Machine learning to analyze Fitbit data for predicting states of energy and fatigue.

Predicting Energy and Fatigue Levels Using Machine Learning Techniques on Fitbit Data

Veedhi Solanki*

*Department of Animal Biosciences

University of Guelph, Guelph, Ontario, N1G 2W1, Canada

ABSTRACT

Energy and fatigue levels are crucial indicators of an individual's health and well-being, impacting daily functioning and long-term health outcomes. As society fight with increasing stressors and the prevalence of lifestyle-related health issues, understanding and managing these indicators becomes essential. This study explores the area of health informatics, employing machine learning algorithms to predict energy and fatigue levels using data derived from physiological and activity-based parameters from smartwatches. By leveraging K-Means clustering, we identified distinct behavioral and attribute-based groupings within a Fitbit dataset. Subsequent predictive modeling harnessed the strengths of Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression algorithms, aiming to categorize data into various energy and fatigue levels. Through a rigorous evaluation of model performance via accuracy, precision, recall, and F1-score metrics, we revealed compelling insights. The Random Forest classifier emerged as the superior model, delivering near-perfect accuracy in both training and testing phases, which underscores its powerful generalization capabilities. In contrast, the Logistic Regression model exhibited signs of underfitting, as

evidenced by its diminished accuracy and precision rates. The learning curves indicated a saturation point in model learning, especially for KNN and Logistic Regression, suggesting intrinsic limitations within these models. The study's application of unsupervised machine learning to generate target variables is an effective strategy when such variables are not readily available. The study's outcomes not only highlight the potential of machine learning in health informatics but also emphasize the critical role of model selection and refinement in optimizing predictive performance for energy and fatigue assessment.

Key words:

Energy Level Prediction, Fatigue Detection, Health Informatics, Unsupervised Learning, K-Means

List of abbreviations:

ML: Machine Learning

KNN: K-Nearest Neighbors

RF: Random Forest

LR: Logistic Regression

HRV: Heart Rate Variability

API: Application Programming Interface

CV: Cross-Validation

BMI: Body Mass Index

1. INTRODUCTION

The integration of machine learning (ML) into health informatics has transformed the landscape of predictive analytics in healthcare. These technologies not only enhance our understanding of complex health dynamics but also enable the prediction of individual health states like energy and fatigue from extensive physiological and activity data. Such capabilities are crucial for preemptive

healthcare and personal well-being management, significantly altering approaches to health monitoring and intervention (Patel et al., 2015; Obermeyer & Emanuel, 2016).

Energy and fatigue, though often considered mundane aspects of daily life, have profound implications on personal and professional productivity. Their determinants are multifaceted, encompassing physical activity, sleep quality, mental health, and other physiological factors. These conditions not only affect personal health outcomes but are also pivotal in occupational health, influencing productivity and cognitive functions across different demographics and professions (Smith, 2003; Grandner et al., 2014).

The evolving field of wearable technology has enabled continuous collection of health data, providing a rich source for ML models to analyze and predict health states (Patel et al., 2015). This shift towards data-driven health insights promises to revolutionize personalized medicine by enabling early detection and intervention, potentially mitigating severe health outcomes.

Machine learning, with its capacity to identify patterns and predict outcomes from large datasets, is particularly well-suited for this task. Techniques such as K-Means clustering have shown effectiveness in identifying patterns and segments within health-related data, revealing distinct behavioral and physiological profiles that can inform targeted interventions (Jain, 2010). Furthermore, supervised learning models like Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression are extensively used to predict health states based on real-world data. These models offer varying strengths: Random Forest excels in handling complex, high-dimensional data; KNN benefits from its simplicity and effectiveness in classification tasks; and Logistic

Regression offers model interpretability, crucial for clinical and health informatics applications (Breiman, 2001; Cover & Hart, 1967; Hosmer Jr et al., 2013).

Recent studies have applied these models in clinical settings for predictive diagnostics, showcasing the potential of ML to enhance patient outcomes through personalized treatment plans (Obermeyer & Emanuel, 2016). However, the application of such technologies in personal health management is still emerging. This study seeks to fill this gap by evaluating the efficacy of these ML models in predicting energy and fatigue levels using data from wearable devices. Through testing the data, we aim to identify optimal models and configurations that can reliably predict these health states, thereby contributing valuable insights to the field of personalized health informatics.

The broader implications of this research are vast, extending beyond individual health monitoring to inform public health strategies and occupational health policies. By advancing our understanding of how daily activities and health metrics interplay with overall well-being, we can better design interventions that enhance productivity, improve quality of life, and reduce healthcare costs on a societal level.

This research not only extends the existing literature on machine learning applications in health informatics but also explores the practical implications of these technologies in everyday health management and preventive care, paving the way for future innovations in the field.

2. MATERIALS AND METHODS

Data collection

Data Source and Acquisition

The dataset for this study was sourced from a public repository on Kaggle, specifically designed for health-related analytics using wearable technology data. It was collected through Amazon Mechanical Turk between December 3, 2016, and December 5, 2016. Key files include daily activity, calories, sleep, and heart rate information, with file sizes ranging from 18 kB to 89 MB. The data is publicly available for further research and can be accessed at FitBit Fitness Tracker Data on Kaggle.

Dataset Details

Data includes seconds, minutes, hours to daily level outputs for physical activity, heart rate, and sleep monitoring. Data can be identified and accessed through export session ID or timestamp, reflecting individual usage patterns and device-specific output differences. The comprehensive dataset consists of 18 CSV files of which some files specific with daily data were merged to enhance analytical depth and breadth, encompassing over 29 different recorded variables, such as steps taken, calories burned, and heart rate metrics. The primary file with daily activity, contained the bulk of the activity data, while additional data from files with information such as calories, heart rate, sleep and weight were integrated to ensure completeness without redundancy.

Subjects

The study population comprised thirty human subjects who were users of Fitbit fitness trackers. These individuals consented to share data that included physical activity, heart rate measurements, and sleep monitoring details.

Type of Measurements and Technologies Used

Data collected from the Fitbit devices included physical activity levels, timings, intensity, heart rate data, sleep patterns and durations. These measurements were captured using various models of Fitbit trackers, accounting for variations in data output related to the specific features and capabilities of each model. The diverse tracking behaviors and preferences of the users also contributed to the variability in the data collected.

Data Characteristics

The dataset features included various types of data:

Numerical data: For metrics such as steps taken, heart rate, calories burned.

Time-series data: Timestamped data providing insights into activity intensity and sleep quality over specific intervals.

There were total of 943 rows and 29 columns initially. Overall, the dataset summary provided an overview of the health and activity metrics, including average steps taken (approximately 7,652 daily), distance covered (5.5 km/day), and calories burned (2,308 kcal/day). Sleep records indicate an average of around 7 hours of sleep per night. Average weight is reported as 72 kg with a BMI of 25.2, suggesting a population with varied physical activity levels and sleep patterns, which are essential for analyzing energy and fatigue levels. The data also includes heart rate measurements, with an average heart rate of 83 bpm and an HRV average of 15.64, reflecting a diverse sample in terms of cardiovascular fitness and stress.

Data Handling and Integration

The integration process involved merging information from different files to create a unified dataset with comprehensive details on daily activities, caloric output, intensity of activities, and sleep metrics. The merging process was carefully handled to ensure data integrity and to avoid duplication of information across different files.

Data processing

The data processing section of the study was methodically structured to ensure stringent quality control and data integrity, which are important in predictive modeling.

Quality control

The initial stage involved a thorough evaluation of feature columns within the dataset. So, not important columns were dropped including 'TrackerDistance', 'Fat', 'LogId', 'WeightPounds', 'IsManualReport', and 'ActivityDate'. Next step was to act on missing values. Columns with a substantial number of missing entries, particularly those exceeding 50% missing data, were identified for potential exclusion. This step was crucial to maintain the dataset's integrity. However, most data with missing values were important such as weight, heart rate and sleep data, so, imputation with mean values was performed. This method was appropriate for preserving valuable data points without introducing significant bias.

To uphold data consistency and integrity, attention was focused on data columns that presented relevance and consistency, such as 'TotalSteps', 'TotalDistance', 'WeightKg' and 'Calories' as shown in **figure 1**. The **figure 1 (a)** and **(b)** for total steps and total distance graphs shows ddirect correlation and has an S-shaped curve that rises steeply in the middle, showing that most individuals accumulate a moderate number of steps, with fewer reaching very high or very low step counts. The **figure 1 (c),** shows calories graph gradually ascends, suggesting a wide spread of caloric expenditure with most values falling in the middle of the range, pointing to moderate daily calorie burn for the majority. The **figure 1 (d),** distribution shows a steep, almost vertical rise at a certain point, indicating a narrow, prevalent weight range among the majority of the dataset, with sharp declines in frequency at the higher and lower ends, reflecting less common weights. This

was due to the imputation.

A correlation matrix was constructed, providing insight into the interdependencies between variables and ensuring there were no inconsistencies or redundancies that could compromise the analyses as shown in **figure 2.** This matrix was also used to drop features during feature selection.

Furthermore, normalization and scaling techniques were applied to features with considerable numerical range disparities. The use of MinMaxScaler assured that no single attribute would disproportionately influence the model's performance due to scale variations.

Each quality control measure was carefully documented within the Python code, using a transparent and replicable data cleaning process that underpins the integrity of the subsequent machine learning analyses.

Data organization

Data organization involved the careful arrangement of collected datasets to ensure their suitability for the intended analyses. All the Physical activity type, distance and minute data coloumns were kept together. Following that all data on weight, calories, sleep and heart rate data were kept together. This organization provided logical organization of features.

The variety of features were engineered to enrich the dataset further, including sleep efficiency, sleep quality, total active minutes, activity diversity and heart rate spread.. These engineered features were thoughtfully designed to encapsulate the multi-dimensional nature of physical activity and rest and their impacts on health outcomes. New features were also normalized using

the MinMaxScaler and StandardScaler. This standardization ensured that all features contributed equally to the models, preventing any one feature with a large range from disproportionately influencing the model's predictions.

The culmination of these organizational efforts resulted in a clean, comprehensive, and structured dataset, paving the way for the next phases of machine learning modeling and analysis.

Problem definition

The primary objective of this study was to investigate the feasibility of using machine learning techniques to predict individual energy levels and fatigue based on daily health metrics collected from smartwatches. The problem was approached as both a clustering and classification challenge, with the ultimate goal of enhancing personalized health monitoring and intervention strategies.

Inputs: The dataset utilized as input comprised various health metrics recorded daily, including total steps, total distance traveled, sleep duration, heart rate averages, and calories burned, among others. These inputs are derived from the raw data collected via Fitbit trackers, which continuously monitor and log health-related activities.

Outputs: The outputs for this study were twofold:

Clustering: Initially, the dataset was subjected to a clustering analysis using the K-Means algorithm to identify inherent groupings within the data based on health behaviors and physiological metrics. The threshold of clusters helped to define energy and fatigue levels. This unsupervised learning approach helped in uncovering patterns and categorizations in the data without predefined labels, providing insights into different health and activity profiles within the population.

The combination of clustering and classification approaches allowed the study to both explore the natural groupings within the data and develop predictive models capable of classifying new data points into these energy and fatigue categories effectively. This methodology underscores a comprehensive approach to problem-solving in health informatics, emphasizing the utility of machine learning in enhancing predictive accuracy and personalized health management.

Statistical analyses

The statistical analyses in this study were designed to thoroughly assess and understand the patterns and relationships within the health metrics data collected from Fitbit devices. Here are the key statistical methods used to analyze and characterize the data:

1. Descriptive Statistics: Initially, basic descriptive statistics such as mean, median, standard deviation, minimum, and maximum values were calculated for each variable to understand the central tendencies and dispersions. This provided a foundational understanding of the data's characteristics, including the distribution and range of health metrics like steps, sleep duration, and heart rate.

2. Correlation Analysis: To explore the relationships between different health metrics, a correlation matrix was computed. This helped identify which variables had strong associations with each other, which is crucial for understanding the interdependencies within the data and for the subsequent feature selection process in model building.

3. Variance Inflation Factor (VIF) Analysis: Before proceeding with predictive modeling, the Variance Inflation Factor was calculated for each feature to detect multicollinearity. Features with a high VIF were considered for removal or adjustment, as multicollinearity can undermine the statistical significance of an independent variable and inflate the standard errors of the coefficients. No Specific threshold was used to remove features because base features were correlating with

newly created features, so, first infinite VIF value features were removed and the base features were removed while keeping newly created features.

4. Cluster Analysis: K-Means clustering was applied to segment the dataset into distinct groups based on similar health behavior patterns. The optimal number of clusters was determined using the elbow method, which involves plotting the sum of squared distances from each point to its assigned center and identifying the 'elbow' point where the rate of decrease sharply shifts.

5. Feature Engineering: New features were derived from existing data to enhance the models' performance. This included ratios, aggregated summaries (like total active minutes), and interaction terms that might be more predictive of energy and fatigue levels than the original data.

6. Model Evaluation Metrics: Classification models were evaluated using accuracy, precision, recall, and F1-score. These metrics provided a comprehensive view of model performance, especially in handling imbalanced class distributions which is common in health data.

7. Learning Curves: To evaluate the effect of training size on the performance of the models, learning curves were plotted. These curves helped in diagnosing whether the models suffer from high bias or variance, guiding further tuning and iteration.

Each of these analytical steps contributed to a rigorous examination of the data, facilitating the development of robust models capable of predicting individual energy levels and fatigue based on daily health metrics. These methods not only supported the primary research objectives but also ensured that the findings were grounded in statistically sound principles.

Machine learning modelling

ML Algorithms:

1. Random Forest: The Random Forest algorithm was selected for its robustness in handling high-

dimensional data and its ability to produce a reliable classification by averaging the results of numerous decision trees constructed at training time. Its effectiveness lies in reducing overfitting while maintaining high accuracy on the dataset. It operates by building multiple decision trees and merging their outputs to improve general accuracy and control over-fitting, which is particularly effective in complex datasets like the ones used in this study [Breiman, 2001].

2. K-Nearest Neighbors (KNN): The KNN algorithm functions on a simple principle of feature similarity, where the prediction for a new instance is derived by majority vote of its 'k' nearest neighbors, based on a distance metric like Euclidean distance. This algorithm is inherently non-parametric and lazy attributes that make it uniquely flexible to the dataset's nuances without making assumptions about the underlying data structure [Cover & Hart, 1967].

3. Logistic Regression: Employed primarily for binary classification tasks, Logistic Regression was utilized to estimate discrete outcomes (e.g., high vs. low energy levels) from a combination of input variables. By applying a logistic function, this model is capable of providing probabilities that describe the possible outcomes for individual cases, which is invaluable for decision-making processes in health monitoring applications [Hosmer Jr, D.W., Lemeshow, S., & Sturdivant, R.X., 2013].

Model Evaluation:

1. Accuracy, Precision, Recall, and F1-Score: Comprehensive metrics such as accuracy, precision, recall, and F1-score were rigorously calculated to evaluate each model's performance across various aspects of classification effectiveness. These metrics are crucial for assessing the trade-offs between true positive rates and false positive rates, particularly important in medical diagnostic processes where the cost of different types of errors varies greatly.

2. Train-Test Split and Cross-Validation: Data was split into training and testing sets to validate

the model's performance against unseen data. Additionally, techniques like k-fold cross-validation were used where the dataset is divided into k-subsets and the model is trained on k-1 of these subsets, tested on the remaining parts. This method helps in reducing bias and variance in the model evaluation process.

Model Overfitting:

1. Hyper-parameter Optimization: Grid Search CV was extensively used to systematically work through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance. This not only aids in optimizing the algorithm parameters but also prevents overfitting, ensuring that the models generalize well to new data.

2. Learning Curves: Learning curves were plotted to diagnose variance and bias in the models. Observing how model error changes as the training set size increases provided insights into whether adding more data, reducing model complexity, or choosing a different modeling algorithm might improve performance.

Feature Importance:

1. Permutation Importance: After the models were trained, permutation importance was utilized to identify the impact of each feature on the prediction accuracy. This method involved randomly shuffling each feature and observing how much the model's accuracy decreased. The more significant the decrease, the more important the feature is considered for the model. This analysis is particularly useful for interpreting the model, giving insights into which variables are driving the prediction process and potentially guiding further data collection and feature engineering efforts.

These detailed methodologies ensured a robust analytical framework for the study, allowing for

comprehensive insights into the predictive modeling of health-related outcomes based on data from wearable devices.

3. RESULTS AND DISCUSSION

The results and discussion critically examine analytics of energy and fatigue levels using machine learning models applied to Fitbit data. It will encompass the outcomes from K-Means clustering to discern distinct behavioral clusters, followed by a comparative analysis of three predictive models: Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. Each model's efficacy is evaluated through a meticulous assessment of accuracy, precision, recall, and F1-scores. Additionally, the potential overfitting and generalization capacity of these models are scrutinized using learning curves. Through this comprehensive evaluation, we aim to conclude the strengths and limitations of each approach, thereby highlighting the most effective methodologies for predicting states of energy and fatigue from wearable device data.

Clustering Analysis: The application of K-Means clustering on the scaled features of the dataset led to the determination of the optimal number of clusters via the elbow method. The elbow plot, as shown in **figure 3**, indicates a distinct bend at k=5, suggesting that five clusters provide a reasonable balance between the within-cluster sum of squares and the number of clusters, thus minimizing within-cluster variance while avoiding overfitting.

Upon clustering, cluster 0 is characterized by moderate activity levels, high caloric burn, and average sleep efficiency, suggesting a balanced energy expenditure with insufficient sleep. Cluster 1 displays high distances traveled and calorie consumption, accompanied by low sleep sufficiency, implying a highly active lifestyle with potential sleep deficits. Cluster 2 exhibits low activity and

caloric burn, indicating a sedentary profile with average sleep patterns. Cluster 3 is marked by moderate distances, activity, and caloric burn, but with high sleep efficiency and sufficiency, suggesting a healthy balance between activity and rest. Cluster 4 shows moderate activity levels but low calories and very poor sleep efficiency, hinting at inefficient energy use or potential health issues.

For energy levels, the criteria incorporated both 'TotalActiveMinutes' and 'Calories' to define a spectrum of energy states from 'Very High Energy' to 'Low Energy.' This spectrum considered not only the quantity of activity but also the caloric expenditure associated with those activities, which could provide a more comprehensive view of an individual's energy status. The inclusion of 'HRV' (Heart Rate Variability) as a physiological marker added another layer to the classification, distinguishing between high energy due to physical activity and very high energy signified by both activity and favorable physiological conditions. Similarly, for fatigue levels, 'SleepEfficiency' was used as a primary indicator, with different thresholds set for 'High Fatigue' and 'Somewhat Chance of Fatigue.' The conditions of 'HRV' and 'Heart rate Average' were utilized to capture the nuances in the 'Low Fatigue' category, linking physiological stress and restfulness to the concept of fatigue.

The K-Means clustering results present an opportunity to understand lifestyle patterns and their implications on wellness. The clear delineation of clusters based on activity, sleep, and HRV allows for targeted health recommendations. For instance, individuals in clusters with 'Very High Energy' but low sleep sufficiency might benefit from advice on better sleep hygiene to maintain sustainable health behaviors. Meanwhile, the presence of a cluster with low energy expenditure and poor sleep efficiency points to a potentially at-risk group that may require medical evaluation

for underlying conditions such as sleep disorders or metabolic inefficiencies. This analysis highlights the power of unsupervised learning to uncover hidden patterns in lifestyle data.

Random Forest Model: The Random Forest classifier demonstrated high performance on the prediction of Energy Levels, with a training and testing accuracy of about 99% as shown in **figure 5 and 6,** . For Fatigue Levels, the model similarly exhibited high training and testing accuracies, at 99% and 98% respectively. The GridSearchCV process for parameter optimization determined that the best parameters for the Energy Level model were 'max_depth' at None, 'min_samples_leaf' at 2, 'min_samples_split' at 2, and 'n_estimators' at 100. The optimized model retained 100% accuracy on the test data. For the Fatigue Level model, the best parameters were 'max_depth' at 10, 'min_samples_leaf' at 1, 'min_samples_split' at 5, and 'n_estimators' at 200, resulting in a test accuracy of about 98%. The feature importance analysis (**figure 5 a and b**) revealed key variables contributing to the predictions. For Energy Levels, 'TotalDistance', 'TotalActiveMinutes', 'InactiveMinutes' and 'Calories' were amongst the most influential features. In contrast, 'HRV', and 'Heart rate Average' were pivotal for predicting Fatigue Levels. The learning curves (**figure 4 and b**) for both models plateau at high levels of training and cross-validation scores, suggesting good model generalization with minimal overfitting.

The model's performance suggests that the selected features robustly capture substantial information about individual energy states. The high test accuracy indicates effective learning and generalization to new data, without signs of overfitting. The slightly lower accuracy in predicting Fatigue Levels, as compared to Energy Levels, may be due to the more nuanced nature of fatigue as a variable but it can be improved by refining feature selection or further tuning the model

parameters. Despite this, the model's performance remains robust and indicates its utility in practical applications. The careful tuning of hyperparameters reflects the balance between model complexity and generalization. By optimizing the Random Forest's depth and the minimum samples for splitting and creating leaf nodes, we avoided overfitting, ensuring the model's high performance on unseen data. The results from this process indicate that while default parameters provide a strong baseline, fine-tuning is essential for maximizing performance. The feature importance underscores the relevance of activity-related features for energy and sleep-related features for fatigue. These insights align with expectations that higher physical activity correlates with increased energy levels, while sleep quality and duration are crucial for fatigue assessment. The model's reliance on these features affirms their physiological significance. The convergence of training and validation scores at high levels indicates that the models are likely to perform well on new data. The absence of a significant gap between the training and testing curves suggests that the models are not suffering from high variance, which can be a common issue with complex models such as Random Forest.

K-Nearest Neighbors (KNN) Model: The KNN classifier with an improved parameter setting reached a training accuracy of approximately 82% for Energy levels, which only slightly increased to 83% in the testing phase. The KNN model for Fatigue levels presented lower accuracy scores, with 78% for training and 71% for testing (**figure 6**). GridSearchCV identified the optimal parameters for both Energy and Fatigue level predictions to be 5 neighbors with uniform weights and a p-value of 2 (euclidean distance). Post-optimization, the models displayed a marginal improvement for Energy levels, achieving a testing accuracy of 81%, and a more noticeable enhancement for Fatigue levels, reaching a testing accuracy of 71% (**figure 6**). The learning curve

plots (**figures 4 c and d**) show a progressive increase in the cross-validation scores with an increase in the training examples, eventually plateauing. The training scores remain relatively constant, indicating a consistent model across the training sizes.

The distribution across energy levels indicates a central concentration, with room to enhance model performance at the spectrum's extremes.The lower accuracies for Fatigue level predictions indicate that the chosen features and KNN parameters may not capture the complexity of the fatigue states as effectively as they do for energy states. The tuning of hyperparameters through GridSearchCV resulted in slightly improved performance, particularly for Fatigue levels. The inherent simplicity of the KNN algorithm, while being its strength, might also be a limitation in capturing complex relationships, suggesting the need for a more sophisticated model or ensemble methods to improve performance further. The learning curves suggest that adding more data points may not significantly improve model performance, indicating that the current model complexity is suitable for the given data. However, since the cross-validation score is lower than the training score, especially in the case of Fatigue levels, it hints at the model's limitations in generalizing, potentially due to the simplicity of the algorithm. In conclusion, the KNN models for predicting Energy and Fatigue levels perform moderately well, but with room for improvement. The analyses suggest that while KNN is capable of capturing trends in the data, more complex patterns may require advanced modeling techniques or data adjustments to enhance predictive accuracy.

Logistic Regression: For the energy level prediction, logistic regression yielded an accuracy of about 70% on the training set and 71% on the test set. These results were obtained after iterating through a maximum of 1000 times to ensure model convergence. In the case of fatigue level

prediction, the model achieved a higher accuracy of 82% on the training data and 77% on the testing data. The default parameter models were compared against models with best parameters obtained from a grid search. The best parameters for the energy level included {'n_neighbors': 5, 'p': 2, 'weights': 'uniform'}, which interestingly reflect the default parameters of a KNeighborsClassifier, suggesting that for this dataset, the default KNN parameters were close to optimal. For the fatigue level, logistic regression with default parameters achieved 82% accuracy on the training set and 77% accuracy on the testing set, again showing the robustness of the default settings. The classification report for the energy level on the testing set indicates high precision and recall across most classes, with perfect precision and recall for class '5', which denotes the 'Very High Energy' category. However, there were notable discrepancies in class '0', indicating 'Low Energy', which had the lowest precision and recall rates of 10% and 17% respectively. The fatigue level's classification report presented a more varied performance across classes, with the 'No Fatigue' class achieving high precision but low recall, indicating a possible imbalance in the dataset or the model's sensitivity towards this class.

The logistic regression model's performance for energy level prediction is satisfactory, but there is room for improvement, especially in the lower energy classes. The difference in performance between the training and test sets is relatively small, which suggests that the model has generalized well and is not overfitting. The model's performance on fatigue level prediction is strong and indicates that logistic regression is a viable option for this type of binary classification problem. However, the performance discrepancy between training and testing suggests that the model might benefit from additional feature engineering or data collection, especially to improve its recall in certain classes. The confusion matrices corroborate the classification reports, revealing that the

majority of predictions fall on the diagonal, indicating correct classifications. However, misclassifications are present and are more prevalent in certain classes, as indicated by the off-diagonal numbers in the matrices. The results indicate that logistic regression, even with default parameters, can be an effective model for this type of health-related classification task. However, it also suggests that careful consideration of class balance and further model tuning could lead to improved performance, particularly in underperforming classes. Given these findings, future work should explore more complex models, feature selection techniques, and over-sampling methods to address class imbalance, potentially leading to improvements in model accuracy and reliability.

The performance of the three machine learning models—K-Means Clustering, Random Forest, and K-Nearest Neighbors (KNN), as well as Logistic Regression—has been evaluated based on accuracy, precision, recall, and f1-score for both energy and fatigue level predictions. We will also discuss the degree of overfitting or underfitting as revealed by the learning curves.

Model Accuracy Comparison: The comparative accuracy figure illustrates that Random Forest achieved the highest accuracy across both training and testing datasets for energy and fatigue as in **figure 6**. The GridSearch optimization did not significantly affect the testing accuracy, which remained consistently high, reflecting the model's robustness. Logistic Regression exhibited the lowest accuracy among the models. While Random Forest outperforms in accuracy, it's crucial to consider the potential overfitting as its training accuracy reaches 100%. The learning curves suggest that while the training score remains high, the cross-validation score plateaus, indicating a perfect fit to the training data but less so to the validation set. Logistic Regression, while being the least accurate, displays a closer performance between training and testing, indicating a better

generalization than KNN, which has a noticeable drop in testing accuracy.

Model Precision and Recall: The evaluation matrix figure for f1, recall, precision, and accuracy as in **figure 7**, demonstrates that Random Forest not only maintains high accuracy but also achieves commendable precision and recall, particularly in the energy level predictions. This suggests a balanced ability to identify true positives while minimizing false positives and negatives. KNN and Logistic Regression show varied performance, with KNN struggling with recall and Logistic Regression with precision. A high recall in Random Forest for fatigue level indicates its strength in identifying all relevant cases within the dataset, which is vital for applications where missing a positive case has a higher cost. However, the precision trade-off indicates a relatively higher false-positive rate, which could be problematic in a different context. KNN's lower recall implies a higher instance of false negatives, which can be critical if the model is applied in a scenario where failing to identify true cases is costly.

Model Fit and Generalization: The learning curves reveal that Random Forest and KNN tend to overfit, as evidenced by a gap between training and ttesting scores. Logistic Regression shows a more consistent performance, suggesting a better fit without large variances between training and validation scores. The learning curves are indicative of the model's ability to generalize to new data. Logistic Regression shows the most promise in this area, despite lower overall accuracy, which might be improved with feature engineering or alternative regularization techniques. Random Forest, while excellent in performance, requires careful cross-validation to ensure it does not overly fit the training data when deployed in a real-world setting.

Overall, while Random Forest shows exceptional performance metrics, the potential overfitting highlighted by the learning curves must be addressed in deployment. Logistic Regression offers a balanced fit and may serve as a reliable model with further tuning, while KNN's performance suggests it may not be as suitable for this dataset's complexities.

4. CONCLUSIONS

The exploration of machine learning techniques in predicting energy and fatigue levels from Fitbit data has revealed significant insights and highlighted the potential of wearable technology in health informatics. Our study has meticulously compared the performance of K-Means Clustering, Random Forest (RF), K-Nearest Neighbors (KNN), and Logistic Regression (LR) algorithms on a dataset encapsulating various health metrics.

The Random Forest algorithm has emerged as the standout performer, demonstrating near-perfect accuracy in classifying energy and fatigue levels. The models' precision, recall, and f1-scores, particularly for RF, indicate a highly capable system that could be instrumental in health monitoring and preemptive healthcare strategies. Nevertheless, the learning curves suggest a potential for overfitting in the RF models, underscoring the importance of deploying them with caution and the need for continuous validation.

The Logistic Regression model showed commendable generalization capabilities despite its lower accuracy. The KNN algorithm, while straightforward and easy to implement, showed limitations in handling the complexity inherent in the data, reflected in its lower recall rates and the decline in performance from training to testing datasets.

From these findings, we conclude that the application of machine learning algorithms can significantly contribute to the prediction of health states using wearable device data. However, the selection of the appropriate model is critical and should be based on the desired balance between accuracy and the model's ability to generalize. This study confirms the value of employing a combination of machine learning techniques to achieve an optimal predictive system.

Furthermore, the study has broad implications for personalized health interventions and public health strategies. By harnessing the power of machine learning and wearable technology, there is an opportunity to enhance individual well-being and contribute to the collective health of the community.

In future research, we recommend further investigation into hybrid models or ensemble methods to combine the strengths of individual algorithms. Additionally, the exploration of more granular data and alternative feature engineering approaches may yield improved predictive power, especially for categories that are currently challenging to predict with high precision.

The findings from this study serve as a testament to the promise of health informatics and machine learning in advancing the goals of predictive healthcare. By continuing to refine these models and adapt them to the nuances of individual health patterns, we move closer to the vision of a proactive, data-informed healthcare paradigm.

ACKNOWLEDGEMENTS

Dr. Dan Tulpan

AUTHORS' CONTRIBUTIONS

NA

DISCLOSURES

The authors declare no real or perceived conflicts of interest.

LITERATURE CITED

1. Smith, L. L. (2003). Overtraining, excessive exercise, and altered immunity: is this a T helper-1 versus T helper-2 lymphocyte response?. Sports medicine, 33, 347-364.Grandner, M.A., et al., "Sleep Symptoms Associated with Intake of Specific Dietary Nutrients," Journal of Sleep Research, vol. 23, no. 1, 2014, pp. 22–34.
2. Alpaydin, E. (2020). Introduction to machine learning. MIT press.
3. Patel, M. S., Asch, D. A., & Volpp, K. G. (2015). Wearable devices as facilitators, not drivers, of health behavior change. Jama, 313(5), 459-460.
4. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216-1219.
5. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.
6. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
7. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.
8. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.4Cover, T. & Hart, P. (1967). Nearest Neighbor Pattern Classification.

FIGURES

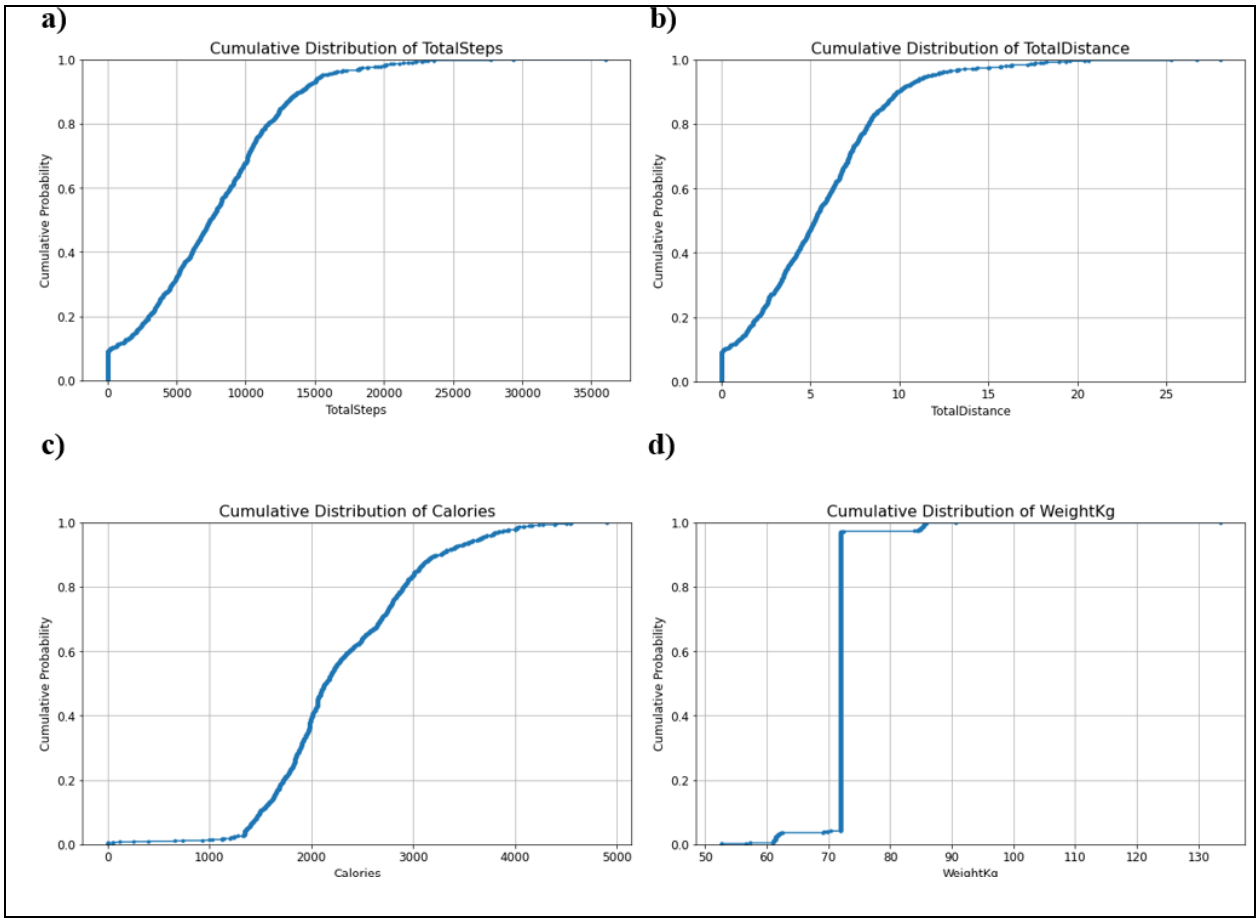


Figure 1: Comparative Cumulative Distributions of Health Metrics from Wearable Device Data: This figure presents the cumulative probability distributions for key health-related metrics: **(a)** Total Steps, **(b)** Total Distance, **(c)** Calories, and **(d)** Weight in kilograms, showcasing the data's dispersion and central tendencies.

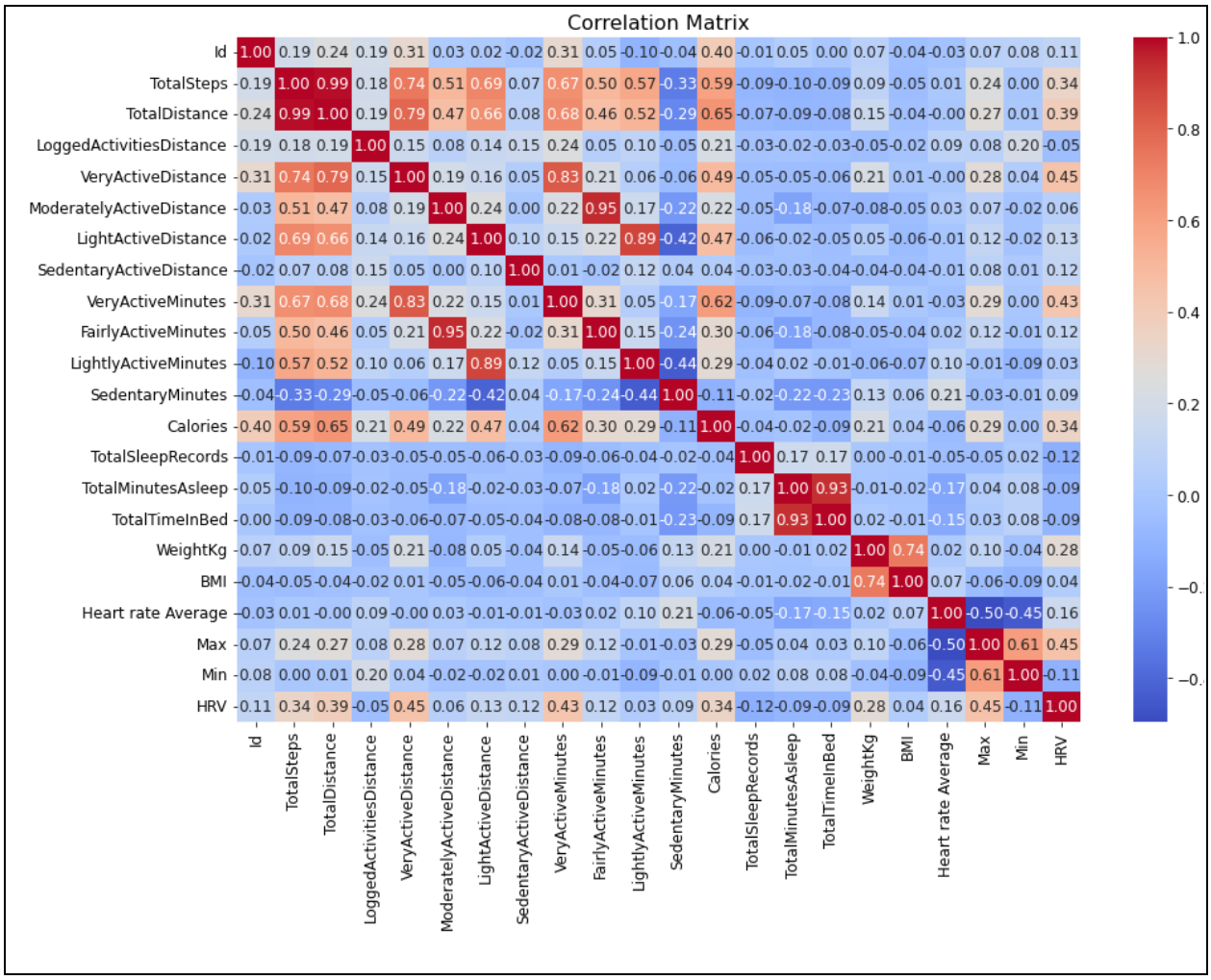


Figure 2: Heatmap of Correlation Coefficients among Multiple Variables. This figure presents a matrix of correlation values ranging from -1 to 1, with 1 indicating a perfect positive correlation, -1 indicating a perfect negative correlation, and 0 signifying no correlation. Each cell represents the correlation between the variables on the corresponding row and column, with red cells highlighting strong positive correlations, blue cells indicating strong negative correlations, and white cells denoting weak or no correlation.

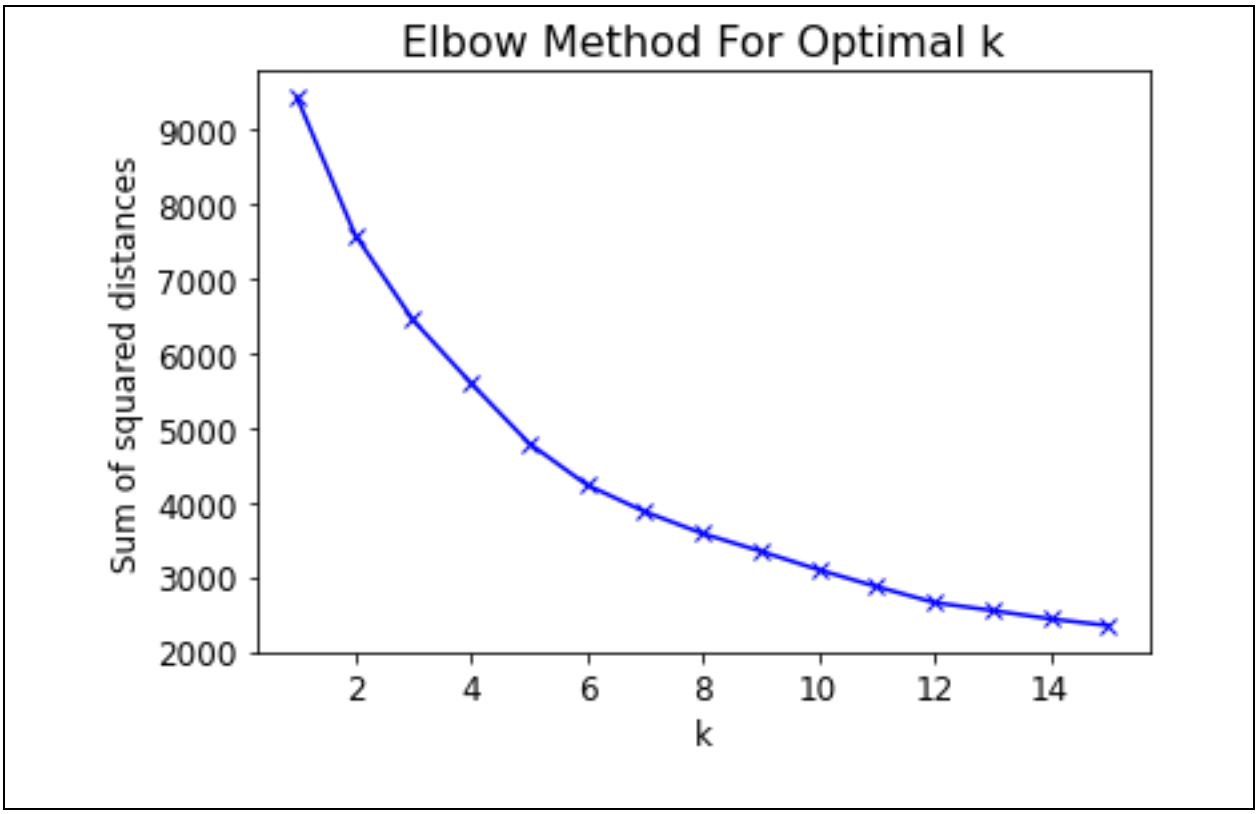


Figure 3 : Elbow plot for determining the optimal number of clusters in KMeans clustering. The bend at k=5 suggests this as the optimal cluster count for minimizing within-cluster variance without overfitting.

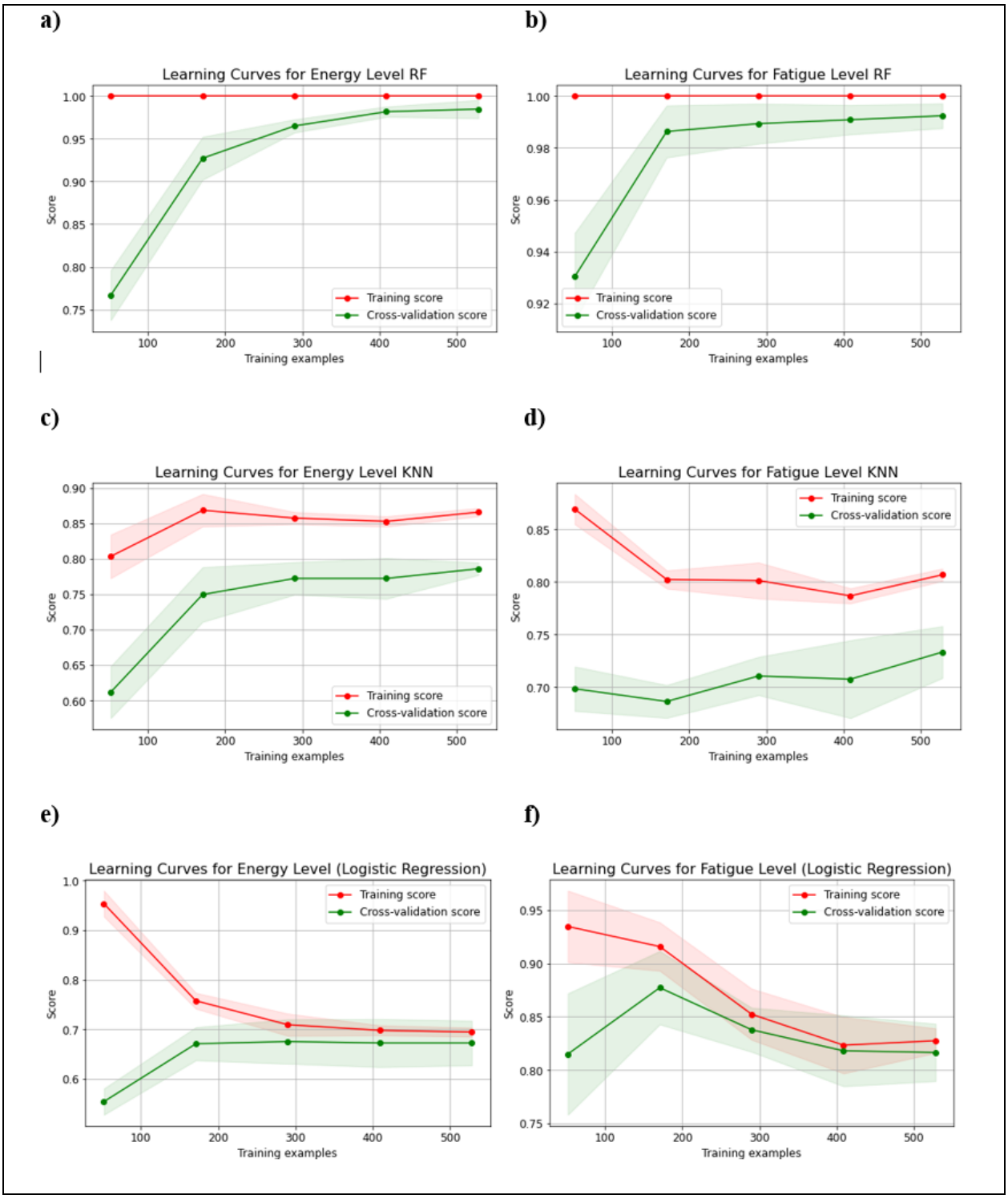
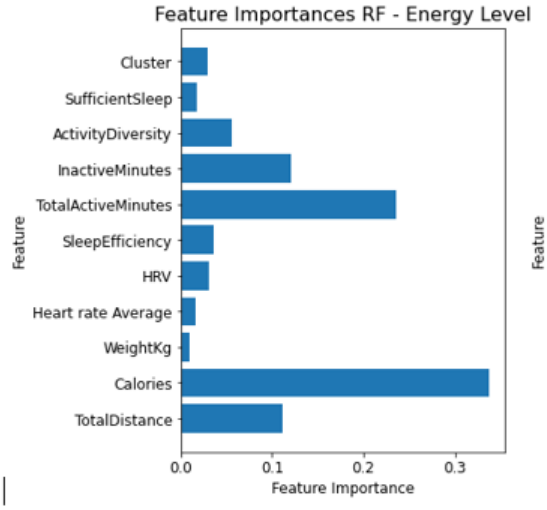
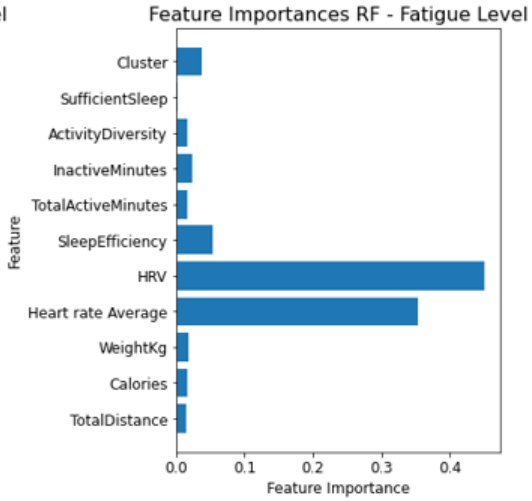


Figure 4: Learning curves for energy and fatigue level predictions using Random Forest **(a, b)**, K-Nearest Neighbors **(c, d)**, and Logistic Regression **(e, f)**, illustrating the models' performance in terms of training and cross-validation scores as the number of training examples increases.

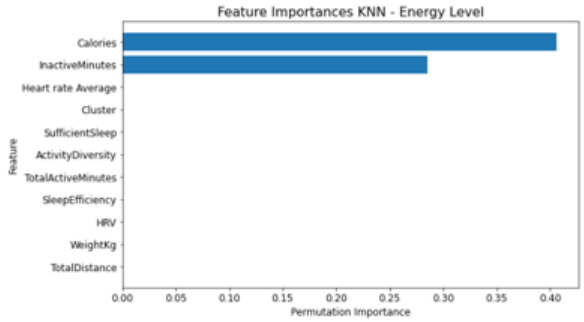
a)


Feature Importances RF - Energy Level
Cluster
SufficientSleep
ActivityDiversity
InactiveMinutes
TotalActiveMinutes
SleepEfficiency
HRV
Heart rate Average
WeightKg
Calories
TotalDistance
Feature
0.0
0.1
0.2
0.3
Feature Importance

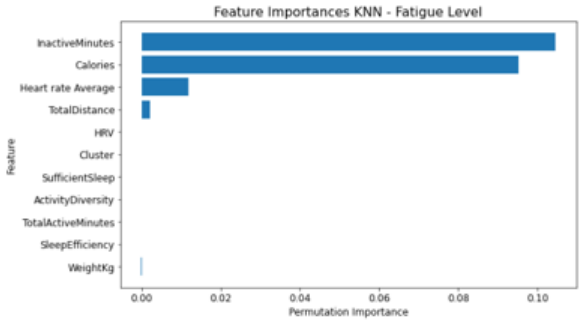
b)


Feature Importances RF - Fatigue Level
Cluster
SufficientSleep
ActivityDiversity
InactiveMinutes
TotalActiveMinutes
SleepEfficiency
HRV
Heart rate Average
WeightKg
Calories
TotalDistance
Feature
0.0
0.1
0.2
0.3
0.4
Feature Importance

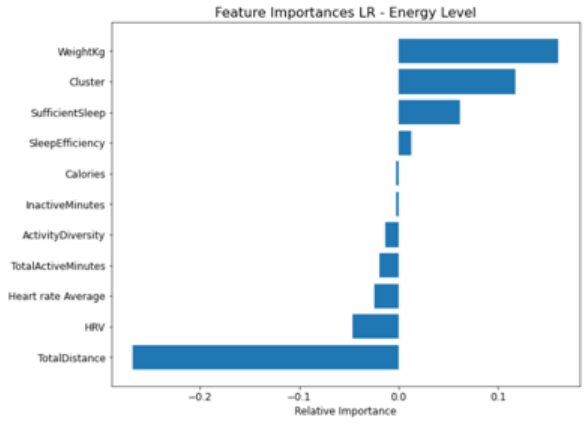
c)


Feature Importances KNN - Energy Level
Calories
InactiveMinutes
Heart rate Average
Cluster
SufficientSleep
ActivityDiversity
TotalActiveMinutes
SleepEfficiency
HRV
WeightKg
TotalDistance
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Permutation Importance

d)


Feature Importances KNN - Fatigue Level
InactiveMinutes
Calories
Heart rate Average
TotalDistance
HRV
Cluster
SufficientSleep
ActivityDiversity
TotalActiveMinutes
SleepEfficiency
WeightKg
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Permutation Importance

e)


Feature Importances LR - Energy Level
WeightKg
Cluster
SufficientSleep
SleepEfficiency
Calories
InactiveMinutes
ActivityDiversity
TotalActiveMinutes
Heart rate Average
HRV
TotalDistance
−0.2
−0.1
0.0
0.1
Relative Importance

f)

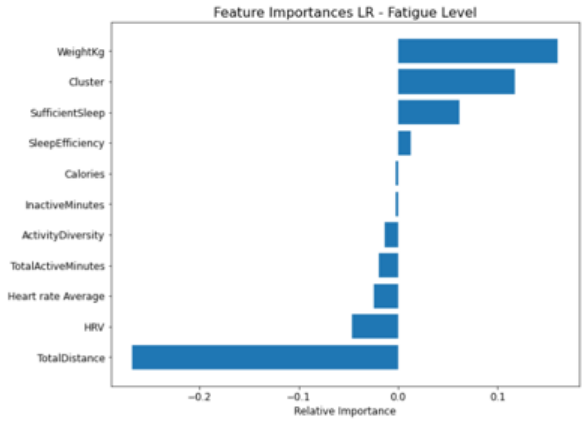

Feature Importances LR - Fatigue Level
WeightKg
Cluster
SufficientSleep
SleepEfficiency
Calories
InactiveMinutes
ActivityDiversity
TotalActiveMinutes
Heart rate Average
HRV
TotalDistance
−0.2
−0.1
0.0
0.1
Relative Importance

Figure 5: Bar charts comparing the feature importance in predicting energy and fatigue levels using Random Forest **(a, b)**, K-Nearest Neighbors **(c, d)**, and Logistic Regression **(e, f)** algorithms, highlighting the variables most critical to each model's predictions.

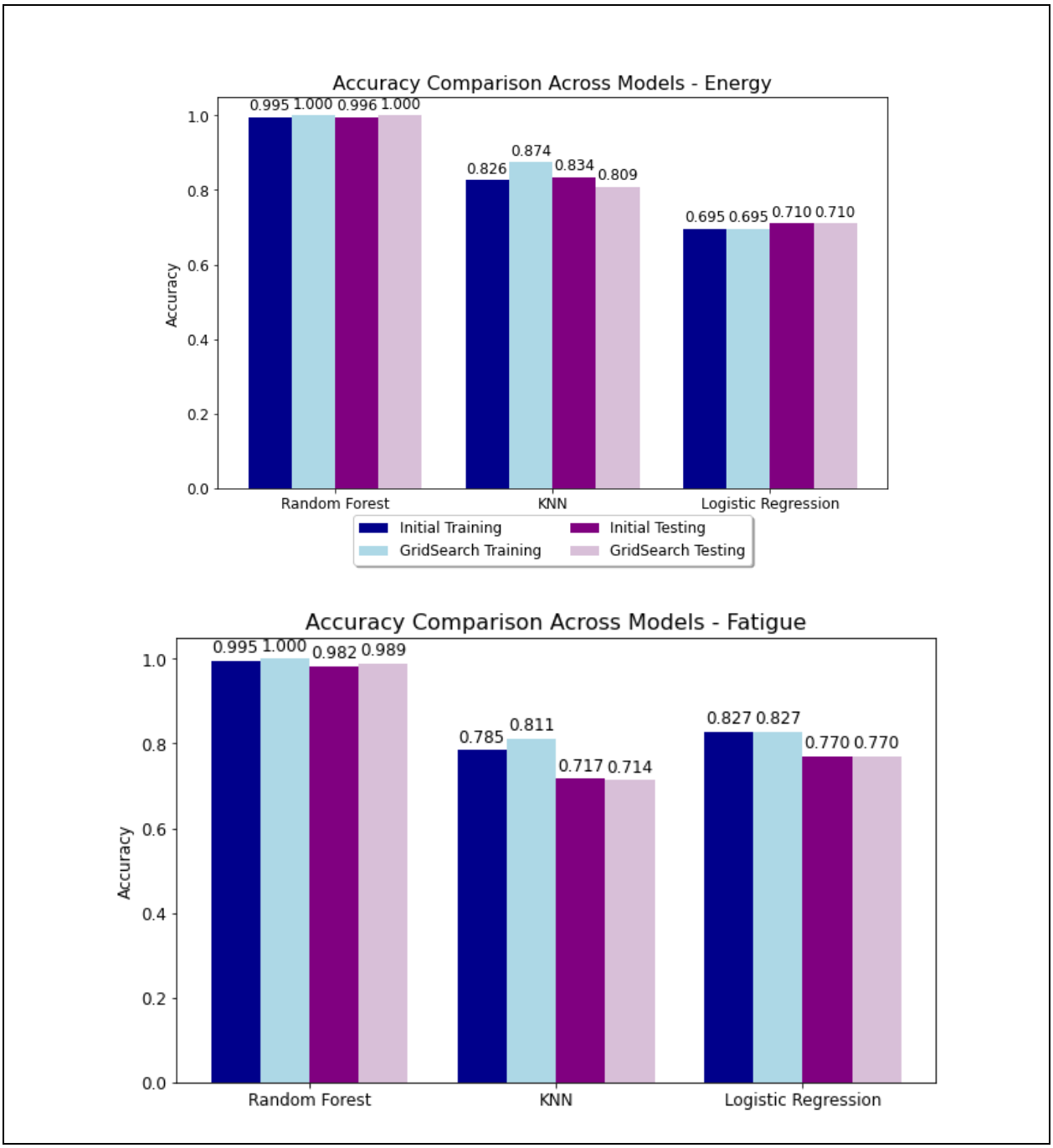


Figure 6: Bar charts displaying the training and testing accuracies of Random Forest, KNN, and Logistic Regression models for energy and fatigue prediction, comparing initial and GridSearch-enhanced results.

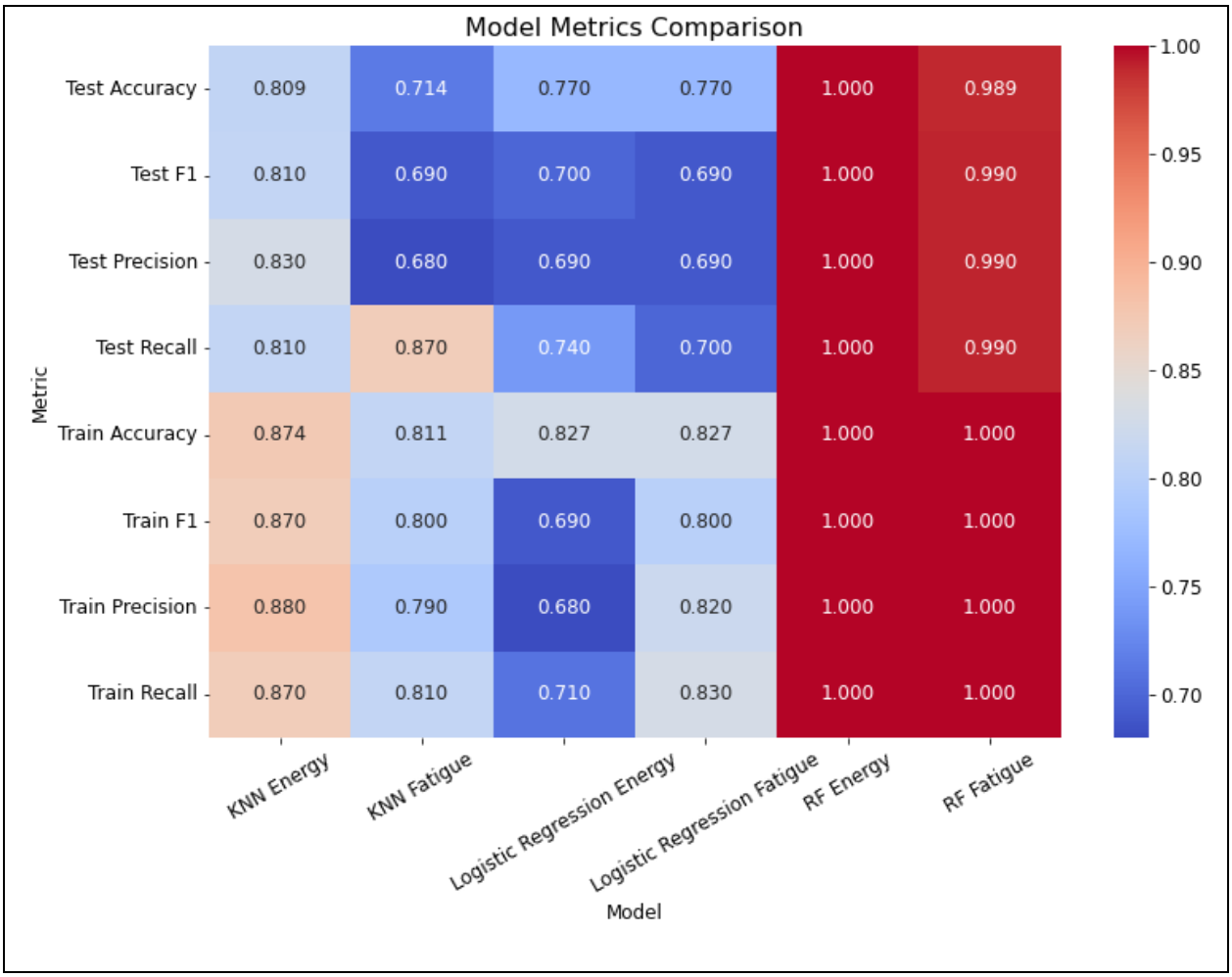


Figure 7: Heatmap displaying performance metrics for machine learning models in predicting energy and fatigue levels, with color gradients indicating the range from lower (blue) to higher (red) scores. Darker shades denote higher accuracy, precision, recall, and F1-scores for each model.