## Artificial

## Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Calories Burnt Prediction

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

**Mohammad Talib Shivam Sharma 2210992316**

**Shivansh Khunger 2210992324**

**Shivean Gupta 2210992332**

**Sehaj Goyal 2210992277**

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,

## Chitkara University, Punjab

**Table of Content**

|  |  |  |
| --- | --- | --- |
| Sl. No | Name | Page No |
| 1 | Introduction | 3-5 |
| 2 | Exploratory Data Analysis | 6-7 |
| 3 | Model Building | 8-9 |
| 4 | Model Evaluation and Comparison | 10-11 |
| 5 | Results | 11-12 |
| 6 | Conclusions | 13 |
| 7 | References | 13 |

**1. Introduction**

**1.1 Project Objectives**

The primary objectives of the "Calories Burnt Prediction" project are:

1. To develop a machine learning model that accurately predicts the number of calories burnt based on individual characteristics such as height, weight, age, and gender.

2. To provide individuals with a tool for better understanding their calorie expenditure during physical activities, aiding in fitness and health management.

**1.2 Scope of the Project**

The scope of the "Calories Burnt Prediction" project encompasses:

* 1. Data collection and pre-processing: Gathering a comprehensive dataset containing relevant variables such as height, weight, age, gender, and calories burnt, and preparing it for analysis.
* 2. Exploratory data analysis (EDA): Conducting an in-depth exploration of the dataset to identify patterns, correlations, and potential outliers.
* 3. Model development: Designing and training machine learning algorithms to predict the number of calories burnt based on the input variables.
* 4. Model evaluation: Assessing the performance of the developed models using appropriate evaluation metrics and techniques.
* 5. Deployment: Implementing the trained model into a user-friendly application or platform for practical use by individuals seeking to estimate their calorie expenditure.
* 6. Continuous improvement: Iteratively refining the model based on feedback and new data to enhance its accuracy and reliability over time.

**1.3 Overview of Techniques**

The "Calories Burnt Prediction" project will utilize a variety of machine learning techniques, including:

1. Regression: Employing regression algorithms to predict the continuous output variable (calories burnt) based on the input features such as height, weight, age, and gender.
2. . Feature engineering: Creating new features or transforming existing ones to improve the predictive power of the model, such as calculating body mass index (BMI) from height and weight data.

3. Cross-validation: Using techniques like k-fold cross-validation to assess the performance of the model and ensure its generalizability to unseen data.

4. Hyper parameter tuning: Fine-tuning the parameters of the machine learning algorithms to optimize their performance and enhance the accuracy of calorie burnt predictions.

5. Ensemble methods: Exploring ensemble techniques like random forests or gradient boosting to combine multiple models for improved prediction accuracy.

6. Interpretability techniques: Employing methods to explain and interpret the predictions of the model, enhancing transparency and trustworthiness, such as feature importance analysis.

These techniques will be applied iteratively throughout the project to develop an accurate and reliable model for predicting calorie expenditure.

**1.4 Data Features Overview**

The data features for the "Calories Burnt Prediction" project include:

1. Height: The height of the individual in centimetres or inches.

2. Weight: The weight of the individual in kilograms or pounds.

3. Age: The age of the individual in years.

5. Calories Burnt: The number of calories burned by the individual during a specific physical activity or time period.

These features serve as the input variables for the machine learning model, with "Calories Burnt" being the target variable that the model aims to predict based on the other features. By analysing these features collectively, the model can learn patterns and relationships to make accurate predictions of calorie expenditure.

**2. Exploratory Data Analysis (EDA)**

**2.1 Dataset Description**

During the exploratory data analysis (EDA) phase of the "Calories Burnt Prediction" project, the dataset will be examined to gain a comprehensive understanding of its characteristics. The dataset description includes:

1. Number of samples: The total number of instances or observations in the dataset.

2. Number of features: The total number of variables or attributes included in the dataset.

3. Data types: The data types of each feature, indicating whether they are numerical (e.g., integer, float) or categorical (e.g., string, boolean).

4. Summary statistics: Descriptive statistics such as mean, median, standard deviation, minimum, and maximum values for numerical features, providing insights into the central tendency and dispersion of the data.

5. Distribution of target variable: The distribution of the target variable "Calories Burnt," including histograms or density plots to visualize its spread and identify any potential outliers.

6. Correlation analysis: Examining the pairwise correlations between features, particularly between the independent variables and the target variable, to identify significant relationships and potential predictors of calorie expenditure.

By conducting a thorough dataset description as part of the EDA process, the project team can better understand the structure, quality, and potential challenges of the data, informing subsequent steps in model development and analysis.

**2.2 Data Preprocessing**

Before conducting exploratory data analysis, it is essential to preprocess the dataset to handle missing values, outliers, and categorical variables. Preprocessing steps include imputation, normalization, encoding categorical variables, and feature scaling. Additionally, we perform feature engineering to create new features or transform existing ones to better capture the

underlying patterns in the data.

**2.3 Data Visualization**

Exploratory data analysis (EDA) involves visualizing the dataset to uncover patterns, trends, and relationships between variables. We utilize various visualization techniques such as scatter plots, histograms, box plots, and heatmaps to gain insights into the distribution of features and their correlation with house prices. Visualization helps us identify outliers, understand the distribution of data, and discover potential relationships that may exist in the dataset.

**3. Model Building**

**3.1 Introduction to Machine Learning Models**

In this section, we introduce the machine learning models used in the project for house price prediction: linear regression, decision tree, random forest, and support vector machine (SVM). Each model is selected based on its suitability for regression tasks and its ability to capture complex relationships in the data.

**3.2 Linear Regression Model**

* Linear regression is a simple yet powerful algorithm for regression tasks. It models the relationship between the independent variables (features) and the dependent variable (calories burnt) by fitting a linear equation to the observed data points. The model assumes a linear relationship between the features and the target variable, making it suitable for datasets with linear dependencies.
* In the context of calories burnt prediction, linear regression can be used to estimate the effect of each feature on the house price. By analysing the coefficients of the linear equation, we can determine which features have the most significant impact on the calories burnt feature.
* The implementation of the linear regression model involves fitting a line to the data that best represents the relationship between the independent variables (features) and the dependent variable. Model evaluation is performed using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared to assess the accuracy and generalization ability of the model.

**3.3 Decision Tree Model**

Decision trees serve as a non-linear algorithm that segments the feature space into regions based on input feature values. Each internal node in the tree represents a decision based on a feature, while each leaf node represents a predicted calorie burn value. Decision trees are valued for their interpretability and capacity to capture nonlinear relationships within the data, which is crucial in predicting calorie burn accurately.

For our project, decision trees can effectively model the intricate interactions among various individual characteristics (like height, weight, age, and gender) and their influence on the calories burnt. The algorithm recursively partitions the dataset into subsets based on the feature that optimally increases the information gain, creating a hierarchical structure that reflects the underlying patterns in the data related to calorie expenditure.

Hyper parameter tuning is critical in decision tree modelling for our project to avoid overfitting and enhance model performance. Adjusting parameters like the maximum depth of the tree, minimum samples per leaf, and splitting criteria enables us to fine-tune the model's accuracy and its ability to generalize to unseen data, ultimately improving the precision of calorie burnt predictions.

**3.4 Random Forest Model**

Random forests serve as an ensemble learning technique that combines multiple decision trees to predict calorie expenditure. Each tree in the forest is trained on a random subset of the dataset and independently predicts calorie burn. The final prediction is then obtained by aggregating the predictions of all the trees in the forest.

For our project, random forests can effectively capture complex interactions among different individual characteristics (such as height, weight, age, and gender) and improve model performance through ensemble aggregation. By averaging the predictions of multiple trees, random forests reduce the variance of the model and enhance its resilience to noise and outliers in the data, resulting in more accurate predictions of calorie expenditure.

Feature importance analysis remains a valuable tool in random forest modelling for understanding which features play the most significant roles in predicting calorie burn. By assessing the relative importance of each feature, we can gain insights into the factors driving the predictions and identify the most influential predictors of calorie expenditure, aiding in better understanding and interpretation of the model's outputs.

**3.5 Support Vector Machine (SVM) Model**

* Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. SVM aims to find the hyperplane that best separates the data into different classes or predicts the continuous target variable. SVM is known for its ability to handle high-dimensional data and capture complex relationships in the data.
* In the context of house price prediction, SVM can effectively model the nonlinear relationships between features and house prices. By mapping the input features to a higher-dimensional space using kernel functions, SVM can find the optimal hyperplane that maximizes the margin between different classes or predicts the target variable with minimal error.
* Model evaluation is crucial in SVM modelling to assess the accuracy and generalization ability of the model. Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared are used to measure the model's performance and compare it against other models.

**4. Model Evaluation and Comparison**

**4.1 Evaluation Metrics**

In this section, we discuss the evaluation metrics used to assess the performance of the models. These metrics provide insights into how well the models are performing and help us identify areas for improvement.

* **Mean Absolute Error (MAE)** measures the average absolute difference between the predicted and actual values. It provides a straightforward interpretation of the model's performance, with lower values indicating better accuracy.
* **Root Mean Squared Error (RMSE)** is similar to MAE but penalizes large errors more heavily. It provides a measure of the spread of the errors and is useful for evaluating the overall performance of the model.
* **R-squared (R^2**) measures the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data.

**4.2 Model Comparison**

* We compare the performance of the different models based on the evaluation metrics discussed earlier. This allows us to identify the strengths and weaknesses of each model and determine which model is best suited for house price prediction.
* Linear regression is a simple yet effective model that performs well when the relationship between the features and the target variable is approximately linear. However, it may struggle to capture complex nonlinear relationships in the data.
* Decision trees are versatile models that can capture nonlinear relationships and interactions between features. They are easy to interpret and visualize, making them suitable for exploratory analysis. However, decision trees are prone to overfitting, especially with complex datasets.
* Random forests address the overfitting issue of decision trees by aggregating the predictions of multiple trees. They are robust and perform well on a wide range of datasets. Random forests also provide a measure of feature importance, allowing us to identify the most influential predictors of house prices.
* Support Vector Machine (SVM) is a powerful model that can capture complex nonlinear relationships in the data. It performs well in high-dimensional spaces and is robust to noise and outliers. However, SVMs can be computationally expensive, especially with large datasets.
* Overall, each model has its advantages and limitations, and the choice of model depends on the specific requirements of the problem at hand. In the next section, we discuss the implications of our findings and provide recommendations for further research.

**5. Results**

In this section, we introduce the evaluation metrics for each machine learning model utilized in the "Calories Burnt Prediction" task. These metrics offer valuable insights into the performance of each model and aid in evaluating their accuracy and generalization capability.

**5.1 Linear Regression Evaluation Metrics:**

Results for Linear Regression:

MAE: 3.418894796899015e-14

MSE: 1.9135451568673097e-27

RMSE: 4.3744087107485846e-14

R2 Score: 1.0

* 1. **Decision Tree Evaluation Metrics:**

Results for Decision Tree Regression:

MAE: 0.004666666666666667

MSE: 0.012

RMSE: 0.10954451150103323

R2 Score: 0.999996879169109

* 1. **Support Vector Machine (SVM) Evaluation Metrics:**

Results for SVR:

MAE: 1.714442306748752

MSE: 22.251417817865043

RMSE: 4.717140852027321

R2 Score: 0.9942130906587088

**5.4 Random Forest Evaluation Metrics:**

Results for Random Forest Regression:

MAE: 0.005040000000000002

MSE: 0.02845406666666678

RMSE: 0.16868333250996312

R2 Score: 0.9999925999724809

**6. Conclusion**

In this study, we delved into the task of predicting calories burnt using machine learning algorithms. We examined four distinct models: linear regression, decision tree, support vector machine (SVM), and random forest. Through comprehensive evaluation, we gauged their performance using various metrics.

The findings indicate that the linear regression model surpassed the others in terms of mean absolute error and R-squared score. Nevertheless, all models displayed shortcomings, including elevated mean absolute percentage error and relatively lower R-squared scores, suggesting potential hurdles in accurately predicting calorie expenditure.

Despite these challenges, our study underscores the importance of machine learning in health and fitness analytics. By harnessing advanced algorithms and methodologies, stakeholders can gain valuable insights into calorie expenditure dynamics, facilitating informed decision-making.

In future research, improving model performance and addressing issues such as high mean absolute percentage error will be paramount. Moreover, exploring ensemble techniques and incorporating more sophisticated feature engineering methods could further enhance predictive accuracy.

**7. References**

* ChatGPT. (n.d.). OpenAI. Retrieved from https://openai.com/chatgpt/

* Kaggle. (n.d.). Kaggle: Your Home for Data Science. Retrieved from <https://www.kaggle.com/>
* Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.