Machine Learning Project on Knee vibroarthrography Dataset

1. Introduction

The goal of this project is to analyze and predict knee conditions based on vibrography signal features. We worked with the dataset **`vag_dataset.csv`** containing 2500 samples with features related to knee condition signals and corresponding labels:

- **Output variables:**
  - `knee_condition`
  - `severity_level`
  - `treatment_advised`
- **Input features:**
  - `rms_amplitude`
  - `peak_frequency`
  - `spectral_entropy`
  - `zero_crossing_rate`
  - `mean_frequency`

---

2. Data Exploration & Preprocessing

- **Missing Values:**

    - Found missing values in `severity_level` (~795 missing).
    - Imputed missing values with mode (most frequent class) to avoid losing data.
- **Duplicate Rows:**
    - Checked and confirmed no duplicate rows in the dataset.
- **Data Encoding:**
    - Converted categorical outputs (`knee_condition`, `severity_level`, `treatment_advised`) into numeric labels using **Label Encoding** for modeling.
- **Feature Scaling:**
    - Applied **StandardScaler** to all numeric input features to normalize data for better model performance.
- **Outlier Detection:**
    - Used **Z-score method** to detect outliers across numeric features.
    - Found **no significant outliers** (Z-score > 3) in any feature, so no removal was performed.

---

3. Modeling

Models Used:

- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Random Forest Classifier (RF)

Targets:

- Each output (`knee_condition_encoded`, `severity_level_encoded`, `treatment_advised_encoded`) modeled separately.

---

4. Model Tuning and Evaluation

- Used **GridSearchCV** to find best hyperparameters for each model per output variable.
- **Best Hyperparameters Found:**

Model	knee_condition_encoded	severity_level_encoded	treatment_advised_encoded
KNN	n_neighbors = 11	n_neighbors = 11	n_neighbors = 11
SVM	C = 10, kernel = linear	C = 10, kernel = rbf	C = 10, kernel = linear
RandomForest	max_depth = 10, n_estimators = 100	max_depth = 20, n_estimators = 200	max_depth = 10, n_estimators = 200

---

5. Performance Results (5-fold Cross-Validation)

Model	knee_condition	severity_level	treatment_advised
KNN	86%	54%	72%
SVM	87%	54%	75%
RandomForest	86%	52%	74%

---

6. Feature Importance (Random Forest)

For all outputs, the top three important features are:

1. `rms_amplitude`

2. `spectral_entropy`

3. `mean_frequency`

Other features (`zero_crossing_rate`, `peak_frequency`) contribute very little.

---

7. Observations and Insights

- **Knee condition** is predicted with **high accuracy (~86-87%)** across all models, indicating good discriminative power of the features.

- **Treatment advised** shows moderate prediction accuracy (~72-75%), suggesting some complexity or overlap in treatment groups.

- **Severity level** is challenging to predict (~52-54% accuracy), which may indicate noisy labels, data imbalance, or insufficient features for this output.

---

8. Next Steps

- **Plot confusion matrices** for all outputs to better understand specific class-wise prediction errors and misclassifications.

- Explore **multilabel or multioutput models** to simultaneously predict all three outputs and capture possible relationships between them.

- Investigate advanced feature engineering or adding domain knowledge features to improve severity level prediction.

- Consider other classifiers or ensemble methods and compare results.

- Experiment with different missing data imputation strategies or balancing classes for severity level.

---

9. Summary

We have completed the initial data cleaning, preprocessing, feature scaling, outlier detection, and applied three classification models (KNN, SVM, Random Forest) with hyperparameter tuning. Knee condition classification shows promising results, while severity level remains challenging. Further exploration and refinement are planned.