RESEARCH ARTICLE

A STUDY ON CALORIES BURNT PREDICTION USING RIDGE AND RANDOM FORESTREGRESSION METHOD

Prakash Rajaram Chavan

Department of Statistics,

Smt. Kasturbai Walchand College of Arts & Science, Sangli, Maharashtra, India

Affiliated to Shivaji University, Kolhapur

*Corresponding author E-mail: prchavan83@gmail.com, prchavan83@kwc.in

DOI: https://doi.org/10.5281/zenodo.16832519

Abstract:

The overarching idea of this research study is to make a comparative study of machine learning algorithms to predict the calories burn during the workout. In this research study we first build a machine learning system that can predict the amount of calories burnt during exercise. In current century many people are doing workout according to weight loss plan that they have taken and calculates how much calorie do they burn once they workout. To solve this problem, we use ML algorithms such as Random Forest regression and Ridge Regression.

Keywords: Prediction, Calories, Random Forest, Ridge Regression, Weight.

Introduction:

In this research study, let's predict the calories burn using Machine learning & let's have a healthy & a happier life. This research study is about calorie prediction with machine learning using python. We will predict calorie based on some features. We eat foods to provide energy so that our bodies can function. This means that we need to eat a certain amount of calories just to sustain life. The risk gaining weight is increases if we take in too many calories. So, there is need to burn Calories, for burning calories we doing exercises and more. For know how much calories we have burn during exercise we are going to build a machine learning model that predict calories.

Based on the same data. When we exercise, the body temperature and the heartbeat willbe rise. The variables time scale is taken for which the individual carrying out the workout training, the average beats per minute and temperature. Then we additionally take the height, weight, gender and age of the person to predict how tons energy the person may be burning. Suvarna S R. et al. (2022) predicted calorie burn using Machine Learning.

In present research paper, we want to predict calories burned using machine learning random forest regression algorithm and ridge regression algorithms on the independent variables, namely, duration, body temperature, height, weight and age of the person.

Material and Methods:

This research study is all about the collection of appropriate set to teach our machinelearning models in order that it will find out what is the amount of calories that the individual

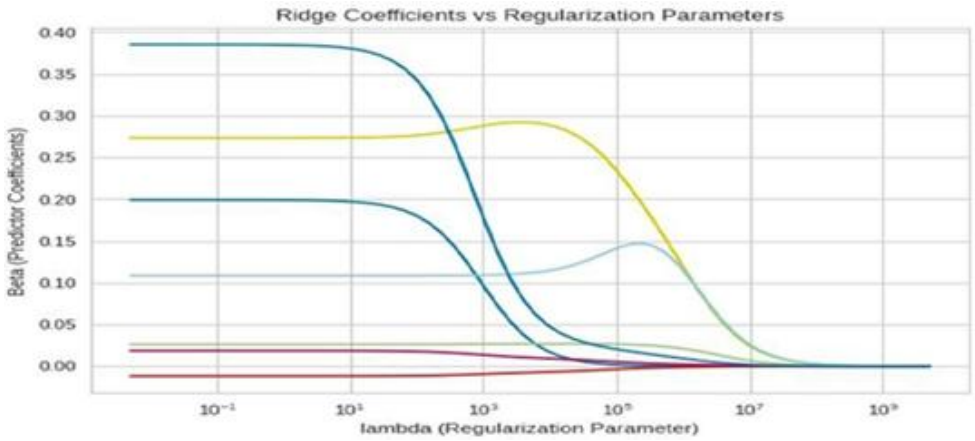
goes to burn. Before feeding procedure the statistics via records pre-processing need to be done. After that data analysis is carry out where we use some visualization techniques to arrange thedata in plots and graphs. Afterwards divide the data set into training and test set. Here we use Random Forest regression and Ridge regression as machine learning models for comparison and then evaluate MS-Excel and python.

Statistical Analysis:

Ridge Regression:

Ridge regression is the method used to analysis the multi co-linearity in multiple regression. When the data contains more number of independent variables then it is most suitable.

It is also known as, L2 Regression adds a penalty to the existing model. Ridge regression adds penalty to loss function which makes the model have a smaller coefficients value. i.e, it shrinks coefficients of the variables of the model that don't contribute much to themodel. Based on the Sum of Square Error it penalizes the model**.** But it prevents from beingexcluded from model by letting them have towards zero as coefficients value



Interpretation:

From the plot as alpha increases the coefficients convert to smaller values of their original. The power of ridge regression is to make the coefficients smaller to limit the co-linearity between predictors.

Ridge regression can be framed as follows:

Ridge = loss + (lambda *L2_penalty)Ridge = loss+ ($\lambda * \Sigma\beta j^2$)

$$\textbf{Loss} = \Sigma(yi - \hat{y}i)2$$

- **Σ** : A Greek symbol that means sum
- **yi** : The actual response value for the i^{th} observation
- **ŷi** : The predicted response value based on the multiple linear regression model
- where *j* ranges from 1 to P predictor variables and $\lambda \geq 0$.
- This second term in the equation is known as a shrinkage penalty. we select the valueof λ that produces the lowest possible test mean squared error.

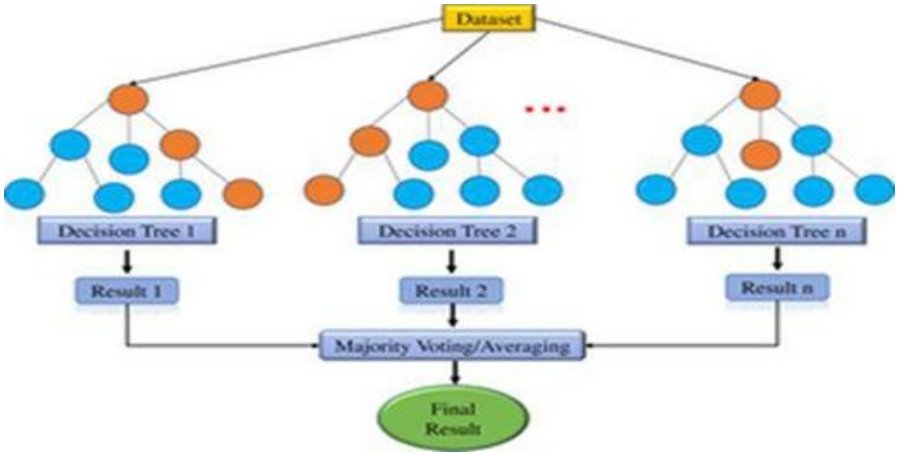
Random forest regression:

Random forest is a type of supervised machine learning algorithm based on ensemble learning. The Random Forest algorithm combines multiple algorithms of the same type. i.e. The multiple decision trees resulting in forest of trees called asRandom Forest. For both the regression and classification tasks, the random forest algorithm can be used.

Algorithm for random forest:

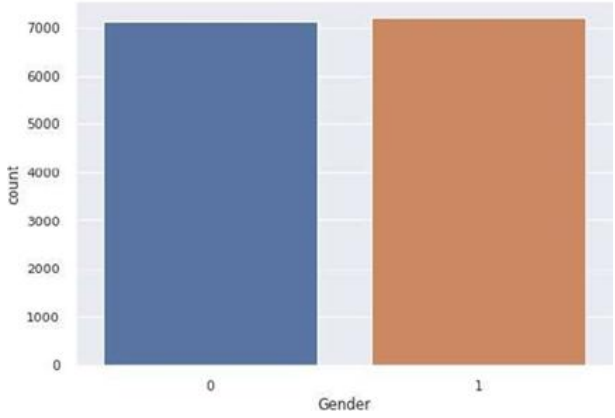
The following are the basic steps involved in performing the random forest algorithm:

- From the dataset, Pick N random records.
- Based on these N records build a decision tree.
- Select the number of trees that you want in algorithm and repeat steps-1 and 2.
- In case of a regression problem, for a new record, each tree in the forest predicts a value forY (output). By taking the average of all values predicted by all the trees in forest, the final valuecan be calculated. In case of a classification problem, each tree in forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that winsthe majority vote.



Data visualization:

sns.countplot(combined new['Gender'])



Interpretation:

From the above plot, we can observe that the data is evenly distributed forboth males and females.

Now we plot heat map for understanding correlation:

Heat map gives colors based on values, and these values are calculated based on therelationship between the data. Each column would be compared to the other column and if thevalue is 1 than two columns are positively correlated if the value is 0 than two columns are notcorrelated, if less than 0 than two columns are negatively correlated. So from graph it can beseen duration, heart rate and body temperature are positively related to calories burnt. plt.figure(fig size=(10,10)) sns.heatmap(correlation, cbar=True, square=True, fmt='.1f', annot=True, annot_kws={'size':8}, cmap='Purples')

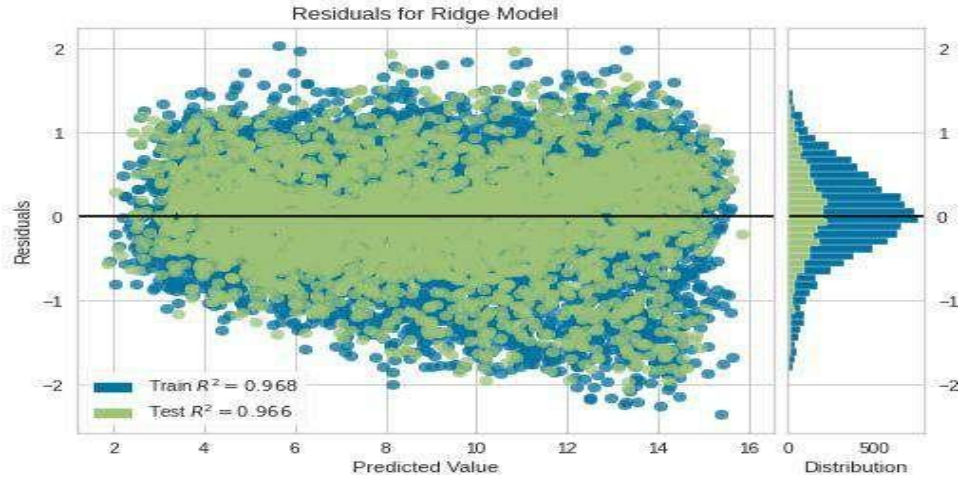
Interpretation:

From above plot we can see that Duration, Heart Rate & Body Temp is highlycorrelate in target variable (Calories). Here, we can see that Height and weight are highly correlated with each other, and Duration,Heart Rate & Body Temp are highly correlated with each other.

Therefore multi co- linearity exists. From this we get that the height and weight are positively correlated and the duration, heartrate and body temperature are highly positively correlated with calories.

Residual plot of ridge regression:

from yellowbrick.regressor import ResidualsPlot visualizer = ResidualsPlot(ridge_model) visualizer.fit(x_train, y_train) visualizer.score(x_test, y_test) visualizer.show()



Interpretation:

In the case above, we see a fairly random, uniform distribution of the residualsagainst the target in two dimensions. This seems to indicate that our linear model is performingwell. We can also see from the histogram that our error is normally distributed around zero, which also generally indicates a well fitted model.

Result:

The analysis of this dataset was done to predict the calories burned depends on theduration of workout and also based on the gender, height, weight, age body temperature and heart rate at some stage in the exercise. By using these machine learning algorithms, we are looking for a machine learning model with greater R^2, which gives more accurate results. By comparing the two models, Random forest regression and Ridge regression we get that the Random forest regression gives the more accurate results of the calories burned with R^2 0.9995515067354402 than the Ridge regression.

Machine Learning Model	R^2	Adj R^2
Ridge Regression	0.9875484041534515	0.9875239960966254
Variable selection	0.9660611777097421	0.9660326975791489

Machine Learning Model	Input data	Predicted Calorie result	Expected Calorie result
Random Forest Regression	Female{1},21,139.0,43.0,17.0,	9.45825413	9.539392
Variable selection	98.0, 40.2 17.0,98.0,40.2	9.52504734	9.539392
Ridge Regression	Female{1},21,139.0,43.0,17.0,	8.87111566	9.539392
Variable selection	98.0,40.2 17.0,98.0,40.2	9.28860214	9.539392

Conclusions:

From the analysis we met with a conclusion that the Random Forest Regression has more accurate results than the Ridge regression model. The R^2 value that is getting in Random forest regression is 0.9995 which is very close to 1 than the Ridge regressionmodel R^2 value 0.9875. It means that the Random forest model R^2 value is larger than the Ridge regression R^2 value. Therefore we can conclude that the best model for the calories burnprediction is Random Forest Regression

References:

1. Suvarna, S. R., & Vidya, S. (2022). Calorie burn prediction using machine learning. *International Advanced Research Journal in Science, Engineering and Technology, 9*(6), 781–787.
2. Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM.*
3. MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.
4. Goukens, C., & Klesse, A. K. (2022). Internal and external forces that prevent (vs. facilitate) healthy eating: Review and outlook within consumer psychology. *Current Opinion in Psychology*, 101328. https://doi.org/10.1016/j.copsyc.2022.101328

5. Khan, A. W., et al. (2022). Factors affecting fitness motivation: An exploratory mixed method study. *IUP Journal of Marketing Management, 21*(2).
6. Roberts, K. C., Shields, M., de Groh, M., Aziz, A., & Gilbert, J. A. (2012). Overweight and obesity in children and adolescents: Results from the 2009 to 2011 Canadian Health Measures Survey. *Health Reports, 23*(3), 37–41.
7. Jadhav, K., et al. (2023). Human physical activities based calorie burn calculator using LSTM. In *Intelligent cyber physical systems and internet of things: ICoICI 2022* (pp. 405–424). Springer International Publishing.
8. Tayade, A. R., & Katesari, H. S. (n.d.). A statistical analysis to develop machine learning models: Prediction of user diet type.