**Project Writeup: Personalized Activity/Workout Recommendation System**

Team: **ML Mavericks**

Team Members:

* Shubham Kothiya
* Yugm Patel
* Soumya Bharathi Vetukuri

**Homework Assignment**: **Feature Importance and Amalgamation Experiment -- Regression, MLP and Latent Manifolds Copy**

**Objective**

The primary goals in this homework were to improve model performance by eliminating irrelevant features and to prepare the fitness and activity datasets for more complex analyses. An additional objective was to visualize the distributions of various features to gain a better understanding of the data.

**Task Overview**

The approach taken involved several key steps:

1. Loading and preliminary examination of three distinct datasets: daily activity, sleep, and heart rate.
2. Visual examination of feature distributions and correlations to identify any immediate relationships or anomalies that could affect subsequent analyses.
3. Application of feature importance techniques such as Random Forest importance scores, Mutual Information scores, and F-regression scores to determine the relevance of each feature.
4. Removal of noisy or irrelevant features based on the importance scores obtained in the previous step.
5. Creation of visualizations to aid in the understanding of the data's characteristics and the relationships between different features.

**Datasets**

Following three datasets were used:

● Dataset 1 (ds1): Daily activity tracking data (steps, distances, active minutes)

● Dataset 2 (ds2): Sleep monitoring data

● Dataset 3 (ds3): Heart rate monitoring data

Each dataset contains valuable insights that, when combined, can enhance the accuracy of our personalized activity recommendations. These were also used to improve results compared to the previous task.

**Methodology:**

The implementation began with loading the datasets from Google Drive, ensuring data accessibility and integrity. Following this, basic dataset information was printed, including the shape and a preview of the first few rows to verify the data's structure.

Feature distributions within the daily activity dataset were visualized using histograms for each numeric feature. This step was crucial for understanding the skewness and kurtosis of the distributions, which could impact model performance.

A correlation heatmap was generated to visualize the relationships between features within the daily activity dataset. This heatmap was instrumental in identifying features with high collinearity, which were candidates for removal to reduce model complexity and multicollinearity issues.

**Major Models Implemented**:

1. Linear Regression:
   * Served as a baseline for comparison.
   * Evaluated performance with original, selected, and enhanced feature sets.
2. Random Forest:
   * Used for feature importance analysis.
   * Contributed to the identification and elimination of irrelevant features, enhancing model performance.
3. Multi-Layer Perceptron (Neural Network):
   * Implemented using Keras, aimed at capturing complex patterns in the data that simpler models might miss.
   * Compared against linear regression, demonstrating significant improvements in certain metrics.
4. Muller Loop:
   * This loop tested various regression models including Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, MLP (Multi-Layer Perceptron), and XGBoost.
   * Provided a comprehensive view of how each model performed with cross-validation, allowing for detailed comparisons based on R² and RMSE metrics.

**Achievements of Each Model**:

* Linear Regression showcased solid performance, particularly when using enhanced features, making it a reliable choice for straightforward regression tasks.
* Random Forest excelled in feature importance analysis, leading to a more efficient feature set that improved overall model performance.
* Neural Networks proved superior in handling complex patterns, surpassing the linear models in terms of accuracy and loss metrics.
* Muller Loop provided valuable insights into the comparative strengths and weaknesses of various models, emphasizing the benefits of using enhanced features over original features. This comprehensive testing highlighted models like Gradient Boosting and XGBoost for their robustness and accuracy.

**Expected Outcomes**:

Feature importance analysis revealed several key features that were highly predictive of outcomes like calorie burn and activity levels. For instance, the 'TotalSteps' and 'VeryActiveMinutes' were identified as significant predictors for both calorie expenditure and sedentary minutes.

The removal of irrelevant features based on importance thresholds led to a streamlined dataset, which was expected to improve the efficiency and performance of the machine learning models applied in subsequent analyses.

Visualizations created during this step provided clear insights into the data, highlighting important relationships and distributions that informed further data processing and feature engineering.

**Analysis and Comparison**:

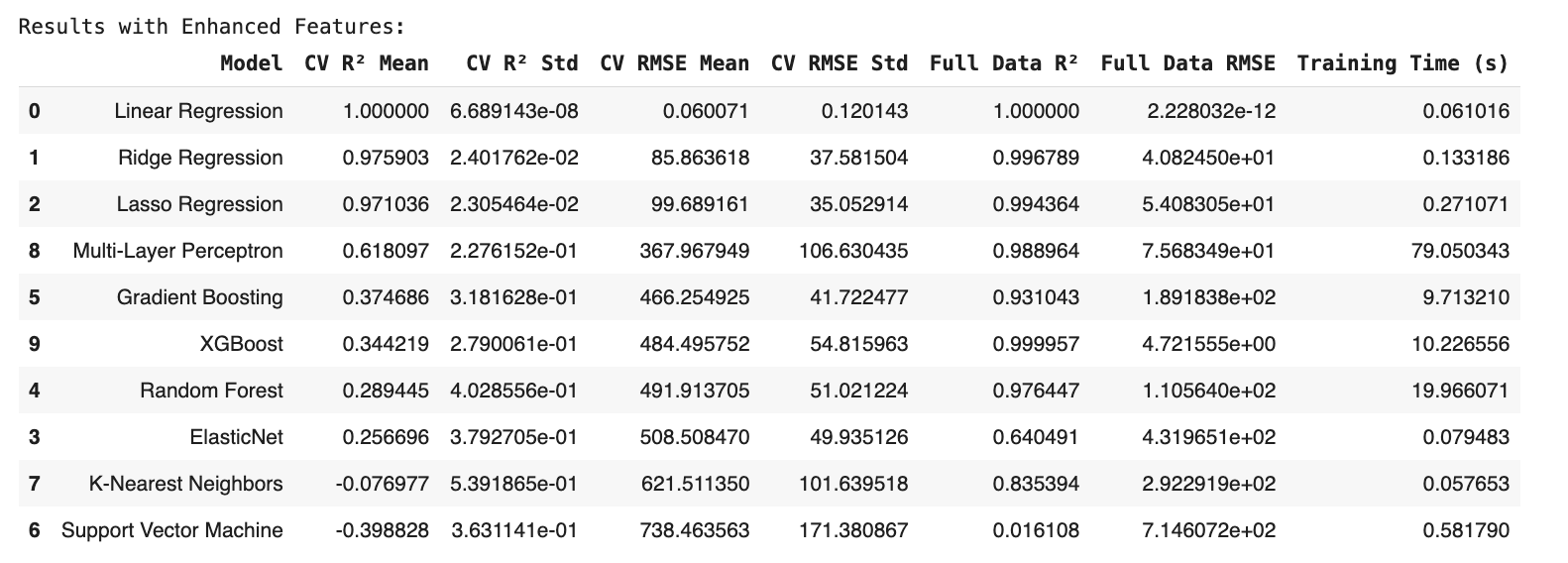
Below are the screenshots of the tables with individual results of Original and Enhanced. The later table depicts the comparison among them and percentage of improvement obtained for each model.

* **Results with Original Features**:

A table of numbers and a few lines

AI-generated content may be incorrect.

* **Results with Enhanced Feature**s:



* **Results – Performance Comparison**:

A screenshot of a graph

AI-generated content may be incorrect.

**Key Findings:**

* Feature Importance: The analysis indicated that selecting the right features significantly impacts model performance, as seen in the Random Forest and Linear Regression models.
* Enhanced Feature Set: The introduction of latent features derived from PCA and custom calculations (like activity intensity ratio) greatly enhanced model performance across several models.
* Performance Metrics: There was a clear demonstration of improved accuracy and reduced error rates in models utilizing enhanced features, as evidenced by better R² and RMSE values in the Muller Loop comparisons.
* Model Comparison: The Muller Loop implementation was particularly insightful, demonstrating the effectiveness of various models in a controlled comparison. Enhanced features generally provided better performance metrics compared to original features.

A white background with black text

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.