CSL2050 - Pattern Recognition and Machine Learning
Project Mid-Report

Stroke Prediction Using Machine Learning

Problem Statement
Stroke is a leading cause of death and disability worldwide, making early detection critical for effective intervention. This project aims to develop a ML model to predict stroke occurrence based on patient data, including demographic, lifestyle, and health-related features. Accurate prediction can help healthcare professionals identify at-risk individuals, prioritizing sensitivity (recall) to minimize missed cases while balancing precision to reduce false alarms. The challenge lies in handling a highly imbalanced dataset, where stroke cases are rare compared to non-stroke cases, requiring robust techniques to ensure reliable performance on the minority class.

Dataset
The dataset, sourced from Kaggle as "Healthcare Dataset Stroke Data" (healthcare-dataset-stroke-data.csv), contains 5,110 patient records with 12 initial features:

- `id`, `gender`, `age`, `hypertension`, `heart_disease`, `ever_married`, `work_type`, `Residence_type`, `avg_glucose_level`, `bmi`, `smoking_status`, and `stroke` (target variable).

After preprocessing, the dataset is saved as `cleaned_dataset.csv` with 16 features and no missing values.

Preprocessing Steps

- **Data Cleaning:** Dropped the `id` column, imputed missing `bmi` values (201 instances, 3.93%) with the median, and retained all 5,110 rows.
- **Encoding:** Converted categorical features (`gender`, `ever_married`, `work_type`, `Residence_type`, `smoking_status`) to numerical values using `LabelEncoder`.
- **Feature Engineering:** Added categorical features:
    - `age_group` (Young, Middle-Aged, Senior, Elderly)
    - `bmi_category` (Underweight, Normal, Overweight, Obese)
    - `glucose_category` (Low, Normal, Prediabetes, Diabetes, High Risk)
    - Created interaction terms: `age_bmi_interaction` and `hypertension_heart_disease`.
- **Scaling:** Standardized numerical features (`age`, `avg_glucose_level`, `bmi`) using `StandardScaler`.
- **Target Distribution:** The dataset is imbalanced, with 95.13% "No Stroke" (4,861 instances) and 4.87% "Stroke" (249 instances), posing a challenge for model training.

Exploratory Analysis

- **Stroke Distribution:** Only 4.9% of patients had a stroke, confirming severe imbalance.
- **Categorical Insights:** Higher stroke risk observed in:
  - Married individuals (5.97%)
  - Self-employed workers (7.80%)
  - Former smokers (7.63%)
- **Numerical Features:** Older age and higher glucose levels linked to increased stroke risk.
- **Correlation:** Moderate correlations exist between:
  - `age` and `stroke` (0.25)
  - `hypertension` and `heart_disease` (0.13)

Early Results
Two models were implemented and evaluated on a test set of 1,022 samples (20% of the dataset, stratified split).

k-Nearest Neighbors (k-NN)

Implementation:

- Pipeline with `StandardScaler`, `OneHotEncoder`, `SMOTE`, `SelectKBest` (ANOVA F-test), and `KNeighborsClassifier`.
- Tuned `n_neighbors`, `weights`, `metric`, and feature selection `k`, optimizing for ROC AUC.

Results:

- **Accuracy:** 0.8474
- **Precision (Stroke):** 0.23, **Recall (Stroke):** 0.92, **F1-Score (Stroke):** 0.37
- **Specificity:** 0.8436, **ROC AUC:** 0.9561, **Average Precision:** 0.6635
- **Confusion Matrix:** [[0.84, 0.16], [0.08, 0.92]]

Analysis:
k-NN excels at recall (0.92), identifying 92% of stroke cases, and achieves a high ROC AUC (0.9561), indicating strong discriminative power. However, precision (0.23) is low due to many false positives, reflecting SMOTE's effect on the imbalanced data.

Gaussian Naive Bayes (GNB)

Implementation:

- Initial pipeline with `OneHotEncoder`, `SMOTE`, `SelectKBest` (mutual information), and `GaussianNB`, optimized for ROC AUC.
- Re-trained with `ImbPipeline`, `SMOTE` adjustments (`sampling_strategy=0.5`), and `scoring='f1'` in `GridSearchCV`, tuning `var_smoothing` and `k`.

Updated Results (F1 Optimization):

- **Accuracy:** 0.7671
- **Precision (Stroke):** 0.15, **Recall (Stroke):** 0.78, **F1-Score (Stroke):** 0.25
- **Specificity:** 0.7665, **ROC AUC:** 0.8205, **Average Precision:** 0.1973
- **Confusion Matrix:** [[0.77, 0.23], [0.22, 0.78]]

Analysis:
Post-retraining, GNB's recall (0.78) remains strong, catching 78% of strokes, and its F1-score improved to 0.25 from 0.19. Precision (0.15) is still low, but specificity (0.7665) and accuracy (0.7671) increased significantly from 0.6276 and 0.6389, respectively, showing better balance.

Comparison and Prioritization

- **Recall:** k-NN (0.92) outperforms GNB (0.78), critical for minimizing missed strokes.
- **Precision/F1:** k-NN (0.23, F1: 0.37) is better than GNB (0.15, F1: 0.25).
- **AUC:** k-NN (0.9561) surpasses GNB (0.8205), suggesting superior ranking ability.
- **Accuracy:** k-NN (0.8474) beats GNB (0.7671), though less relevant due to imbalance.

Which Is Better?

- **Medical Priority (High Recall):** k-NN is preferable due to higher recall (0.92 vs. 0.78), but both models have low precision, leading to false positives.
- **Balanced Performance (F1-Score):** k-NN's higher F1-score (0.37 vs. 0.25) indicates a better balance, though GNB's retraining shows progress.

Proposed Approaches

1. **Further Refine GNB**
    - Adjust `SMOTE` ratios (0.3–0.7) and decision thresholds (0.6) to improve precision and F1-score.
2. **Enhance k-NN**
    - Modify `SMOTE` settings and test different thresholds (e.g., 0.7) to balance recall and precision.
3. **Explore Advanced Techniques**
    - Implement `SVM` (with class weights), `Decision Trees` (with pruning), and `ANN` to better capture complex patterns.

Next Steps
Immediate focus is on refining GNB and k-NN with further SMOTE and threshold adjustments. Subsequent steps involve implementing SVM, Decision Trees, and ANN, alongside enhanced feature engineering. Performance will be evaluated using recall, precision, F1-score, and AUC, aiming for a robust stroke prediction model by the project's conclusion.