Machine Learning Models for Multi Query Classification

Theory Concepts

Ensemble Learning

Ensemble is a method of combining two or more models into one to increase prediction accuracy. Generally, it involves combining weak learners to form a strong learner.

Types of Ensemble Methods

- **Voting Classifier**: Used in classification tasks. It combines predictions from multiple models.
  - **Hard Voting**: The class that gets the majority of votes is selected. For example, if two models predict Class A and one model predicts Class B, Class A is chosen.
  - **Soft Voting**: The final prediction is based on the average of predicted probabilities. The class with the highest average probability is selected.

Pipelining

Pipeline combines preprocessing and modeling into a single object, chaining multiple steps together. In the pipeline:

- **StandardScaler** is used to scale features, ensuring better and faster model performance.

Random Under Sampling

Random under sampling is used to balance the class distribution by reducing the number of samples in majority classes, enabling balanced predictions.

TF-IDF Vectorizer

TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer identifies important words in sentences.

- **Term Frequency (TF)** = (Number of times a word appears in a sentence) / (Total number of words in the sentence)
- **Inverse Document Frequency (IDF)** $= \log\left(\frac{\text{Total number of sentences}}{\text{Number of sentences containing the word}}\right)$
- **TF-IDF** = TF $\times$ IDF

Other Concepts

- **Counter**: Used to count the frequency of strings or elements.
- **Alpha (Smoothing Factor)**: Used to prevent overfitting by adding a smoothing effect during model training.
- **Coerce**: While converting text to numeric, non-numeric values are replaced with NaN (Not a Number).
- **Flatten**: Used to convert multi-dimensional arrays into a 1-dimensional array.

Model Accuracy Table

ML Model Name	**Accuracy (%)**
Naive Bayes (NB)	65.14
Extreme Gradient Boosting (XGBoost)	53.58
Logistic Regression (LR)	65.50
Support Vector Machine (SVM)	62.18
Random Forest (RF)	54.98
Stochastic Gradient Descent (SGD)	64.87

Table 1: Accuracy of individual ML models

Ensemble	Accuracy (%)
NB + LR + SVM + SGD + XGB	65.68
NB + LR + SVM + RF + SGD	64.92
NB + LR + RF + SGD + XGB	64.91
NB + SVM + RF + SGD + XGB	64.70
NB + LR + SVM + RF + XGB	64.64
LR + SVM + RF + SGD + XGB	63.99

Table 2: Top 5-Model Ensemble Accuracies

Top 5-Model Ensemble Results

All 3-Model Ensemble Results

Ensemble	Accuracy (%)
NB + LR + SGD	65.70
NB + SGD + XGB	65.66
NB + LR + XGB	65.65
NB + SVM + XGB	65.43
NB + SVM + SGD	65.40
NB + LR + SVM	65.39
LR + SVM + SGD	65.16
LR + SVM + XGB	65.00
SVM + SGD + XGB	64.96
NB + LR + RF	64.93
NB + RF + SGD	64.80
NB + SVM + RF	64.61
NB + RF + XGB	64.05
LR + SGD + XGB	63.96
LR + SVM + RF	63.86
SVM + RF + SGD	63.76
LR + RF + SGD	63.11
SVM + RF + XGB	62.91
LR + RF + XGB	60.93
RF + SGD + XGB	59.92

Table 3: All 3-Model Ensemble Accuracies (Top to Bottom)

Data Processing Flowchart