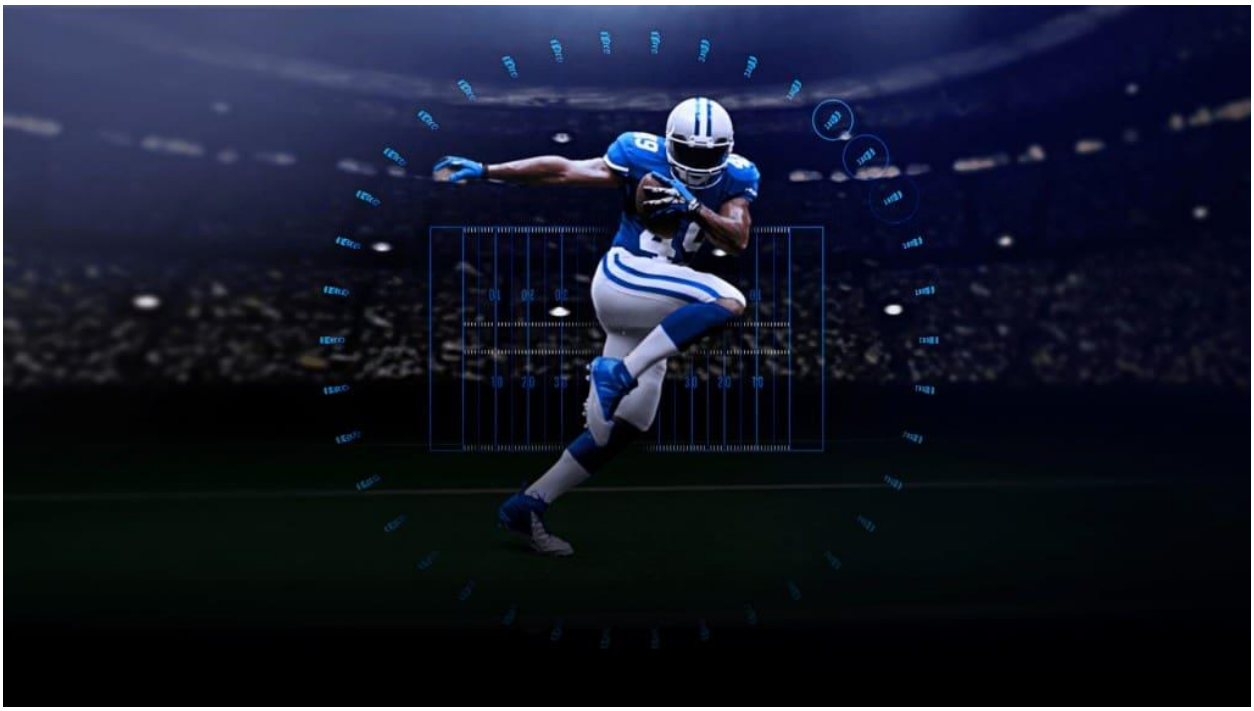


TOPIC: DAILY SPORTS ACTIVITY RECOGNITION

IE 7275 DATA MINING IN ENGINEERING

PRO. SRINIVASAN RADHAKRISHNAN

PTOJECT BY OMKAR NARKAR

INTRODUCTION

The dataset utilized for this project comprises motion sensor data from 19 daily and sports activities that were performed by 8 subjects for a duration of 5 minutes each. The unique aspect of this dataset lies in its capturing of variations in the performance of these activities by individuals, allowing for a more comprehensive analysis of the movements involved. The activities range from basic tasks like sitting and standing to more complex exercises such as rowing and playing basketball.

The subjects, consisting of 4 females and 4 males aged between 20 and 30, were instructed to execute the activities in their own preferred style, without any specific guidelines. Consequently, this led to variations in the speeds and amplitudes of the activities among the subjects, adding an additional layer of complexity to the dataset. The activities took place at different locations, including a sports hall, an engineering building, and a flat outdoor area on campus.

The dataset contains data gathered from five Xsens MTx units that were worn on the torso, arms, and legs of the subjects. Each unit is equipped with 9 sensors that measure accelerations, gyroscopes, and magnetometer readings along the x, y, and z axes. The data was recorded at a sampling frequency of 25 Hz and segmented into 5-second intervals, resulting in a total of 480 segments for each activity.

The activities encompassed in the dataset span a wide spectrum of movements, ranging from straightforward tasks like sitting and standing to more intricate exercises like cycling, rowing, and playing basketball. Additionally, the subjects were tasked with activities such as walking on a treadmill at varying speeds and inclinations, as well as using exercise machines like a stepper and a cross trainer.

The dataset's structure is organized based on activity, subject, and segment, with each text file containing data from all sensors over a 5-second period. The columns in the text files correspond to the different sensors on each unit, providing the basis for a detailed examination of the movements associated with each activity.

PROBLEM STATEMENT

The problem at hand involves predicting the type of activities performed by individuals based on the data collected from all sensors. This task requires the development of a predictive model that can effectively analyze the motion sensor data and classify the activities accurately. By leveraging machine learning algorithms and techniques, the goal is to create a robust predictive model that can automate the recognition and classification of daily and sports activities. The challenge lies in processing the multivariate and time-series data from the sensors to identify distinct patterns associated with each activity, overcoming variations in how the activities are performed by different individuals. Ultimately, the objective is to harness the potential of the dataset to improve activity recognition and analysis through the development of accurate and efficient predictive models.

OBJECTIVE

The primary objective of this project is to develop predictive models capable of accurately predicting the type of activity based on sensor data. This entails implementing and training multiple machine learning algorithms to analyze the multivariate and time-series data collected from the sensors. The project aims to explore the potential of the dataset in automating the recognition and classification of daily and sports activities through the application of advanced machine learning techniques. Additionally, the project involves refining data collection processes to ensure the quality and relevance of the data used for training the predictive models. By leveraging the diverse dataset and employing a variety of machine learning approaches, the project seeks to enhance activity recognition capabilities and contribute to the advancement of automated analysis in the field of human motion studies.

DATA SORCE

The data source "Daily sports activity recognition"
Citation: Billur Barshan and Kerem Atun
Intention Dataset: UCI Machine Learning Repository
https://archive.ics.uci.edu/dataset/256/daily+and+sports+activities

DATA DESCRIPTION

Each of the 19 activities is performed by eight subjects (4 female, 4 male, between the ages 20 and 30) for 5 minutes. Total signal duration is 5 minutes for each activity of each subject. The subjects are asked to perform the activities in their own style and were not restricted on how the activities should be performed. For this reason, there are inter-subject variations in the speeds and amplitudes of some activities. The activities are performed at the Bilkent University Sports Hall, in the Electrical and Electronics Engineering Building, and in a flat outdoor area on campus. Sensor units are calibrated to acquire data at 25 Hz sampling frequency. The 5-min signals are divided into 5-sec segments so that 480(=60x8) signal segments are obtained for each activity.

The 19 activities are:
sitting (A1), standing (A2),
lying on back and on right side (A3 and A4),
ascending and descending stairs (A5 and A6),
standing in an elevator still (A7) and moving around in an elevator (A8),
walking in a parking lot (A9),
walking on a treadmill with a speed of 4 km/h (in flat and 15 deg inclined positions) (A1 0 and A11),
running on a treadmill with a speed of 8 km/h (A12),
exercising on a stepper (A13),
exercising on a cross trainer (A14),
cycling on an exercise bike in horizontal and vertical positions (A15 and A16),
rowing (A17),
jumping (A18), and playing basketball (A19).

File structure:
19 activities (a) (in the order given above)
8 subjects (p)
60 segments (s)
5 units on torso (T),
right arm (RA),
left arm (LA),
right leg (RL),
left leg (LL)

9 sensors on each unit (x,y,z accelerometers, x,y,z gyroscopes, x,y,z magnetometers) Folders a01, a02, ..., a19 contain data recorded from the 19 activities.

Folders a01, a02, ..., a19 contain data recorded from the 19 activities.

For each activity, the subfolders p1, p2, ..., p8 contain data from each of the 8 subjects.

In each subfolder, there are 60 text files s01, s02, ..., s60, one for each segment.

In each text file, there are 5 units x 9 sensors = 45 columns and 5 sec x 25 Hz = 125 rows. Each column contains the 125 samples of data acquired from one of the sensors of one of the units over a period of 5 sec. Each row contains data acquired from all of the 45 sensor axes at a particular sampling instant separated by commas.

Columns 1-45 correspond to:
T_xacc, T_yacc, T_zacc, T_xgyro, ..., T_ymag, T_zmag, RA_xacc, RA_yacc, RA_zacc, RA_xgyro, ..., RA_ymag, RA_zmag, LA_xacc, LA_yacc, LA_zacc, LA_xgyro, ..., LA_ymag, LA_zmag, RL_xacc, RL_yacc, RL_zacc, RL_xgyro, ..., RL_ymag, RL_zmag, LL_xacc, LL_yacc, LL_zacc, LL_xgyro, ..., LL_ymag, LL_zmag.

Therefore, columns 1-9 correspond to the sensors in unit 1 (T), columns 10-18 correspond to the sensors in unit 2 (RA), columns 19-27 correspond to the sensors in unit 3 (LA), columns 28-36 correspond to the sensors in unit 4 (RL), columns 37-45 correspond to the sensors in unit 5 (LL).

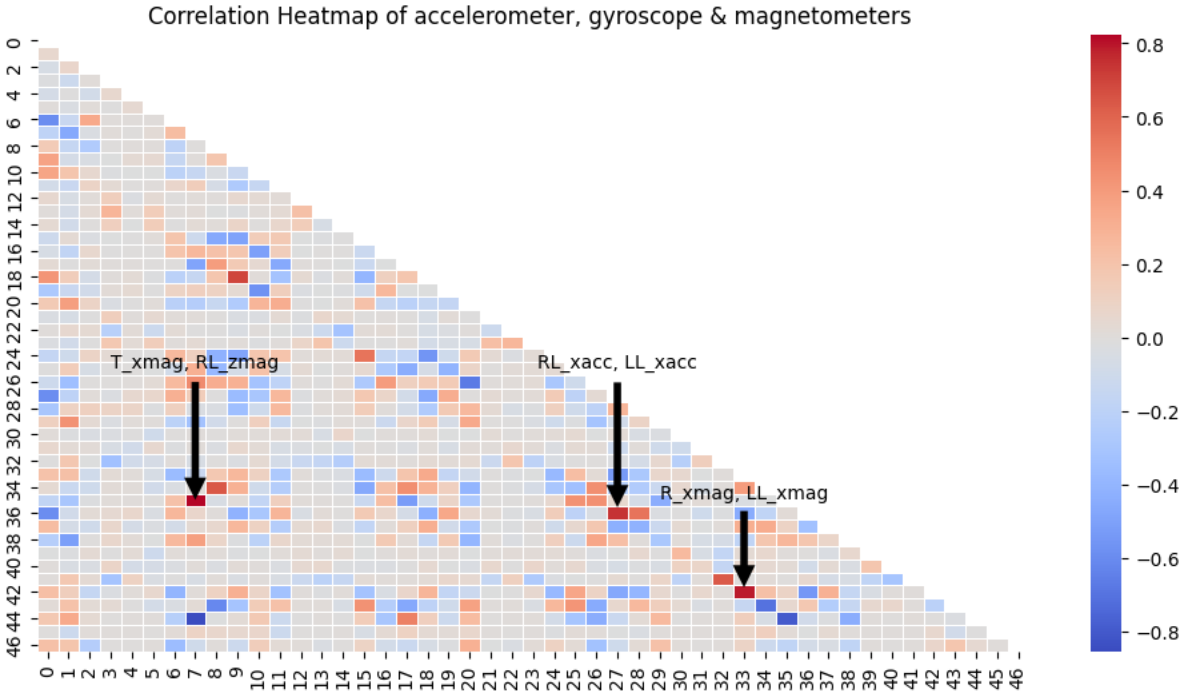
Data Pre-Processing

Data Collection:

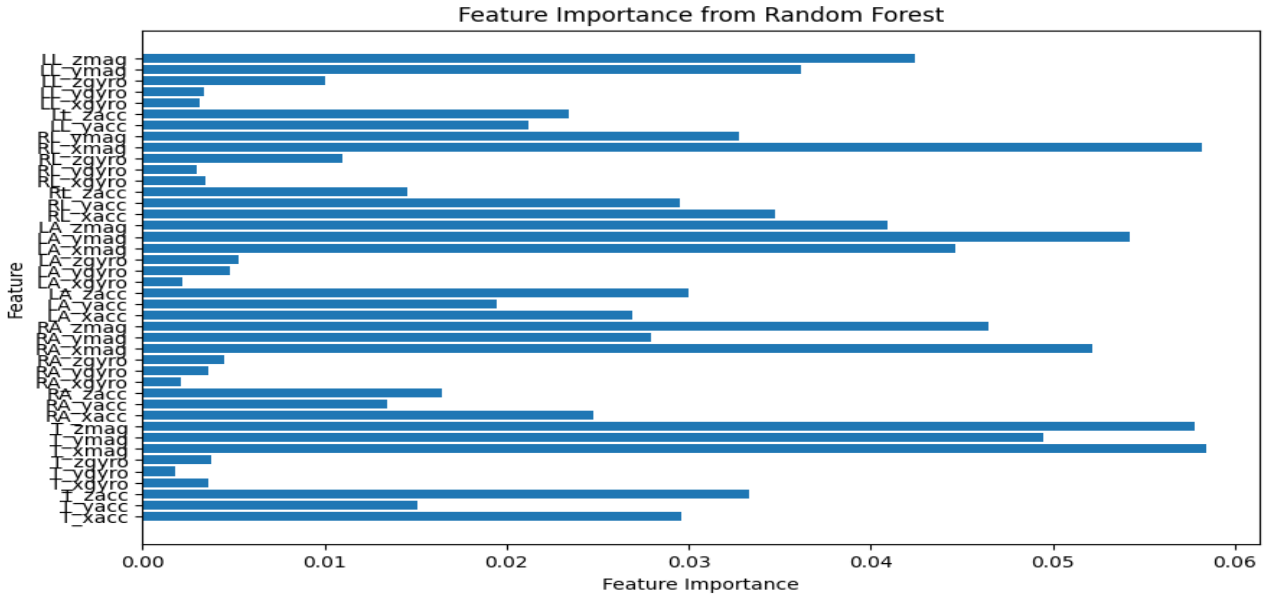
The data, initially stored in individual folders in txt format for each person and activity, was consolidated using Python libraries such as Pandas and NumPy. By merging and transforming the text file data into a CSV format, we aimed to enhance performance. Subsequently, to optimize efficiency further, the CSV files were converted into Parquet files. This approach streamlined the data storage and retrieval process, facilitating smoother operations for subsequent analyses and machine learning modeling.

Feature Selection:

a) In the data preprocessing stage, a correlation matrix was constructed to identify highly correlated features. Subsequently, six features exhibited strong correlations, leading to the decision to eliminate three of them to reduce ambiguity in machine learning modeling. This step aimed to enhance the accuracy and interpretability of the predictive models by eliminating redundant or highly correlated features that could potentially introduce noise and hinder the effectiveness of the machine learning algorithms.



b) After applying PCA for feature selection, I identified six features with minimal predictive contribution and subsequently removed them before training the KNN model. This streamlined dataset was then utilized for modeling, focusing on the most informative attributes for accurate predictions. This approach optimized the feature set by discarding less relevant variables, enhancing the efficiency and effectiveness of the KNN algorithm in activity classification based on motion sensor data.



Modeling Performance

A) K-Nearest Neighbors (KNN)

	precision	recall	f1-score	support
0	1.00	1.00	1.00	11913
1	1.00	1.00	1.00	11946
2	1.00	1.00	1.00	12037
3	1.00	1.00	1.00	12142
4	0.91	1.00	0.95	12006
5	0.97	0.93	0.95	11999
6	0.99	1.00	1.00	12299
7	0.96	0.97	0.96	11926
8	0.97	1.00	0.99	11960
9	0.97	0.99	0.98	11999
10	0.98	0.98	0.98	12078
11	0.97	0.99	0.98	11973
12	0.97	0.99	0.98	12081
13	0.98	1.00	0.99	11950
14	1.00	1.00	1.00	11941
15	1.00	1.00	1.00	11962
16	1.00	1.00	1.00	11889
17	0.95	0.94	0.95	12097
18	0.96	0.81	0.88	11802
accuracy			0.98	228000
macro avg	0.98	0.98	0.98	228000
weighted avg	0.98	0.98	0.98	228000

Precision: This metric measures the accuracy of positive predictions made by the model. For instance, activities like sitting (class 4) and walking (class 5) have slightly lower precision scores (0.91 and 0.97 respectively), indicating that the model is less precise in identifying these activities compared to others. Precision values close to 1.00 (e.g., classes 0, 1, 2, 14, 15, 16) signify high accuracy in predicting those specific activities.

Recall: Recall indicates the model's ability to identify all relevant instances of a class. Lower recall scores for certain activities (e.g., class 5 with a recall of 0.93) suggest that the model may miss some instances of these activities in the dataset.

F1-Score: This is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. High F1-scores (close to 1.00) indicate a good balance between precision and recall for most activities, while lower scores (e.g., class 5 with an F1-score of 0.95) highlight areas where the model's performance can be improved.

Support: The support column indicates the number of instances (samples) for each activity in the dataset. Activities with larger support values (e.g., classes 3, 6, 12) have more data available for training and evaluation.
Accuracy: The overall accuracy of 98% suggests that the KNN model performs well in classifying the activities, considering all classes and their respective distributions in the dataset.

B) Decision Tree

Accuracy: 0.94

	precision	recall	f1-score	support
0	1.00	1.00	1.00	11913
1	1.00	1.00	1.00	11946
2	1.00	1.00	1.00	12037
3	1.00	1.00	1.00	12142
4	0.94	0.90	0.92	12006
5	0.88	0.94	0.91	11999
6	0.97	0.98	0.97	12299
7	0.87	0.88	0.87	11926
8	0.94	0.95	0.95	11960
9	0.91	0.91	0.91	11999
10	0.91	0.92	0.92	12078
11	0.94	0.93	0.94	11973
12	0.93	0.93	0.93	12081
13	0.95	0.96	0.96	11950
14	0.99	0.99	0.99	11941
15	0.99	0.99	0.99	11962
16	1.00	1.00	1.00	11889
17	0.88	0.87	0.87	12097
18	0.75	0.71	0.72	11802
accuracy			0.94	228000
macro avg	0.94	0.94	0.94	228000
weighted avg	0.94	0.94	0.94	228000

Accuracy: The overall accuracy of 94% indicates that the Decision Tree model correctly predicts the activities for a significant portion of the dataset.

Precision and Recall: Precision measures the accuracy of positive predictions, while recall reflects the model's ability to identify all relevant instances of a class.
Activities like sitting (class 4), walking (class 5), and using exercise machines (class 17, class 18) exhibit lower precision and recall scores compared to other activities. This suggests challenges in accurately classifying these activities with the Decision Tree model.

F1-Score: The F1-score, which is the harmonic mean of precision and recall, provides a balanced measure of the model's performance for each activity. Lower F1-scores (e.g., activities with precision, recall, and F1-score around 0.90 or lower) indicate areas where the model may struggle to make accurate predictions.

Support: The support column indicates the number of instances (samples) for each activity in the dataset. Activities with higher support values have more data available for evaluation.

Confusion Matrix:

```
Confusion Matrix:
[[11898     0     0     0     0     0     0     2     0     1     0     1
      2     0     4     0     1     0     4]
 [    0 11922     0     0     2     3     0     7     3     1     1     0
      0     0     0     0     0     4     3]
 [    0     0 12028     0     0     0     0     0     0     0     0     0
      3     0     0     0     3     1     2]
 [    0     0     0 12141     0     0     0     0     0     0     0     0
      0     0     0     0     0     0     1]
 [    0     5     0     0 10801   606    15   143    48   119    61     3
      6     3     1     5     0    57   133]
 [    0     3     0     0     0 11241    32   277    32     0     0     0
      1     0     0     1     0   318    94]
 [    0     4     0     0    15    14 12013   232     6     0     0     0
      1     0     0     0     0     3    11]
 [    1     3     0     0   137   271   288 10444    58    11    15     5
     36     9    14    11     0   138   485]
 [    0     1     0     0    47    30     4    61 11348   115    72     8
     16     7     0     1     0    89   161]
 [    0     0     0     0   178     0     0    11   125 10885   489    36
     72    24     1     8     0    22   148]
 [    0     0     0     0    80     0     3     9    63   482 11119    41
     29    45     3     6     3    16   179]
 [    1     0     0     0     4     0     0    11    13    44    57 11190
     60    97     7    12     9    62   406]
 [    8     0     0     0     5     0     2    26    29    59    73    89
  11231   136     5     1     3    15   399]
 [    0     0     0     0     6     0     0    12     4    20    63    94
    144 11423     7    11     0     3   163]
 [    2     0     0     0     2     1     2     9     0     1     1     4
      2     7 11857    26     0     3    24]
 [    0     0     0     0     7     0     0    11     2     7     6     7
      3    16    21 11812     0    15    55]
 [    0     0     2     1     0     0     0     0     0     1     6     8
      6     0     0     0 11851     0    14]
 [    2     5     0     0    68   480    10   180   102    15    19    54
     33     7     7    29     0 10525   561]
 [    7    16     4     0   179   113    17   582   222   204   227   417
    456   189    34    64    25   720  8326]]
```

- The confusion matrix reveals that the Decision Tree model performs well for some activities (e.g., classes 0, 1, 2, 3, 14, 15, 16) with high diagonal values, indicating accurate predictions.
- Certain activities show noticeable misclassifications (e.g., classes 4, 5, 7, 8, 9, 10, 11, 12, 13, 17, 18), as seen in the off-diagonal elements. For example, class 4 has significant misclassifications as class 5, class 8 has misclassifications as class 4 and class 9, and class 17 has misclassifications as class 4 and class 5.
- The confusion matrix provides insights into specific areas where the Decision Tree model struggles to differentiate between similar activities or where there may be overlapping features leading to misclassifications.
- This analysis can guide further model refinement efforts, such as feature engineering, parameter tuning, or exploring ensemble methods, to improve the model's ability to correctly classify activities with higher precision and recall.

In summary, the confusion matrix offers a comprehensive view of the Decision Tree model's performance in predicting daily activities based on motion sensor data, highlighting strengths and areas for improvement in activity recognition and classification.

C) Random Forest

	precision	recall	f1-score	support
1	1.00	0.00	0.00	17850
2	0.00	0.00	0.00	17911
3	1.00	0.99	1.00	18019
4	1.00	1.00	1.00	17915
5	0.00	0.00	0.00	18010
6	0.00	0.00	0.00	17948
7	0.00	0.00	0.00	17989
8	0.75	0.00	0.00	18197
9	0.91	0.00	0.00	18064
10	0.29	0.83	0.43	17990
11	0.00	0.00	0.00	18028
12	0.73	0.08	0.14	18008
13	0.17	0.93	0.28	17861
14	0.90	0.00	0.01	17971
15	0.16	0.99	0.27	18171
16	0.82	0.95	0.88	18091
17	1.00	0.82	0.90	17990
18	0.73	0.08	0.14	17978
19	0.59	0.01	0.02	18009
accuracy			0.35	342000
macro avg	0.53	0.35	0.27	342000
weighted avg	0.53	0.35	0.27	342000

- The confusion matrix highlights specific areas where the Random Forest model struggles to differentiate between activities based on motion sensor data.
- Improvement strategies may include feature engineering, model tuning (e.g., adjusting hyperparameters), or exploring alternative algorithms to enhance the model's ability to accurately classify activities with higher precision and recall.

In summary, the Random Forest model's performance, as reflected in the classification metrics and confusion matrix, underscores the need for further refinement and optimization to improve its accuracy and reliability in predicting daily activities from motion sensor data.

Best Performing Model

Accuracy:

Decision Tree: Achieved an accuracy of 94%.
KNN: Achieved an accuracy of 98%.
Random Forest: Achieved an accuracy of 35%.
Analysis: Among the three models, KNN achieved the highest accuracy (98%), indicating that it overall made the most correct predictions compared to the other models.
Precision, Recall, and F1-Score:
These metrics provide insights into the models' abilities to correctly classify activities (precision), capture all relevant instances of each activity (recall), and balance between precision and recall (F1-score).

Analysis: Decision Tree: Shows a balanced performance across precision, recall, and F1-score for most activities, though it struggled with certain activities (e.g., class 4, class 5).

KNN: Demonstrates high precision, recall, and F1-score across most activities, indicating robustness in activity classification.
Random Forest: Exhibits varying performance with low precision, recall, and F1-score for several activities, suggesting challenges in accurate prediction.

Confusion Matrix Insights: The confusion matrix reveals specific areas of strength and weakness for each model in classifying different activities.

Analysis: Decision Tree: Shows decent performance but struggles with certain activities, leading to misclassifications.

KNN: Generally, performs well across all activities with high precision and recall.
Random Forest: Faces significant challenges in accurately predicting many activities, resulting in low performance metrics.

Conclusion

This project explored the application of machine learning algorithms for automated recognition and classification of daily and sports activities using motion sensor data. Various models including KNN, Decision Tree, and Random Forest were evaluated based on accuracy, precision, recall, and F1-score across 19 activities. KNN emerged as the top-performing model with 98% accuracy, effectively classifying activities based on diverse motion patterns. Future steps involve refining the model through hyperparameter tuning and feature engineering to optimize performance. The research underscores the potential of machine learning in health monitoring, sports analytics, and human-computer interaction, showcasing how these technologies can automate activity recognition tasks with high accuracy and reliability, paving the way for practical applications in real-world settings.