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 R^2 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**:

Results with Original Features:

	Model	CV R^2 Mean	CV R^2 Std	CV RMSE Mean	CV RMSE Std	Full Data R^2	Full Data RMSE	Training Time (s)
1	Ridge Regression	0.510666	0.289429	407.396279	83.473691	0.790895	329.439550	0.183159
2	Lasso Regression	0.502875	0.290225	412.682889	95.841397	0.791565	328.911524	0.400847
0	Linear Regression	0.478356	0.257900	431.442695	103.318272	0.792477	328.191470	0.507911
8	Multi-Layer Perceptron	0.278813	0.242552	521.994985	105.854190	0.814707	310.115400	95.242887
5	Gradient Boosting	0.174195	0.501009	530.543706	83.337504	0.847622	281.225300	12.571720
9	XGBoost	0.145029	0.481350	545.293750	101.544123	0.964561	135.624050	6.287483
4	Random Forest	0.135079	0.518293	543.458559	82.236549	0.934810	183.943943	20.708141
3	ElasticNet	0.087373	0.524891	558.703537	60.139162	0.548315	484.185164	0.116918
7	K-Nearest Neighbors	-0.238855	0.623415	666.497725	111.963664	0.727976	375.748363	0.091121
6	Support Vector Machine	-0.400962	0.365365	738.897722	171.282576	0.014597	715.155854	1.210654

- **Results with Enhanced Feature**s:

Results with Enhanced Features:

	Model	CV R^2 Mean	CV R^2 Std	CV RMSE Mean	CV RMSE Std	Full Data R^2	Full Data RMSE	Training Time (s)
0	Linear Regression	1.000000	6.689143e-08	0.060071	0.120143	1.000000	2.228032e-12	0.061016
1	Ridge Regression	0.975903	2.401762e-02	85.863618	37.581504	0.996789	4.082450e+01	0.133186
2	Lasso Regression	0.971036	2.305464e-02	99.689161	35.052914	0.994364	5.408305e+01	0.271071
8	Multi-Layer Perceptron	0.618097	2.276152e-01	367.967949	106.630435	0.988964	7.568349e+01	79.050343
5	Gradient Boosting	0.374686	3.181628e-01	466.254925	41.722477	0.931043	1.891838e+02	9.713210
9	XGBoost	0.344219	2.790061e-01	484.495752	54.815963	0.999957	4.721555e+00	10.226556
4	Random Forest	0.289445	4.028556e-01	491.913705	51.021224	0.976447	1.105640e+02	19.966071
3	ElasticNet	0.256696	3.792705e-01	508.508470	49.935126	0.640491	4.319651e+02	0.079483
7	K-Nearest Neighbors	-0.076977	5.391865e-01	621.511350	101.639518	0.835394	2.922919e+02	0.057653
6	Support Vector Machine	-0.398828	3.631141e-01	738.463563	171.380867	0.016108	7.146072e+02	0.581790

- **Results – Performance Comparison**:

Final Results Table – Muller Loop Performance Comparison:

	Model	Original R^2	Enhanced R^2	R^2 Improvement (%)	Original RMSE	Enhanced RMSE	RMSE Reduction (%)
3	ElasticNet	0.087	0.257	193.8	558.7	508.5	9.0
9	XGBoost	0.145	0.344	137.3	545.3	484.5	11.1
8	Multi-Layer Perceptron	0.279	0.618	121.7	522.0	368.0	29.5
5	Gradient Boosting	0.174	0.375	115.1	530.5	466.3	12.1
4	Random Forest	0.135	0.289	114.3	543.5	491.9	9.5
0	Linear Regression	0.478	1.000	109.0	431.4	0.1	100.0
2	Lasso Regression	0.503	0.971	93.1	412.7	99.7	75.8
1	Ridge Regression	0.511	0.976	91.1	407.4	85.9	78.9
6	Support Vector Machine	-0.401	-0.399	-0.5	738.9	738.5	0.1
7	K-Nearest Neighbors	-0.239	-0.077	-67.8	666.5	621.5	6.7

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 R^2 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.

```
========== STEP 7 SUMMARY ==========
Muller Loop Implementation

Key Findings:
1. The best performing model was Linear Regression with R² = 1.000 and RMSE = 0.1
2. Enhanced features improved R² scores by an average of 90.7%
3. Enhanced features reduced RMSE by an average of 33.3%
4. ElasticNet showed the most improvement with 193.8% better R² when using enhanced features

Conclusion:
- Feature importance and latent feature engineering significantly improved model performance
- Linear Regression and Ridge Regression achieved strong results with enhanced features
- The feature engineering approach succeeded in reducing dimensionality while improving predictive power
- PCA components and activity intensity metrics were particularly valuable features
- The Muller Loop proved valuable for comparing multiple regression approaches
=======================================
```