**UNIVERSITY OF SUNDERLAND**

**ASSIGNMENT COVERSHEET**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student ID: 219572048** | | **Student Name/ Names of all group members:**  **Abhishek Yadav** | |
| **Programme: Computer System`s Engineering** | | **Module Code and Name: CET313 Artificial Intelligence** | |
| **Module Leader/ Module Tutor: Anil Bhujel** | | **Due Date: 01/06/2023**  **Hand in Date: 01/06/2023** | |
| **Assessment Title: Intelligent Prototype Development** | | | |
| **Learning Outcomes Assessed: (number *as appropriate*)** | | | |
| |  |  | | --- | --- | |  | **Mark** | | **Areas for Commendation** |  | | **Areas for Improvement** |  | | **General Comments** |  | | | | |
| **Assessor Signature:** | **Overall mark (subject** to ratification by the assessment board**)** | | **Moderator Signature** |

*I confirm that in submitting this assignment that I have read, understood and adhered to the University’s Rules and procedures governing infringements of Assessment Regulations.*

**PRINT Student Name**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Faculty Stamp *(date/time)*

**Student Signature**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Module Code and Name**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Name of Module Tutor** : \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Introduction**

AI refers to the simulation of human intelligence in machines that have been designed to think and act like humans. The author in (Dobrev, 2005), defined AI as a program which can learn and interact with its environment in a way similar to how humans do so also the program should be able to handle this situation no worse than a human would, whether it be natural or artificial. These intelligent machines can be trained to perform various tasks by processing large amounts of data and recognizing patterns within the data. AI has the ability to learn and adapt to new situations, which makes it a powerful tool for solving problems that are too complex for humans to tackle.

In order to achieve a specific goal, AI researchers and practitioners rely on a number of different techniques and approaches, including knowledge representation and reasoning, search, natural language processing, planning, and machine learning. Almost all of these approaches were imparted to us throughout this module through task activities and lab exercises. Knowledge representation and reasoning involves development of algorithms that enable intelligent systems to make logical inferences based on incomplete or uncertain information. Search algorithms are used to explore the space of possible solutions to find the best one. Natural language processing (NLP) is a subfield of artificial intelligence that focuses on how to enable computers to understand, interpret, and generate human language. It has a wide range of applications, including language translation, text summarization, text classification, and information extraction and as an important component of many AI systems, to communicate with humans in a natural and intuitive way. Machine learning algorithms are used to analyze large datasets in order to identify patterns and relationships that can be used to make predictions or decisions. Some of the key techniques used in machine learning include supervised learning, unsupervised learning, and reinforcement learning. Artificial neural networks are a type of machine learning algorithm that is inspired by the structure and function of the human brain and are composed of interconnected nodes, or "neurons," which are used to process and transmit information.

The primary goal of this project is to use machine learning to address a real-world issue. The project, Calories burnt prediction, is based on supervised Machine learning technique, where Python programming and its libraries are used. The problem domain of this project is about unhealthy lifestyle of humans. Modern day people lead stressful and busy lives which obviously alters the eating habit and most of them rely on junk and fast foods, which lead to an unhealthy way of life. Thus, one of the major problems in contemporary life is an unhealthy lifestyle and the Calorie burnt prediction prototype addresses this problem by providing users, the amount of calorie they burn everyday for maintaining a healthy lifestyle. Prediction system are able to learn patterns, handle large amount of data and can be automated which makes them one of the most successful application areas of machine learning.

**Section1: Prototype Identification and Planning**

**Section 1.1: Literature Review on Prototype Identification**

Accurate calorie burnt prediction can help individuals track their caloric intake and expenditure, allowing them to make informed decisions about their diet and exercise routine in order to achieve specific health and fitness goals. Several studies have relied on the machine learning techniques like linear regression and decision trees, to model the relationship between physical activity and calorie expenditure. Besides these, complex models, such as neural networks are used to capture the non-linear relationships between these variables. MIT had worked on this during prior days, and the authors of MIT in (Tapia, 2008), propose a system that uses machine learning algorithms to analyze data from a wearable sensor and make predictions about the user's activity level and energy expenditure. The system uses a variety of sensors, including accelerometers and gyroscopes, to collect data about the user's movements. The data is then fed into machine learning algorithms that are trained to recognize different activities, such as walking, running, and climbing stairs. They propose using machine learning to predict energy expenditure based on the user's activity level and other factors such as their age, weight, and height.

(First Teddy Surya Gunawan, 2018) discusses the use of a generalized regression neural network (GRNN) to predict food intake calories. The authors propose using a GRNN to analyze data collected from wearable sensors and other sources in order to make predictions about the number of calories a person consumes. The GRNN is a type of neural network that is well-suited for regression tasks, such as predicting numerical values. It is trained using a set of input-output pairs, and can then be used to make predictions for new input data. The authors describe how they used a GRNN to analyze data collected from wearable sensors and other sources in order to predict food intake calories. They conducted experiments to evaluate the performance of the proposed system and found that it was able to accurately predict food intake calories.

The research by (Megha Rathi, 2018), used data mining techniques to investigate health fitness and heart status. The authors propose using a variety of different techniques to analyze data collected from wearable sensors and other sources, including decision trees, association rules, clustering, neural networks, and support vector machines (SVMs). They used linear discrimination analysis and k nearest neighbors for classification of data and developed a dataset using the basal metabolic rate formula. Their study showed that for losing 1 pound around 500 calories should be shaved off daily.

Recently, the authors in (Marte Nipas, 2022) claims to have been able to predict the calories burnt with a high accuracy rate using the Random Forest Regression model. To choose the optimal model for the investigation, K-fold validation was used for model training and testing. The study's best model, Random Forest regression, has an accuracy of 95.77%, according to the average accuracy calculation. The researchers in (Marte Nipas, 2022) advised that some of the parameters be changed in order to observe if the score significantly changes for the researchers who might use this study as reference. Additionally, they advised testing out existing regression methods that might offer a better match than the ones employed in this study.

Likewise, numerous regression models such as Linear and logistic regression are used to predict the value of calories burnt accurately but logistic model is generally used to predict a binary outcome. I decided to go with XGBoost Regressor model as it is highly efficient and accurate machine learning algorithm that has been widely used in a variety of prediction tasks. It is particularly well-suited for handling large datasets and can handle missing values and other types of data noise. Also, the authors in (Audace B. K. Didavi, 2021) claims that the machine learning algorithms like XGBoost, Random Forest and Decision Tree perform well in predicting accurate values and that the XGBoost has the best performance among them with least mean square error value. The prototype for our project uses a similar classification of steps which involve selection of datasets, training the XGBoost Regression algorithm with training data, record and evaluate the prediction results and let the ML module predict the result from the new input data.

**Section 1.2: Reflection on the Prototype Identification**

With the rise of wearable fitness devices and mobile apps that track physical activity, there is a growing need for accurate and reliable methods of estimating the number of calories an individual burns during various types of physical activity. Based on the prior work, relatable methodology used in calorie burnt prediction projects can be observed in the research. These research highlights that the study used reliable data with quantitative predictive approaches to complete a prediction on calorie burnt. XGBoost is often considered to be a good choice for prediction systems because it is efficient, accurate, fast, scalable, and flexible. Derived from the research, the algorithm chosen for this project is XGBoost regression, which is commonly used algorithm for predictive systems.

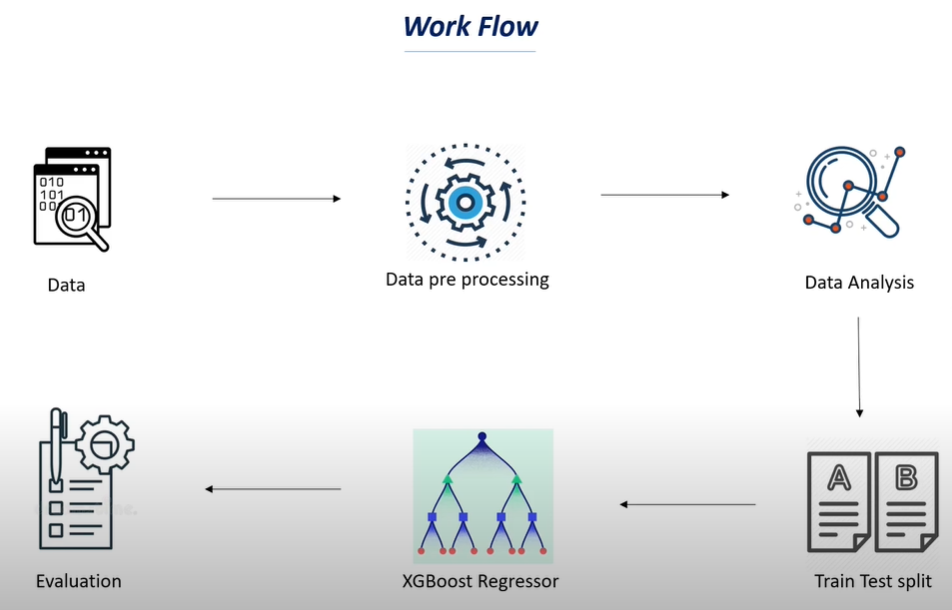
Aside from the literature review, the knowledge I gained during the course of the project was mostly in python programming and different libraries used in classification, which helped me develop the project. I gained a basic understanding of programming concepts in the initial weeks, including variables, classes, dictionaries, and tuples used in the management of big data frames. In addition, I was able to grasp the idea of data frames, Pandas, and NumPy, which the library employs to handle multi-dimensional data frames and analyze data. Later, I became familiar with the visual tools, which enabled me to visualize the data that the program had retrieved. I made use of particular technologies like seaborn and matplotlib. During the mid of the module, I learnt about functions and classes used to train and evaluate machine learning model like logistic regression for prediction of a continuous outcome and got to know about resources like Kaggle and other machine learning repositories to look for datasets. In the project I accomplished the calorie burnt prediction with high accuracy using XGBoost regressor that helps overcome the challenges created by unhealthy lifestyle for users by predicting the amount of calorie they burn.

**Section 2: Development**

The prototype is designed to accurately predict the calorie burnt by the user. The approaches are classified and it is being built in python. The dataset’s CSV file was fetched from the Kaggle repository which was then processed and visualized and multiple libraries were employed to complete the project. The code and the respective procedures are comprehensively written below:

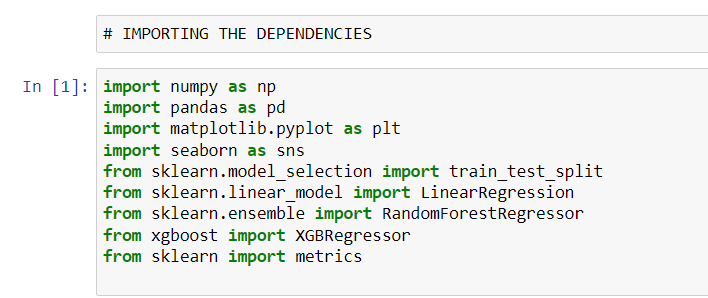
**Work flow**

The work flow of the Calorie burnt prediction prototype is shown in the Figure 1.

****

**Figure 1: Work Flow of calorie burnt prediction**

**Using libraries/dependencies**

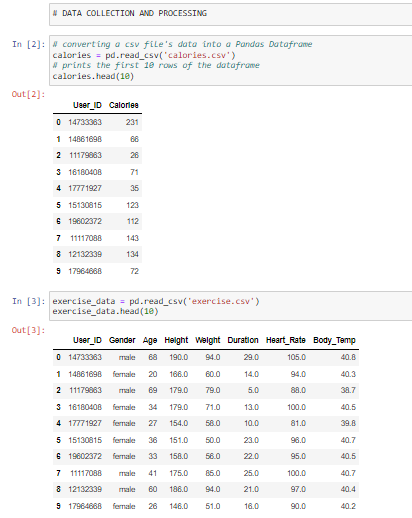
****

**Figure 2: Importing the dependencies**

Dependencies are the packages or modules that are required by another packages or modules in order to function properly. Importing the dependencies is the first step in any Python program. In the Figure 1, all the necessary dependencies for data collection, visualization, ML model training, prediction and evaluation have been imported. NumPy, which is imported here as ‘np’, is a software library in Python that is used for numerical computing. It provides a variety of functions and methods for working with large and complex arrays of numerical data. Pandas, which is imported here as ‘pd’, is a software library in Python that is used for data manipulation and analysis. It provides a variety of data structures and data analysis tools that make it easy to work with large and complex datasets. Matplotlib is a data visualization library in Python that is used to create static, animated, and interactive visualizations in a variety of formats. It provides a range of functions and methods for creating a wide variety of plots and charts, including scatter plots, line plots, bar plots, and more. Seaborn is a data visualization library in Python that is built on top of Matplotlib. It provides a higher-level interface for creating statistical plots and charts, and is particularly well-suited for working with dataframes and statistical data. Matplotlib and Seaborn are very popular libraries in the Python data science ecosystem and are often used in conjunction with other libraries such as NumPy and pandas for data analysis and visualization tasks. The train\_test\_split function that is used to split a dataset into a training set and a testing set, is imported from the model\_selection module in the scikit-learn library. Additionally, imported machine learning models include XGBRegressor, Linear Regression and RandomForestRegressor. Finally, metrics that are imported from Sklearn are used to evaluate the model, determining whether it is operating effectively or not.

**Data collection and processing**

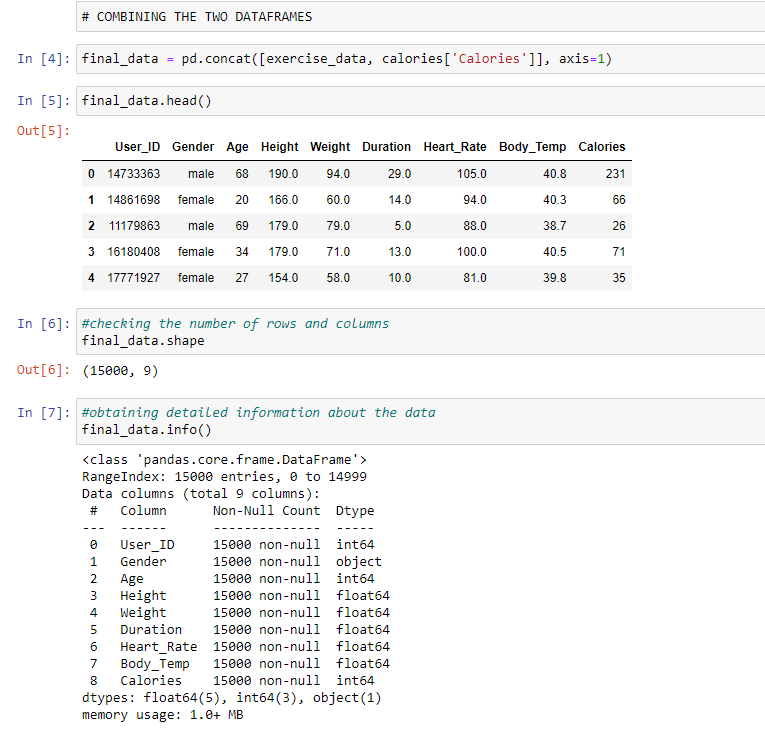
I found a reliable dataset on Kaggle comprising of two datasets namely calories and exercise to evaluate the calorie burnt. In Figure 3, I loaded all the data from a csv file, which is in the form of a comma (,) separated value and cannot be analyzed. Consequently, we set the csv file to the pandas data frame in order to examine the data. After that, using the .head() function the topmost 10 rows of the data frame is printed to check and compare the datasets.



**Figure 3: Data collection and processing**

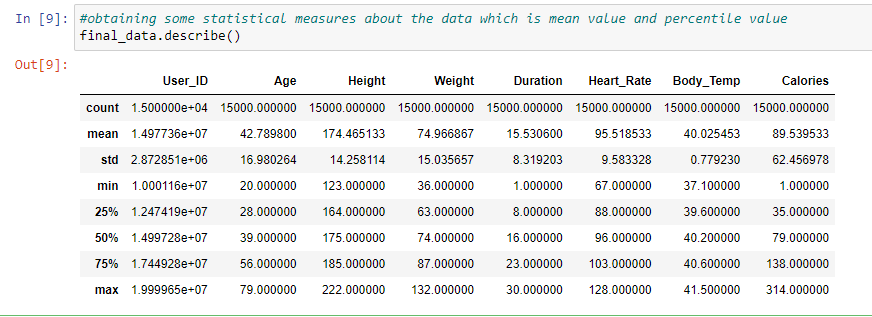
**Organizing the data**

The two datasets are not useful when used separately, so we need to combine these datasets to get more information about the data. The shape function is used to check the number of rows and columns in the dataset. The Figure 4 shows that there is 15000 rows and 9 columns in the final dataset which is a reliable dataset size. Furthermore, as we can see, the use of info function to display the datatype, which includes the user ID, age, height, weight, duration, heart rate, calories and gender, here body temperature in the form of a float, which has a decimal value.



**Figure 4: Organizing the data**

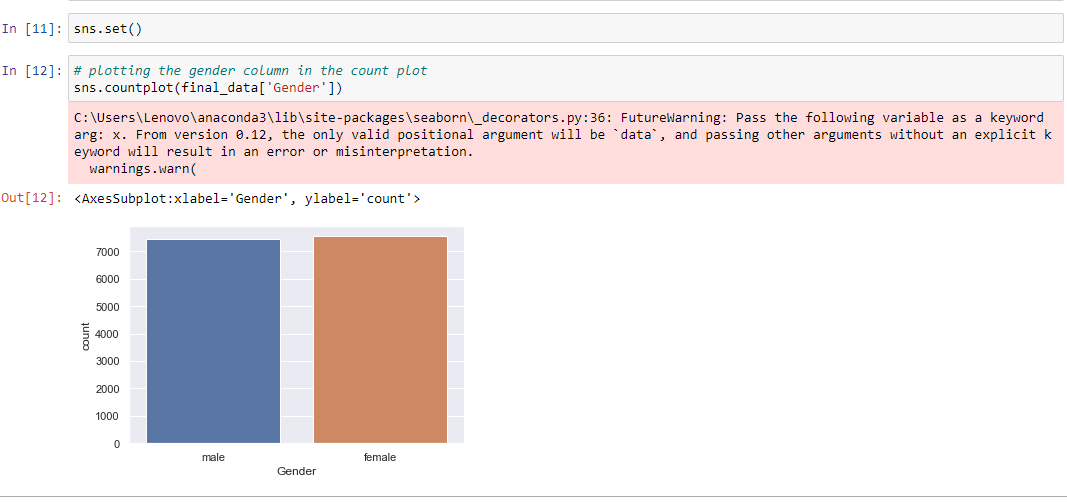
**Data Analysis and Visualization**

****

**Figure 5: Getting statistical measures using the describe function**

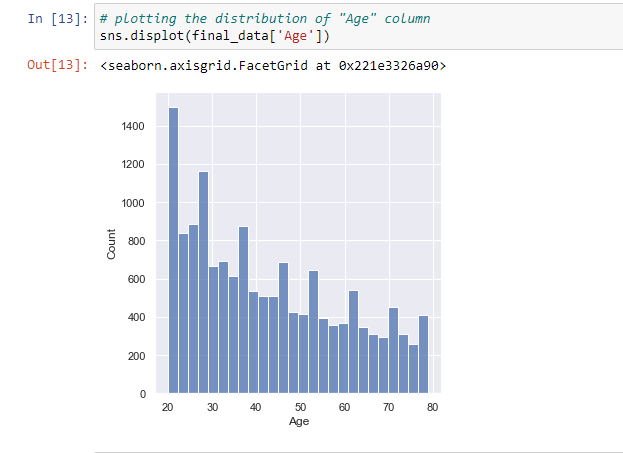
In Figure 5, the describe function is used to get statistical measures about the data which displays the different statistical measure values like mean, standard deviation and minimum and maximum values for each column in the dataset.

Next, I tried to plot different columns separately to visualize the dataset and analyze them.



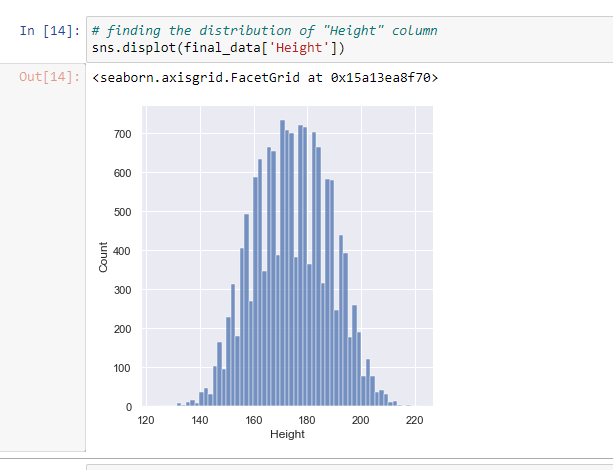
**Figure 6: Plotting the gender column**

For gender being and example of categorical column, the countplot is used to plot this column as shown in Figure 6. The figure shows the equal distribution of fame and female which confirms the uniformity of dataset.



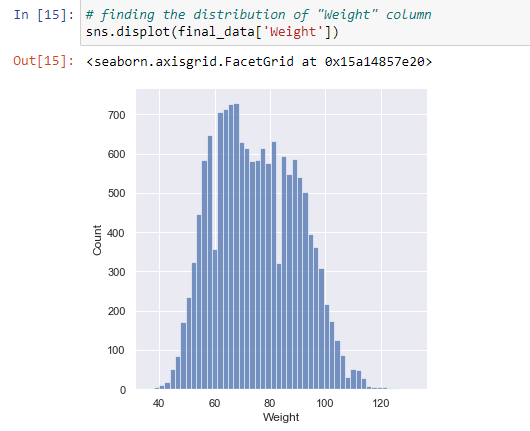
**Figure 7: Plotting the distribution of age column**

The figure 7 shows the use of displot function to check the distribution of age column and it showcases that there is dense distribution of age column between the age of 20 and 30.

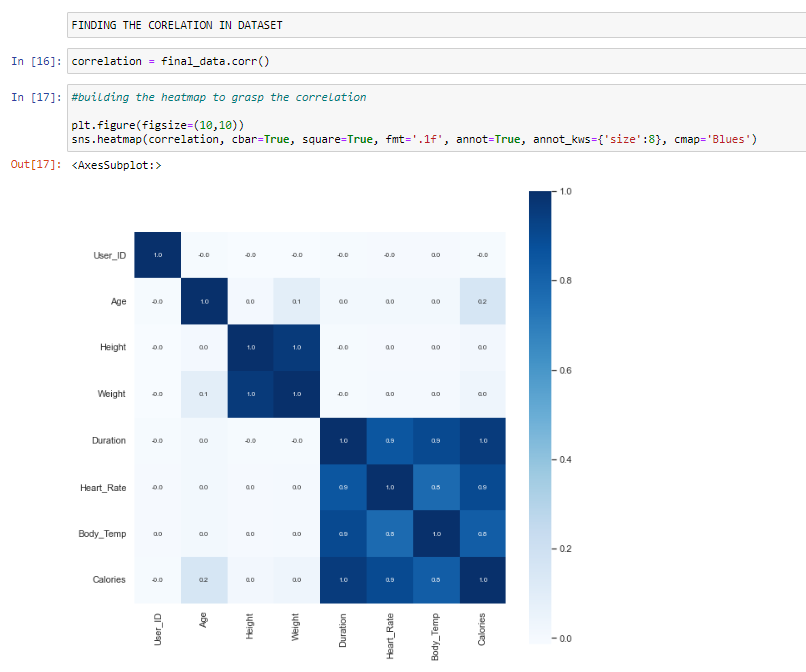


**Figure 8: Plotting the distribution of height column**

In Figure 8, the distribution of height column is shown and the height seems to me more concentrated between 160 and 190. Similarly, in Figure 9, the distribution of weight column is shown through the use of displot function of sns library.



**Figure 9: Plotting the distribution of weight table**

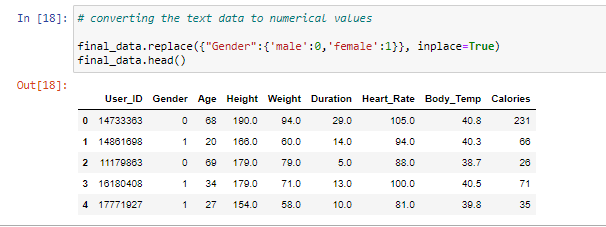


**Figure 10: Heatmap to understand correlation**

In the Figure 10, a heatmap is constructed to visualize the correlation value to identify which columns are positively correlated and which ones are not. The two types of correlation are identified by larger values when the two columns are positively correlated and vice versa. The figure shows that the duration, heart rate and body temperature are positively correlated to the calories value and I shall use these to find out the value of calories burnt in this project.

**Converting text data to numeric values**

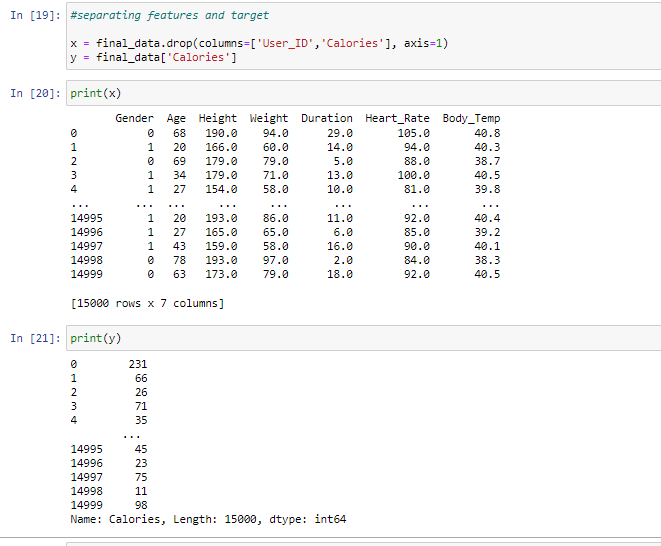
In the Figure 11, the values from the gender column are taken and replaced by 0 and 1, respectively for male and female. The parameter inplace=True is added to ensure the permanent change in the dataset.



**Figure 11: Converting text data to numeric values**

**Separating features and target**

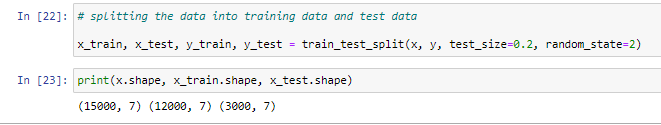
In the Figure 12, the final dataset is separated into features and target before the Train Test Split. The feature columns which comprise of gender, age, height, weight, heart rate and body temperature is stored in x and the target column which comprises of calories column is stored in y and finally, both the variable are printed to showcase the results.



**Figure 12: Separating features and target**

**Train Test Split**

Next, the data is fragmented into train data and test data. For this, the train\_test\_split function imported from the sklearn model is used and we need to create four arrays which are x\_train, x\_test, y\_train and y\_test. Here, the training data is used for training our machine learning model and test data which is 20% of total final data, would be used to evaluate our model. The test set then tests the model’s predictions based on what it learned from the training set. Lastly, the shape of the training data and the test data is printed as shown in Figure 13.



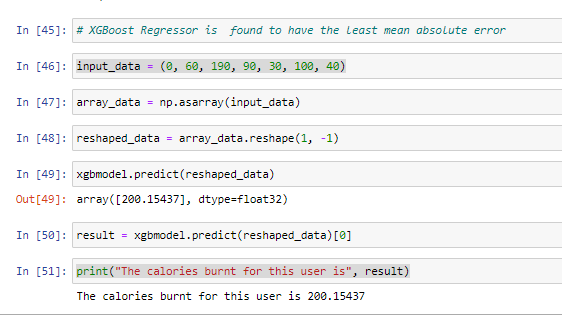
**Figure 13: Train Test split**

**Model training**

In the project, I have used the XGBoost regression model which we had imported at the start. The model is then trained with x\_train and y\_train data. The model uses x\_test data to find out the calorie burnt and then it is compared to y\_test to evaluate the effectiveness of the prototype. The mean absolute error value is found to be 1.5 and the R squared value as 0.99 as shown in Figure 14. Later a predictive system was built over this model to predict the value of calories burnt from the user input features data as shown in Figure 15.



**Figure 14: XGBoost Regressor Model**



**Figure 15: Prediction system using XGBoost Regressor Model**

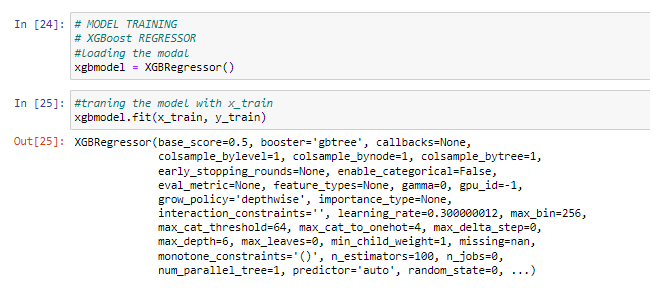
Finally, the input data were given to the XGBoost model and it predicted the amount of calories burnt.

**Section 3: Evaluation**

The prototype helps the user to maintain a healthy lifestyle by accurately predicting the number of calories burnt and let them track their health activities to keep themselves healthy. For the accurate Calorie burnt prediction, I have tested three regression models namely XGBoost Regressor, Linear regressor and RandomForestRegressor, all focusing on predicting the value of continuous outcomes through the usage of various regression algorithms. MAE is a measure of how far the predicted values are from the true values. It is calculated by taking the absolute difference between the predicted value and the true value, and then taking the mean of these differences over all data points. A lower MAE indicates that the model is making predictions that are closer to the true values. R Squared, also known as the coefficient of determination, is a measure of how well the model fits the data. It is calculated by taking the sum of the squared differences between the predicted values and the mean of the true values, and dividing by the sum of the squared differences between the true values and the mean of the true values. An R Squared value of 1 indicates that the model fits the data perfectly, while a value of 0 indicates that the model does not fit the data at all. I have used these two metrics to evaluate the regression algorithms selected in the project.

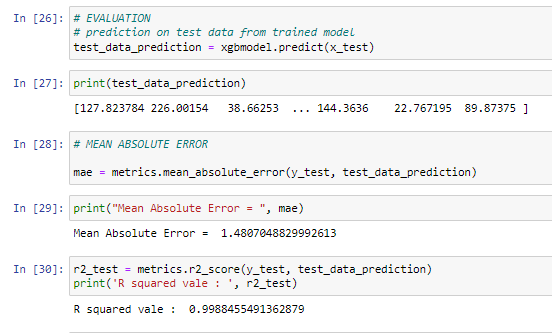
**Testing the XGBoost Regressor**

From the prior research done, I was able to choose this model as this is widely used for regression tasks, such as predicting numerical values. It is an implementation of gradient boosting, which is a method for building models that involves training a sequence of simple models, such as decision trees, and then combining them to create a more powerful model. According to the authors in (Audace B. K. Didavi, 2021), the XGBoost algorithm is known for its efficiency, accuracy, and speed, and has been used to win many machine learning competitions. It is particularly well-suited for handling large datasets and can handle missing values and other types of data noise.



**Figure 16: Loading the XGBoost Regressor Model**

The Figure 16 clearly shows that the XGBRegressor model is loaded first to the set variable. Next part comprises with the training of this model with the training data. The fit function was used to train the x\_train data and y\_train data. When we feed the model with test data it predicts the accurate amount of calorie burnt as it finds the pattern among the data to give the most accurate predictions.

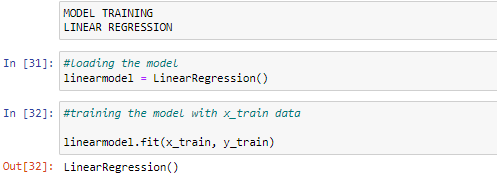


**Figure 17: MAE and R squared value for XGBRegressor**

Here, the model predicted the test data, using the x\_test data and prints down the results of the test. Later, I calculated the mean absolute error and r squared value by comparing the value predicted by the model with the original value i.e. y\_test. The MAE value is 1.4 while the R squared value is 0.99.

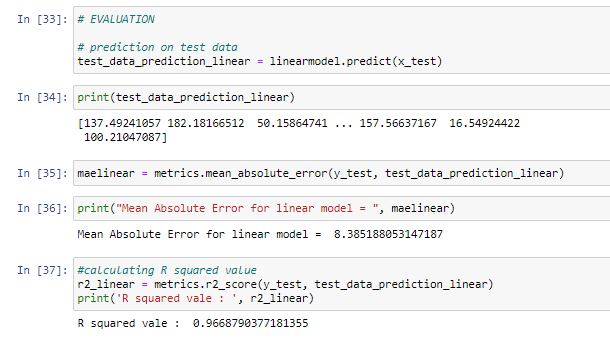
**Testing the Linear Regression Model**

The second machine learning model I approached was the Linear regression model. Linear regression works by finding the line of best fit through a scatterplot of the data. The line of best fit is the line that minimizes the sum of the squared differences between the predicted values and the true values. Once this line has been found, the linear regression model can be used to make predictions for new data by calculating the corresponding value on the line of best fit.



**Figure 18: Loading the linear model**

In the Figure 18, at first the linear model is imported and loaded in a variable name linearmodel. Next, the model is trained with the training data using the fit function which I had done in previous models too. The x\_train and y\_train data are fit into the model to train it. These data are given to a linear regression model, which subsequently identifies any patterns in the data. For instance, the model can comprehend that as activity time increases, so do the calories burned. These things are automatically learned by our model.

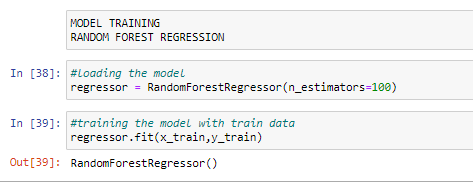


**Figure 19: MAE and R squared value for linear model**

It is known from the figure that the prediction was carried out on the test data by the model using predict function and the results on the test data is printed. The predicted value is then compared with the original value i.e y\_test. The mean absolute error was found to be 8.3 which shows that the predicted value has a huge error and our test results are flawed and the R squared value as 0.966 which showcases that the model fits the data to a high extent.

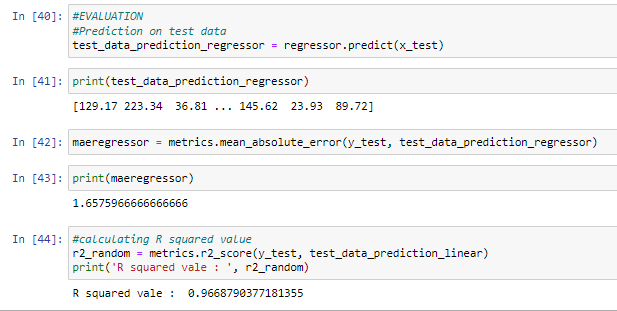
**Testing the Random Forest Regressor**

A supervised learning technique called Random Forest Regression leverages the ensemble learning approach for regression. The ensemble learning method combines predictions from various machine learning algorithms to provide predictions that are more accurate than those from a single model. Also, the authors in (Marte Nipas, 2022) claims that their result shows that Random Forest regression is the best model for the study with an accuracy of 95.77%. This led me to test the prototype for this model too.



**Figure 20: Loading the Random Forest Model**

From the sklearn package containing ensemble learning, the RandomForestRegressor is imported, an variable is assigned to it’s instance. The parameter n\_estimators determines the number of trees in the random forest. We train the model using the .fit() function, modifying the train data in accordance to improve the accuracy.



**Figure 21: MAE and R squared value for Random Forest Regressor**

The.predict() method is used after our model has completed training and is prepared to make predictions. The predicted value is then compared with the original value i.e y\_test. The MAE was found to be 1.65 which showcases that it’s prediction is close to the actual value and is comparable to the XGB regressor whereas the R squared value was found to be 0.966 which confirms the data to be best fit for the model.

Among the three model being tested, the XGBoost Regressor has been found to have the best fit data for prediction as its R squared value is close to 1 i.e 0.99, also the mean absolute error value is just 1.4 which is the least among all three models and the predictions are done with less erorr. This showcases that the XGBoost model is best fitted for this project prototype and thus, I have used this regression model to predict the number of calories burnt by the users.

**Conclusion**

The idea that machines can learn new things in the same manner that people do has been around for a very long time. It is crucial to keep in mind that AI is designed to make human lives better. It could be seen as a tool to enable us to advance past our circumstances. As AI technology develops, we might need to adapt and assess ourselves. I was astonished from the advancement of machine learning algorithms in research I carried out. The module and the prototype development has intrigued me at higher level and as of now I am more interested in Data Science techniques and machine learning techniques.

The literature review for identifying the prototype was done and as conclusion the effective model was chosen. As per the given assessment, the solution to the problem of inappropriate prediction of calories burnt and the unhealthy lifestyle is solved using the XGB regressor model to predict the calories burnt with high accuracy. After I had chosen the problem domain from the research, the system methodology was classified. Python was used as the programming language as we were taught python in our module. Next, the important dependencies like numpy, pandas, seaborn, matplotlib and all the regression model used in the project were imported. Machine learning models were created for the main objective, and their effectiveness was evaluated using mean absolute value and R squared value, using the dataset's CSV that was acquired from the Kaggle repository. The data was processed and visualized using various methods. The processed data was then split into training and test data for training our model and to evaluate it respectively. While training the model, I thought of going with the logistic regression taught in the model, but later found out that it shows a huge error and this model is specially designed to output binary results rather can prediction numbers. From the prior research and the module contents, I was able to choose the XGB regressor model and came to surprise when this model was found to be the best fit in comparison to other tested modules. This model was found to give the least mean absolute error and the R squared value being so close to 1 showcases the best fit of given data for this model.

This academic module had helped me a lot learning the concepts of machine learning and the use of some prediction models. This increased my confidence while tackling the machine learning models and new methodology. The Python programming taught in this module and some of the internet resources I learnt from helped me to code the complete project on my own. The analyzing of data for visualization and separation was quite challenging for me at first, but later the concepts were made easy by our model tutor and the insights I got from my research made me capable to use various visualization techniques.

# References

Audace B. K. Didavi, R. G. A. a. M. A., 2021. 'Comparative study of Decision Tree, Random Forest and XGBoost performance in forecasting the power output of a photovoltaic system.' *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART).*

Dobrev, D., 2005. 'A definition of Artificial Intelligence.' *Mathematica Balkanica,* Volume 19, pp. 67-74.

First Teddy Surya Gunawan, M. K. N. A. M. a. N. I., 2018. 'Food Intake Calorie Prediction using Generalized Regression Neural Network.' *2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA),* pp. 1-4.

Marte Nipas, A. G. A. J. N. M. M. A. F. M. J. A. B. S. a. J. S. G., 2022. 'Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm. '*2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T),* pp. 1-4.

Megha Rathi, D. G. M. G. a. V. S., 2018. 'Prediction of Health Fitness and Heart Status Using Data Mining Techniques.' *roceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018 held at Malaviya National Institute of Technology, Jaipur (India) on March 26-27, 2018.*

Tapia, E. M., 2008. 'Using Machine Learning for Real-time Activity Recognition.' *Massachusetts Institute of Technology,* Issue 2008, pp. 481-493.