Comparison of Machine Learning Methods for Calories Burn Prediction

Alfred Tan Jing Sheng[1], Zarina Che Embi[1*] and Noramiza Hashim[1]

[1]Faculty of Computing and Informatics, Multimedia University, Malaysia

Corresponding author: (zarina.embi@mmu.edu.my, ORCiD: 0000-0001-9378-7380)

Abstract - This paper focuses on the prediction of calories burned during exercise using machine learning techniques. Due to a growing number of obesity and overweight people, a healthy lifestyle must be adopted and maintained. This study explores and compares several machine learning regression models namely LightGBM, XGBoost, Random Forest, Ridge, Linear, Lasso, and Logistic to assess their calories burned prediction performance that can be used in systems such as fitness recommender systems supporting a healthy lifestyle. Our findings show that the LightGBM for predicting calorie burn has a good accuracy of 1.27 mean absolute error, giving users reliable recommendations. The proposed system has a good potential in assisting users in reaching their fitness objectives by offering precise and tailored advice.

Keywords— Exercise, machine learning, prediction, calories burn, recommender system

I. INTRODUCTION

Over the last few years, technology has improved the convenience and connectivity of our lives, but it has also had a negative impact on our health and wellbeing. Increased rates of obesity, heart disease, and other chronic health issues are a result of long workdays, sedentary lives, and a lack of physical activity. These disorders can have a substantial negative effect on general health and quality of life, as well as raise the price of healthcare. According to [1], the environment and lifestyle factors, like physical activity and eating habits, are also thought to be of vital relevance, even though genetics play a significant part in obesity. In addition, people who are less active may experience muscle deterioration and slowed metabolism, which makes it more difficult for them to maintain a healthy Body Mass Index (BMI). In an article by [2], there are ways to reduce overweight and obesity such as engaging in regular physical activity for 150 minutes a week for adults. Therefore, exercise is very important to everyone, especially for those who are obese and overweight.

According to the World Health Organization (WHO) [3], Malaysia has the highest rate of obesity and overweight among Asian countries with 64% of men and 65% of women being fat or overweight. As stated in [4], overweight and obesity cases are rising throughout the nation at an alarming rate right now. Exercise is an essential part of a healthy lifestyle and to have a good BMI. Exercise increases energy expenditure, which directly affects the number of calories burned by the body. Regular exercise can help balance calorie intake, regulate weight, and enhance general health.

However, due to a lack of information, motivation and support, many people with obesity find it difficult to start and stick with an exercise programme.

Knowing how many calories are burned while exercising might help people make more informed diet and exercise choices. Calories burned can be estimated by a number of methods such as metabolic equations [5], heart rate [5], wearable fitness trackers [6], and advanced sensors [7] by taking into account characteristics such as age, gender, weight, height, activity type, heart rate, exercise duration and intensity. Prediction of calories burned can be made using machine learning predictive models.

In this paper, we propose the use of LightGBM for predicting the amount of calories burned during exercises. It is compared to several machine learning algorithms namely XGBoost Regression, Random Forest Regression, Ridge Regression, Linear Regression, Lasso Regression and Logistic Regression. These machine learning algorithms will be applied on a dataset that is obtained from Kaggle. By accurately predicting the calories burned during exercise, individuals can have a better understanding of their physical activity and make informed decisions to achieve their health goals. This study aims to provide individuals with a reliable and accurate tool for estimating the calories expended during physical activities.

II. LITERATURE REVIEW

In multiple attempts to assist the consumer in losing their weight and gaining a healthy BMI, there have been several machine learning techniques introduced to predict calories burn.

A. Calories Burned Prediction Models

Maintaining a healthy lifestyle has grown more crucial in today's fast-paced environment. Managing our calorie intake and making sure we engage in adequate physical activity to burn enough calories is a crucial component of living a healthy lifestyle. However, it can be difficult and confusing to anticipate how many calories would be burned during diverse activities. In order to overcome this obstacle, scientists and health enthusiasts have created cutting-edge techniques for estimating calorie expenditure based on several variables like body weight, exercise type, duration, and intensity. These techniques are intended to provide people with a better knowledge of how much energy they use and to help them set reasonable fitness goals, create efficient workout schedules, and monitor their progress.

This study's goal is to provide an overview of calorie burned prediction based on machine learning regression models namely XGBoost, Linear, Random Forest, Lasso, Logistic, Ridge and LightGBM, which will be investigated. By analysing the advantages and disadvantages of these methods, the study aims to provide valuable insights into the current state of calorie burn prediction.

B. XGBoost Regression

Due to its remarkable performance and adaptability in regression problems, the XGBoost regression has attracted significant interest and popularity in the field of machine learning. For calories burned prediction, it has been used in works by [8], [9] and [10]. Extreme Gradient Boosting, or XGBoost, is an approach for ensemble learning that combines regularisation methods with the strength of gradient boosting algorithms. According to [8], the XGBoost algorithm performs well since it has robust handling of many varieties of data types, relationships, distributions, and the many hyperparameters that you can fine-tune.

A wide variety of regression issues can be solved using XGBoost because of its capacity to handle various data formats and recognize sophisticated patterns. XGBoost keeps all intermediate calculations in cache so that we don't have to do the same calculation again and again [8]. XGBoost Regression is also known for its performance of managing missing values, handling category and numeric features, and efficiently handling overfitting. XGBoost has regularly shown higher performance than other regression models such as linear regression [9,11]. Although the XGBoost regression is a good model, there are still some limitations to it. XGBoost requires very careful tuning of hyperparameters to get good performance. Other than that, XGBoost makes use of a collection of decision trees, which can use up a lot of memory, especially when working with deep trees or big feature areas.

C. Linear Regression

Linear Regression is a supervised learning algorithm and has been used for calories burned prediction by [8], [11] and [12]. A dependent variable is related to one or more independent variables using the basic statistical modelling approach known as linear regression. Linear regression model is used to identify the linear relationship of the input variables and output variables [11]. The dependent variable can be predicted using the independent variables and the model's estimated coefficients for the slope and intercept of the line. The coefficients are estimated using a variety of methods, such as gradient descent, maximum likelihood estimation, and ordinary least squares (OLS).

By reducing the sum of squared residuals, these methods help the model accurately represent the underlying linear relationship. Numerous industries, including economics, finance and healthcare use linear regression extensively. There are some limitations in this machine learning model such as the assumptions of linearity, independence of mistakes, and homoscedasticity. Additionally, issues with multicollinearity, outliers, missing data, and residual non-normality are also some of the limitations that will affect the validity dependability of linear regression analysis.

D. Random Forest Regression

Random Forest Regression is a powerful ensemble learning method that combines the advantages of decision trees and the efficiency of ensemble techniques. Ensemble learning is a method that combines predictions from different machine learning algorithms. [11] states that it achieves a higher prediction accuracy than using a single model. Due to its adaptability and reliable performance, Random Forest Regression has found extensive applications across industries such as banking, healthcare and marketing.

High prediction accuracy, robustness against outliers and noise, and the capacity to handle big datasets with high-dimensional feature spaces are just a few benefits of Random Forest Regression. According to [11], Random Forest has outperformed linear regression and ridge regression with the lowest mean absolute error of 1.81 calories burned predictions. The limitations of random forest regression are memory hungry, parameter sensitive and limited extrapolation capability. Random forest regression can be memory hungry when working with huge datasets. The number of resources required to build numerous decision trees and store their structures in memory may restrict the algorithm's capacity to scale in some situations.

E. Lasso Regression

A popular regression method that combines the advantages of regularisation with linear regression is called lasso regression, also known as least absolute shrinkage and selection operator or L1 regularisation. This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination [10]. By adding a regularisation term to the conventional linear regression model, lasso regression produces a more reliable and economical solution.

Due to its capacity to manage high-dimensional data and choose pertinent characteristics, lasso regression has found extensive use in a variety of fields, including calories burned prediction [10]. Lasso regression has been used in the world of finance for credit scoring, risk prediction, and portfolio optimization.

F. Logistic Regression

Authors in [10] also compare another model called logistic regression for calories burned prediction. This model is a useful analysis method for classification problems to determine if a new sample fits best into a category [10]. Logistic regression is known for modelling binary or categorical outcomes. It estimates the likelihood of an event happening in the future from a set of predictor factors. The linear combination of predictors is transformed into a probability between 0 and 1 using the logistic function, also referred to as the sigmoid function. The outcome variable and the predictors are related in a certain direction and to a certain extent, as shown by the model's estimated coefficients.

Although effective, logistic regression has certain drawbacks. It presupposes that there is a linear relationship between the predictor factors and the outcome's log-odds. More complicated models might be needed for nonlinear connections. Research is also required to deal with multicollinearity problems and handle missing data in logistic regression.

G. Ridge Regression

A common regularisation method in linear regression is ridge regression. Ridge regression is applied to data with multi-collinearity [11]. Ridge regression adds a penalty term to the least squares objective function to improve as a linear regression. The penalty term, controlled by a tuning parameter lambda (λ), imposes a constraint on the sum of

squared coefficients, shrinking them towards zero. By dispersing the influence of correlated predictors, this method successfully lessens the effect of multicollinearity.

Ridge regression has been used in a wide range of applications due to its ability to manage multicollinearity and improve prediction accuracy. It has been extensively employed in sectors like banking, economics, and healthcare. In [11], it has been used for predicting calories burned. The limitations of ridge regression are sensitive to feature scaling and limited handling of outliers. All predictors are seen as being on a comparable scale in Ridge regression. Different scales or units between the predictors can produce skewed values and impact the regularisation procedure. Ridge regression is sensitive to outliers in the data. Results may be skewed as a result of outliers' excessive influence on the computed coefficients.

H. LightGBM Regression

LightGBM (Light Gradient Boosting Machine) is a popular machine learning algorithm that is a part of the gradient boosting framework. It uses a gradient-based optimization strategy that aims to reduce the loss function by including trees gradually. LightGBM can significantly outperform XGBoost and SGB in terms of computational speed and memory consumption [12]. Due to its capacity to manage big datasets, high-dimensional features, and intricate interactions, the LightGBM regression has become more popular across fields such as finance, e-commerce and marketing. The limitations of LightGBM Regression are potential overfitting, limited handling of missing data and limited performance on small datasets. LightGBM Regression requires sufficient regularisation to prevent overfitting.

Due to the algorithm's capacity for creating complicated models, training data may contain noise or outliers that cause the system to perform poorly on unobserved data. LightGBM regression does not directly deal with missing data and may need to use preprocessing or imputation methods to deal with missing values effectively. In order to prevent biassed results or model instability, missing data should be managed before the method is used. Larger datasets tend to yield better results for LightGBM regression due to its capacity for high-dimensional feature spaces. In comparison to other algorithms that are better suited for small sample numbers, the method may struggle to detect significant patterns or generalise well on tiny datasets.

I. Proposed Work

After conducting the literature review, it was found that there are two limitations that need further improvement. One is the lack of machine learning models to predict calories burned during exercise and another one is the lack of comparison between every machine learning model. Lack of machine learning models in prediction of calories burned is a very crucial problem as there might be another better machine learning algorithm that can make a more accurate prediction for calories burned.

Other than that, comparison between every machine learning model is very important in order to find the difference and the best machine learning model to predict the calories burned. In a previous paper by [8], XGB Regression is often being compared with only one other machine learning algorithm.

After identifying the limitations in previous research, the proposed work aims to provide a solution to the challenges. An existing machine learning repression algorithm, LightGBM, is proposed in this work to predict the calories burned. Based on our literature search, LightGBM regression has not been used for such a purpose. Additionally, we contrast the suggested regression model with six additional models: XGBoost, Linear, Random Forest, Lasso, Logistic, and Ridge.

III. METHODOLOGY

The goal of this study is to gather a suitable dataset for our machine learning models to be trained to predict calorie burn. The workflow is adapted from [13], which is shown in Figure 1. Data preprocessing will be done with the collected records to ensure data quality. After that, the data will be trained using the prepared dataset. In this comparison, XGBoost regression, linear regression, lasso regression, logistic regression, ridge regression, random forest regression and LightGBM regression will be used.
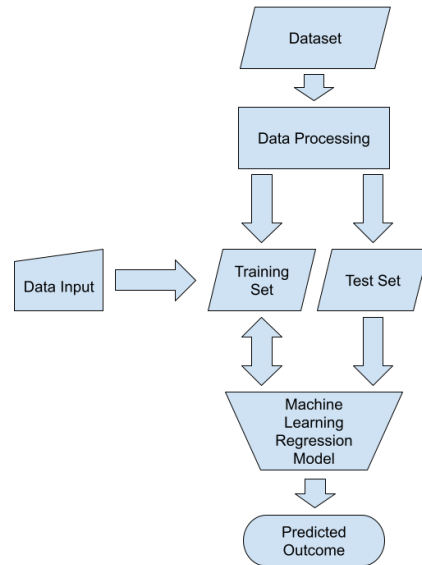
Figure 1: Work flow adapted from [13]

A. Data Collection

As the quality and accuracy of the data collected can have a considerable impact on the study's findings, data collection is an important step in the research process. In this research work, the dataset is collected from a publicly available Kaggle[1]dataset. The data collected is categorical and numerical.

B. Data Preprocessing

In this study, a dataset containing 15000 instances and 7 attributes from Kaggle are used. It includes the person's height, weight, age, gender, body temperature, heart rate and workout duration. The detail of the dataset is depicted in Table 1.

Table 1. Attributes and values for data

Attribute	Value
Gender	Male for 0, female for 1
Age	The person's age in years
Calories	Total amount of calories burned during workout
Height	The person's height in centimeters
Weight	The person's weight in kilograms
Heart_Rate	Average heart rate during workout
Body_Temp	Average body temperature during workout in degree celsius
Duration	Total time used to workout in minutes

---

[1] https://www.kaggle.com/datasets/fmendes/fmendesdat263xdemos

B. Data Analysis

The process of gathering, modelling, and analysing data is known as data analysis [14]. The data analysis phase is an important part of this study since it offers understanding of the dataset, ensures data quality, and makes it easier to choose and assess machine learning algorithms for estimating calories burned during exercise. The outcomes of this phase's research will further our knowledge of the relationships between numerous variables and improve our ability to forecast calorie burn accurately.

C. Gender Distribution

Gender distribution shown for this data in the pie chart is very equally distributed. Male has 50.35% and female has 49.65% as can be seen in Figure 2.
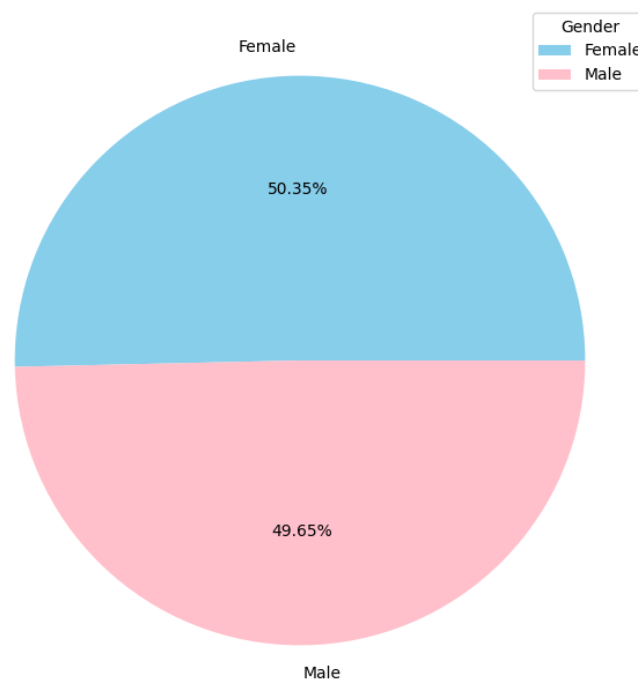

Figure 2: Gender distribution

D. Age Distribution

The age in the plot exhibits a positively skewed distribution, with most data points concentrated between 20 to 50 years. Understanding the characteristics of this column is essential for various analytical purposes and can guide further exploration and analysis of the dataset. From the age plot in Figure 3, it also shows that the older the age, the fewer the people who perform the workout.
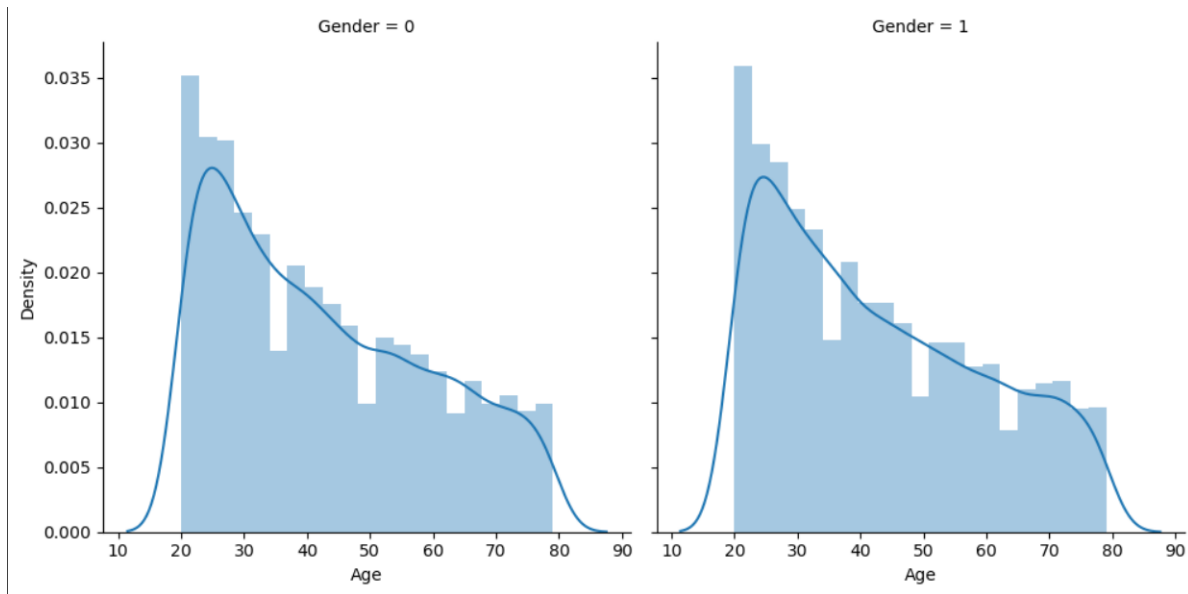
Figure 3: Age distribution

E. Correlation Heatmap

The heatmap (Figure 4) shows a strong positive correlation between calories and body temperature, heart rate, and duration. This means that increased body temperature, increased heart rate, and prolonged physical activity are all linked to increased calorie expenditure.
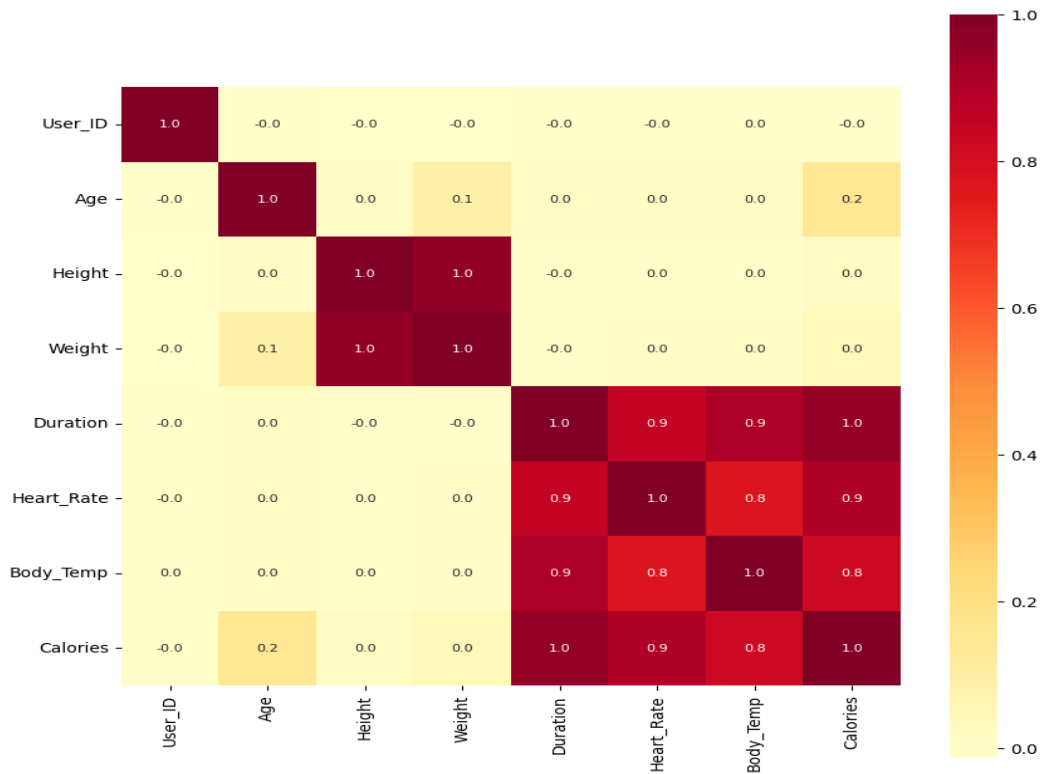

Figure 4: Correlation heatmap

188

F. Model Selection & Model Training

There are a lot of machine learning models that are compared to get the best machine learning models to predict the calories burned based on the data. These models are LightGBM, XGBoost, linear regression, logistic regression, lasso regression, ridge regression, and random forest regression.

G. Evaluation

After the model selection and model training, the performance of the calories burned prediction models is evaluated using a scale-dependent metrics that is based on absolute errors. The mean absolute error (MAE) measures the magnitude of the absolute errors between the predicted value and the actual value. MAE is defined as in Eq (1) [15].

$$MAE = \frac{l}{n}\sum_{l}^{n}|D_{pre} - D_{act}| \tag{1}$$

where n denotes the amount of data collected, D_{act} denotes the variable's actual value, D_{pre} denotes its predicted value. The MAE has a range of 0 to + whereby the lower the MAE value, the prediction model is more accurate [15].

IV. RESULTS

The analysis of this dataset was conducted to forecast the number of calories burned based on the activity time as well as the gender, age, body temperature, and heart rate at various points during the workout. All these seven machine learning algorithms models were compared to find the model with least MAE which provides the most accurate prediction data as shown in Table 2. Therefore, the best model is LightGBM regression with the MAE of 1.27 as can be seen in Table 3.

Table 2. Predicted and expected calories burned for each machine learning model

Machine Learning Model	Data Input	Predicted Calories Burned	Expected Calories Burned
LightGBM Regressor	(1,39,156.0,62.0,28.0,104.0,40.8)	170.63	170.0
XGBoost Regressor	(1,39,156.0,62.0,28.0,104.0,40.8)	170.75	170.0
Random Forest Regression	(1,39,156.0,62.0,28.0,104.0,40.8)	170.88	170.0
Lasso Regression	(1,39,156.0,62.0,28.0,104.0,40.8)	173.60	170.0
Logistic Regression	(1,39,156.0,62.0,28.0,104.0,40.8)	174.0	170.0
Ridge Regression	(1,39,156.0,62.0,28.0,104.0,40.8)	174.33	170.0
Linear Regression	(1,39,156.0,62.0,28.0,104.0,40.8)	174.33	170.0

Table 3. Comparison of different regression models including our proposed model

Machine Learning Model	Mean Absolute Error (MAE) in Calories
LightGBM Regressor	1.27
XGBoost Regressor	1.48
Random Forest Regression	1.68
Ridge Regression	8.39
Linear Regression	8.39
Lasso Regression	8.39
Logistic Regression	15.39

V. CONCLUSION

Various machine learning algorithms to predict calorie burn during exercise as well as the advantages and disadvantages have been compared. To assist people in establishing and maintaining a healthy lifestyle, precise estimations of calories burned during physical exercise are required due to the rising incidence of obesity and overweight around the world.

The literature review provided insights into various machine learning regression models for predicting calories burned, each with advantages and disadvantages in terms of performance, handling of missing data, scaling, and handling of outliers. This work proposed using the LightGBM regression model, which has demonstrated excellent results in handling huge datasets to solve the limitations of earlier studies.

The methodology included collecting data from Kaggle, preparing the data to ensure data quality, analysing the data to learn more about the dataset, choosing a model, and then training the model using the selected machine learning algorithms. The accuracy and efficiency of the models were evaluated using metrics, namely mean absolute error.

By comparing the performance of the machine learning models, valuable insights can be gained by contrasting their performance. The findings can then be used in a recommender system such as [16], [17] and [18] to accommodate individuals who are obese or seeking to live a healthy lifestyle in terms of exercise intensity and duration. It can also provide useful information on other areas such as in predicting travel insurance purchases that has been conducted by [19] as well as for the classification of ECG heartbeat [20].

LightGBM, XGBoost, Linear, Lasso, Logistic, Ridge and Random Forest regression models are used to predict the precise number of calories burned, which depends on a variety of variables. Based on the analysis and the evaluation metric, LightGBM regression has the lowest mean absolute error (MAE) of 1.27. The LightGBM regression is thus the most accurate model for estimating calories burned.

The results of this study will further enhance the understanding of calorie burn prediction and help people manage their weight and lead healthy lives. To improve the precision and reliability of predictions regarding calorie burn, more studies can be done to examine additional machine learning techniques or optimise the current models.

ACKNOWLEDGEMENT

REFERENCES

[1] X. Chen and J. Cui, "The association between BMI and health-related physical fitness among Chinese college students: a cross-sectional study," BMC Public Health, Apr. 5, 2020. [Online]. Available: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-08517-8

[2] V. Martin, "Obesity and overweight," World Health Organization (WHO), Jun. 9, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[3] N. Kaur, "Malaysia and WHO call for more investment in primary health care in the 21st century," World Health Organization (WHO), Apr. 8, 2019. [Online]. Available: https://www.who.int/malaysia/news/detail/08-04-2019-malaysia-and-who-call-for-more-investment-in-primary-health-care-the-21st-century

[4] Bernama, "One of two adults in Malaysia is overweight, says Deputy Health Minister," New Straits Times, Mar. 25, 2022. [Online]. Available: https://www.nst.com.my/news/nation/2022/03/783279/one-two-adults-malaysia-overweight-says-deputy-health-minister

[5] D. D. T. D. Dakshima, K. Seliya Mindula, R. M. S. J. Rathnayake, S. Kasthuriarachchi, A. K. Buddhi Chathuranga and D. Lunugalage, "Health Care – A Personalized Guidance for Non-Communicable Diseases," 2022 4th International Conference on Advancements in Computing (ICAC), Colombo, Sri Lanka, 2022, pp. 447-452, doi: 10.1109/ICAC57685.2022.10025109.

[6] A. Pande, J. Zhu, A. K. Das, Y. Zeng, P. Mohapatra and J. J. Han, "Using Smartphone Sensors for Improving Energy Expenditure Estimation," IEEE journal of translational engineering in health and medicine, vol. 3, 2015. doi: 10.1109/JTEHM.2015.2480082

[7] E. E. Dooley, N. M. Golaszewski, and J. B. Bartholomew, "Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices," JMIR mHealth and uHealth, vol. 5, no. 3, 2017. doi: 10.2196/mhealth.7043

[8] S. P. Vinoy and B. Joseph, "Calorie Burn Prediction Analysis Using XGBoost Regressor and Linear Regression Algorithms," NCECA, 2022. [Online]. Available: https://nceca.in/2022/46_Calorie_Burn_Prediction_Analysis_Using_XGBoost_Regressor_and_Linear_Regression_Algorithms.pdf

[9] N. Meenigea, "Calorie Burn Prediction: A Machine Learning Approach using Physiological and Environmental Factors," Jul. 1, 2014. [Online]. Available: https://www.jetir.org/view?paper=JETIR1701933

[10] R. K. Singh and V. Gupta, "Calories Burnt Prediction Using Machine Learning," May 2022. [Online]. Available: https://ijarcce.com/wp-content/uploads/2022/06/IJARCCE.2022.115105.pdf

[11] M. Nipas et al., "Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm," May 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9776710

[12] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Y. Liu, (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," NIPS papers, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[13] S. S. Ratnakar and V. S, "Calorie Burn Prediction using Machine Learning," Jun. 2022. [Online]. Available: https://iarjset.com/wp-content/uploads/2022/07/IARJSET.2022.961125.pdf

[14] B. Calzon, "What Is Data Analysis? Methods, Techniques, Types & How-To," BI Blog | Data Visualization & Analytics Blog | Datapine, Mar. 9, 2022. [Online]. Available: https://www.datapine.com/blog/data-analysis-methods-and-techniques/#data-analysis-definition

[15] A. Jierula, S. Wang, T. M. Oh, and P. Wang, "Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data," Applied Sciences, vol. 11, no. 5, 2021.

[16] A. Mahyari and P. Pirolli, "Physical Exercise Recommendation and Success Prediction Using Interconnected Recurrent Neural Networks," in 2021 IEEE International Conference on Digital Health (ICDH), Chicago, IL, USA, 2021 pp. 148-153. doi: 10.1109/ICDH52753.2021.00027

[17] M. Roy, "OBESEYE: Interpretable Diet Recommender for Obesity Management using Machine Learning and Explainable AI," IJRAMT, vol. 4, no. 6, pp. 1-7, 2023. [Online]. Available: https://journal.ijramt.com/ijramt/article/view/2733

[18] Preethi Brunda, Sushmitha Aralikatte, and Manjula, "B-Fit: A Fitness and Health Recommendation System," International Research Journal of Engineering and Technology (IRJET), vol. 9, no. 7, Jul. 2022.

[19] S. T. Lim, J. Y. Yuan, K. W. Khaw, and X. Chew, "Predicting Travel Insurance Purchases in an Insurance Firm through Machine Learning Methods after COVID-19," Journal of Informatics and Web Engineering, vol. 2, no. 2, 2023, pp. 43-58.

[20] J. O. Victor, X. Chew, K. W. Khaw, and M. H. Lee, "A Cost-Based Dual ConvNet-Attention Transfer Learning Model for ECG Heartbeat Classification," Journal of Informatics and Web Engineering, vol. 2, no. 2, 2023, pp. 90-110.