Joy Chen(jc10137)

Report on Fetal Health Classification

I. Problem Statement

The objective is to classify the health of a fetus into three categories – Normal, Suspect, or Pathological – using Cardiotocograms (CTGs) data. This classification is pivotal in determining the required medical intervention to prevent child and maternal mortality.

II. Data Sources

The dataset used for this analysis was provided in the form of a CSV file, named `fetal_health.csv`. This dataset consists of features extracted from CTG exams, which were then classified by expert obstetricians into three classes. I got this dataset from Kaggle.

III. Feature Engineering and Preprocessing Done

1. **Replacing Missing Value**: Replace missing values with certain rules. In this dataset, there are no missing values.
2. **Removing Duplicates**: Duplicate rows were identified and removed to ensure the uniqueness of each data entry.
3. **Handling Outliers**: In the context of this dataset, there are important reasons for not automatically removing these outliers:
    A. In medical datasets like this one, what might appear as outliers could represent clinically significant cases, indicating unusual but important medical conditions that are critical for the model to learn from
    B. If we remove too many data points, the remaining dataset might not be representative of the real-world diversity, leading the model to overfit on the reduced dataset.
4. **Feature Selection**: Feature selection is not always strictly necessary, especially with datasets of moderate size and models that handle many features well. Thus, even though I tried feature selection, I did not use it in the construction of models.
3. **Standardization**: Features were standardized to have a mean of zero and a standard deviation of one. This step is crucial for models sensitive to the scale of the data.
4. **Train-Test Split**: The data was split into training and testing sets with a ratio of 80:20, ensuring stratification to maintain the distribution of the target classes.

IV. Model Performance

The following table summarizes the performance of various models (we used the macro average for precision, recall, and F1-Score):

Model	Accuracy	Precision	Recall	F1-Score

KNN	90.78%	84%	77%	80%
SVM	90.31%	85%	78%	81%
Decision Tree	93.62%	94%	86%	89%
Random Forest	95.27%	96%	88%	91%
Gradient Boosting	95.51%	95%	90%	92%

V. Hyperparameter Search Results

- **KNN Best Parameters**:
  - `n_neighbors`: 3
  - `KNN Accuracy with Best Params`: 90.78%

- **SVM Best Parameters**:
  - `kernel`: rbf
  - `C`: 100
  - `gamma`: 0.01
  - `SVM Accuracy with Best Params`: 90.31%

- **Decision Tree Best Parameters**:
  - `min_samples_leaf`: 2
  - `max_depth`: 10
  - `min_samples_split`: 5
  - `Decision Tree Accuracy with Best Params`: 93.62%

- **Random Forest Best Parameters**:
  - `n_estimators`: 200
  - `max_depth`: None
  - `Random Forest Accuracy with Best Params`: 95.27%

- **Gradient Boosting Best Parameters:**
  - `n_estimators`: 300
  - `max_depth`: 5
  - `learning_rate`: 0.1
  - `Gradient Boosting Accuracy with Best Params`: 95.51%

VI. Feature Importance

Understanding which features are most influential in classification can be crucial for making informed decisions. Here, we highlight the top 5 feature importance as determined by the Random Forest and Gradient Boosting models:

- Random Forest Feature Importance:
 - mean_value_of_short_term_variability: 0.13523451
 - abnormal_short_term_variability: 0.13159709
 - percentage_of_time_with_abnormal_long_term_variability: 0.10535975
 - histogram_mean: 0.08989029
 - prolongued_decelerations: 0.05623582

- Gradient Boosting Feature Importance:
 - abnormal_short_term_variability: 2.45910934e-01
 - mean_value_of_short_term_variability: 1.44083254e-01
 - percentage_of_time_with_abnormal_long_term_variability: 1.38451654e-01
 - histogram_mean: 1.33120666e-01
 - prolongued_decelerations: 6.56771352e-02

VII. Conclusions

- **Model Selection:** Among the models tested, the Gradient Boosting model exhibited the highest accuracy, making it a favorable choice for this classification task.
- **Importance of Preprocessing**: Standardization played a key role in enhancing model performance, particularly for models like SVM and KNN.
- **Hyperparameter Tuning**: The results from hyperparameter tuning indicated potential improvements in model performance, highlighting its importance in model optimization.

Further research and model tuning, including advanced feature engineering and experimenting with different algorithms, could potentially lead to even better performance and more robust models for fetal health classification.

VIII. Link to the GitHub repo

https://github.com/joyc7/PA-HW3-fetal-health-prediction