Project progress update

Diabetes Prediction with Ensemble and Advanced Machine Learning Algorithms

Team members:

Neeraja Reddy Renati

Aparna Devabhaktuni

Sudam Rajamolla

Sai Spandana Sabbavarapu

Completed Tasks:

The core programming language for this project will be Python, using libraries such as Scikit-learn for machine learning tools, Pandas for data manipulation, NumPy for numerical computations, and Matplotlib or Seaborn for creating visualizations.

1. Data Collection

The dataset used in this project is the Pima Indians Diabetes Database, sourced from the UCI Machine Learning Repository. This dataset includes several medical predictor variables and one target variable, Outcome. The dataset is loaded using Pandas.

2. Data Preprocessing

Preprocessing involves cleaning the data to ensure it is suitable for analysis and modeling. Key steps include:

Handling Missing Values: Certain columns in the dataset have zero values, which are not medically plausible and need to be replaced with more appropriate statistics.

Removing Duplicates: Ensuring there are no duplicate rows in the dataset.

Checking for Null Values: Confirming there are no missing values after initial cleaning.

3. Exploratory Data Analysis (EDA)

Visualized data distributions and identified outliers using histograms and boxplots.

Explored relationships between features using scatter matrices.

4. Data Splitting

The data post preprocessing is ready to be used in our machine learning analysis. For understanding the performance of the model, we require a testing set of data. Therefore, the current dataset is split into two parts: a training set and a test set. The training set is 80% of the data and is used to train machine learning models. The test set is 20% of the data and will be used for model evaluation purposes. The data is

divided initially into X and y variables as part of attributes and classes. The Outcome variable is the class variable, and all other variables are predicting variables. The splitting is done using the sklearn library to obtain four variables related to training and testing.\

Data and dataset link:

The dataset being utilized is Pima Indians Diabetes Database from Kaggle, following the link: (diabetes.csv)

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data

Challenges:

Encountered Difficulties

Encountered difficulties encompassed understanding feature relevance for classification, particularly discerning global versus local structures, and effectively utilizing advanced techniques like PCA, KPCA, and t-SNE due to dataset complexity. Additionally, interpreting results from sophisticated models posed conceptual misunderstandings.

Addressing Challenges

To address these challenges, a two-pronged approach was adopted. Firstly, further research and a learning journey focused on PCA, KPCA, and t-SNE were initiated, supplemented by seeking guidance from a professor to navigate complex concepts. Secondly, to tackle data issues such as class imbalance, techniques like SMOTE were employed, ensuring a more balanced dataset for analysis.

Collaboration:

Meeting Frequency

The team meets twice a week, both in person and through Zoom. All members feel that contributions are meaningful and well-distributed.

Group Contribution:

Neeraja is responsible for Data Cleaning and Preprocessing Initial Model Training (Baseline Models)

Aparna focuses on Exploratory Data Analysis (EDA), Data Visualization

Sudam contributes Advanced Model Training, Ensemble Methods Implementation

Sai Spandana manages Model Evaluation and Fine-tuning Documentation and Final Report

Next Steps:

Task	Start Date	End Date
Feature Selection	01-06-2024	03-06-2024
Model Selection	04-06-2024	05-06-2024
Model Training	06-06-2024	08-06-2024
Model Evaluation	09-06-2024	10-06-2024
Fine-tuning	11-06-2024	13-06-2024
Final Model Selection	14-06-2024	15-06-2024

Plan to Complete Remaining Tasks:

Feature Selection:

Use advanced techniques to select the most relevant features.

Model Selection and Training: Continue to train and evaluate multiple models, refining them based on performance metrics.

Model Training and Evaluation:

Train various models including ensemble methods such as AdaBoost, Bagging, Extra Trees, Gradient Boosting, and XGBoost

Fine-tuning:

Optimize the selected model by adjusting hyperparameters and retraining.

Evaluation:

Perform comprehensive evaluations using metrics like accuracy, precision, recall, and F1-score.

Potential Challenges:

Fine-tuning the logistic regression model and interpreting results.

Ensuring robustness and generalizability of the final model.

Conclusion:

The project is progressing well, with significant milestones achieved in data cleaning, EDA, and initial model training. The team addresses challenges effectively and is well-coordinated, meeting regularly to ensure steady progress. The next steps are clearly defined, with a detailed plan to accomplish remaining tasks and tackle potential challenges.