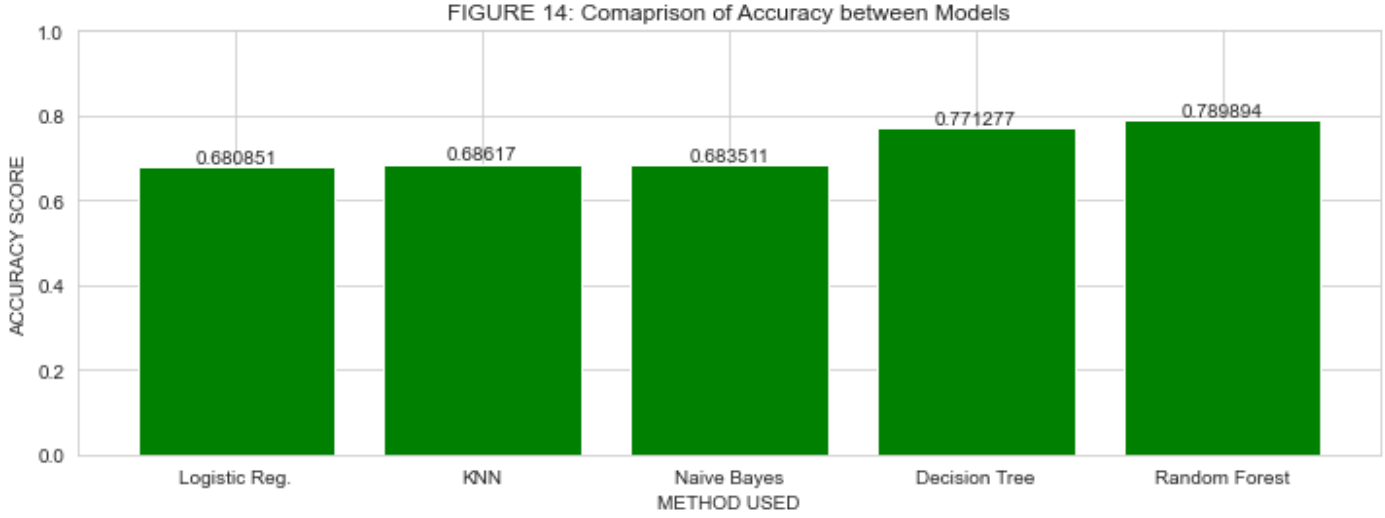


FIGURE 15: Comaprison of Accuracy between Models

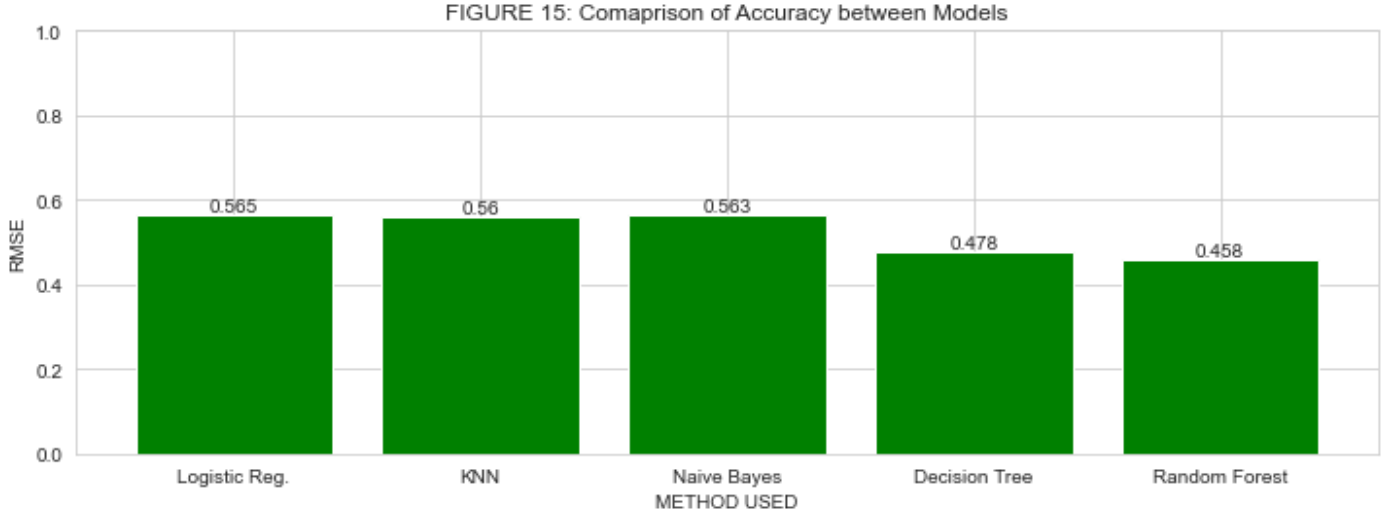
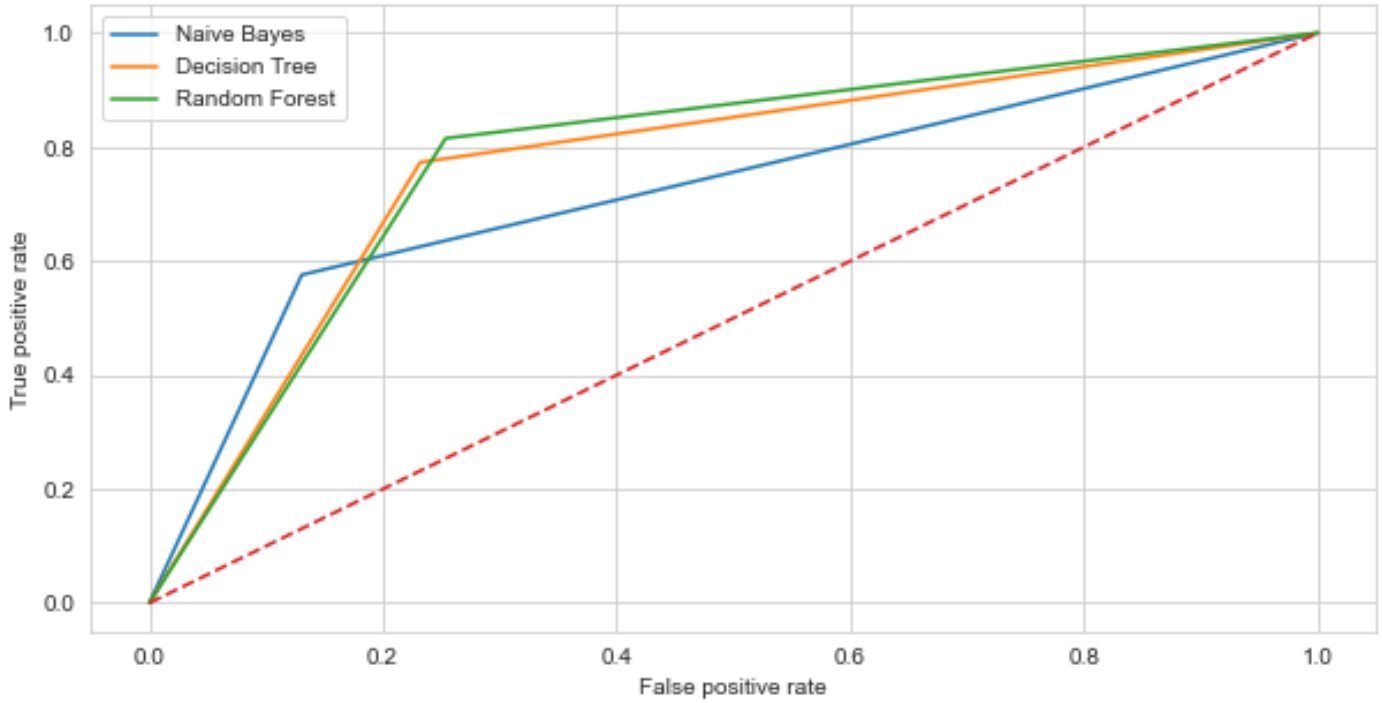


FIGURE 16: Receiver Operating Characterstic Curve



From this we can infer that the best model for this problem is Random Forest classifier since it's accuracy score is the highest among all and the RMSE value is also the least among all. The ROC curve plot also suggests the same inference. The second best model is the Decision Tree classifier and the least accurate model for the problem is logistic regression.

CONCLUSION

In this age of technology and consciousness of body image, a lot has changed. We determine weight loss that is burning out the calories as very important factor to keep our health on check and make us fit. This century has been led by jobs that most of the time makes us sit in front of desktops and have a sedentary lifestyle. These sedentary minutes mainly is due to the office hours spend by the person and the person has a role that makes them sit at one place without any movement (jobs involving information technologies which has 90% of desk jobs which doesn't lead to a person moving around).

To keep everything on track we hit gyms or go for a run or walk around. The more the distance and steps covered the more calories being burned. Our bodies metabolism generally dictates that 2000 calories are burned in a day without much activity. But to keep our weight in check, we have to burn more than that. Hence, the two categories are created.

Based on these two categories various machine learning classification models were run to form the best model which came out to be a Random Forest model. This model will help a person to be classified if the he or she is working out or not based on the category of calories burned.

The limitation of this study is that each person has different metabolism which hasn't been taken into account.

Our daily life revolves around keeping our health in check. Thus, its important to have physical activities daily to burn out calories gained from being sedentary for hours in our desk jobs or in classroom.

REFERENCES

1. https://www.datacamp.com/tutorial/understanding-logistic-regression-python
2. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
3. Lucio, N. D., Salazar, E. V., Figueroa, I. A., Gamez, J. L., Russell, R. D., & Funk, M. D. (2018). Accuracy of Fitbit Charge 2 at estimating VO2max, calories, and steps on a treadmill. In *International Journal of Exercise Science: Conference Proceedings* (Vol. 2, No. 10, p. 11).
4. Bender, C. G., Hoffstot, J. C., Combs, B. T., Hooshangi, S., & Cappos, J. (2017, March). Measuring the fitness of fitness trackers. In *2017 IEEE Sensors Applications Symposium (SAS)* (pp. 1-6). IEEE.
5. Wang, C., Lizardo, O., & Hachen, D. S. (2021). Using Fitbit data to examine factors that affect daily activity levels of college students. *PloS one*, *16*(1), e0244747.
6. www.kaggle.com