# **Project Report**

Name: Aravindh V  
 Course: B.Sc CS (AI & ML) – 2nd Year  
 Registration Number: TU6243202111006

### **Project Title:**

Calorie Burn Prediction Using Physical Activity Data with Linear Regression

## **1. Project Description**

Provide a concise overview of the problem. Explain the real-world context, motivation, and the goal of the project. Mention what the project aims to achieve (e.g., prediction, classification, recommendation, clustering, etc.).

## **2. Learning Objectives**

* Objective 1: Apply regression techniques to health-related data for calorie prediction.
* Objective 2: Evaluate model performance using statistical metrics (R², RMSE).
* Objective 3: Interpret correlations and feature importance to understand the factors influencing calorie burn.

## **3. Timeline**

* Start Date: Oct 1, 2025
* Submission date: Oct 1, 2025

## **4. Algorithm Used**

* Algorithm Name: Linear Regression
* **Explanation:** Briefly describe how the algorithm works, why it was chosen, and its advantages in solving the given problem.

## **5. Tools & Libraries**

* **Programming Language:** Python
* **Libraries Used:**
  + Pandas
  + NumPy
  + Scikit-learn
  + Matplotlib / Seaborn
  + TensorFlow / PyTorch (if used)
  + Joblib

## **6. Dataset Description**

* Source: Kaggle (Calories Burnt Prediction Dataset)
* Size: ~15000 rows, 8–10 features
* Target Variable: Calories
* **Description of Features:** Provide a short description of key features.

## **7. Methodology**

* Data Preprocessing: Missing values handled with mean imputation; categorical encoding for gender; normalization applied.
* Model Training: Dataset split into train (80%) and test (20%); Linear Regression trained on features.
* Evaluation: R² Score, RMSE used.
* Hyperparameter Tuning: Not applied for Linear Regression, but future scope includes Ridge/Lasso.

## **8. Results**

* Performance Metrics: R² Score ~0.82, RMSE ~35 kcal
* Visualizations: Correlation heatmap, scatter plot of actual vs predicted values.
* Insights: Duration and Heart Rate strongly influence calories; age/gender also show measurable effects.

Attach screenshot of outputs and short description

## **9. Questions Answered**

* Q1: Which features affect calorie burn?  
  Q2: How accurate is the model?  
  Q3: What is the correlation matrix?  
  Q4: How to handle missing values?  
  Q5: What is the impact of age/gender?  
  Q6: How to visualize predictions?  
  Q7: What is the role of normalization?  
  Q8: How to evaluate RMSE?  
  Q9: Can we use polynomial regression?  
  Q10: How to deploy the model?  
  Q11: How to split the dataset into training and testing sets?  
  Q12: How to detect and remove outliers from the data?  
  Q13: Can feature selection improve accuracy?  
  Q14: How does cross-validation affect model performance?  
  Q15: How to save and load the trained model for future use?

## **10. Challenges & Improvements**

* Challenges: Missing values, outliers, and Linear Regression limitations in handling non-linear data.
* Future Improvements: Use Random Forest, Gradient Boosting, hyperparameter tuning, and larger datasets.

## **11. References**

* Dataset Links: https://www.kaggle.com/
* Research Papers / Documentation: Scikit-learn Documentation, Seaborn Documentation

## **12. GitHub Link**

https://github.com/aravindhvinayagam2007-crypto/calories