**BUSA545 : Exploring Mental Health Data**

Group 9: Enyojo Phoebe Alabi, Hephzibah Jayaraj, Ginikachi Oyakhare

December 01, 2024

**Abstract**

This project leverages machine learning to predict depression by analyzing a mental health dataset from Kaggle. Multiple machine learning models were developed and evaluated, including Logistic Regression, Decision Trees, Gradient Boosting, LightGBM, and XGBoost. The best-performing model was selected to build a system designed to predict depression with high levels of accuracy. The data was preprocessed by handling missing values, feature engineering, and balancing class distributions with SMOTE. Metrics such as accuracy, precision, recall, and F1-score were used to assess model performance. XGBoost outperformed other models, delivering the highest accuracy and reliable predictions. This project demonstrates the potential of machine learning to contribute significantly to healthcare such as enabling early diagnosis and intervention in healthcare.

**1 Introduction**

Artificial Intelligence (AI) has significantly changed the way the analysis and interpretation of complex data is done across various fields. In healthcare, one of its most impactful applications is the use of machine learning models to predict and identify mental health conditions, including depression. This area has received a lot of attention, with numerous efforts leveraging machine learning algorithms to achieve groundbreaking breakthroughs. The ability of machine learning models to allow the integration of diverse data sources and data types such as demographic and behavioral factors, to build predictive models that can assist in early diagnosis and more customized treatment strategies has made their use very popular and relevant. This project explores the application of a variety of machine learning algorithms such as ensemble learning, a powerful technique that combines multiple models like Logistic Regression, and Decision Trees to develop a model with improved predictive accuracy. More advanced models such as Gradient Boosting, LightGBM, and XG Boost were implemented in a bid to explore the chance of developing more robust models. The algorithms used in this project fall under the domain of Supervised Machine Learning. Supervised Machine Learning algorithms rely on labeled data to predict target outcomes and make them suited for predictive analytics. The emphasis of this work is to develop effective classification models to predict the target variable (depression) based on the various features contained within the dataset. By leveraging the strengths of these algorithms, the project aims to provide a reliable framework for understanding factors influencing depression and aid in more accurate diagnosis.

**1.1 The problem you tried to solve**

Exploring factors that cause/contribute to depression in individuals by developing machine learning algorithms that can accurately identify individuals with depression and predict cases of depression with precision.

**1.2 Related Work**

A lot of work has been done in this area and several journals and articles document results from research and projects addressing this problem. Several scientific papers and other resources were consulted during this project to provide foundational knowledge of the general overview of the scope of work and gain insights into supervised machine learning, classification, ensemble methods, and depression.

In a study that focused on the use of Machine learning models to predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine, Linear regression models, logistic regression models, and SVM classification were used [1]. The Linear regression model was used to predict continuous response variables like anxiety score and the other classification models were used to predict the binary class of whether depression was present or not. The accuracy of the Linear regression model was assessed with R-square values, Mean Squared Error, and Mean Absolute Error. In contrast, the Classification models were judged using Area Under the Precision-Recall Curve (AUPRC), Area Under the Receiver Operating Characteristic curve, Balanced Accuracy, F1 score, and Brier loss [1].

Another enlightening work was done on identifying depression predictors from standard health surveys using machine learning [2]. This work employed 7 traditional machine learning algorithms: AdaBoost, Decision Tree, KNN, SVM, Random Forest, Naive Bayes, and Neural Networks. The Neural Network Model showed the highest accuracy and performance upon model evaluation using standard metrics.

In addition to that, we studied the work done by our fellow competitors on Kaggle and we were able to get insights from their methodologies and approach. Ensemble learning techniques that combined a variety of machine learning models and different boosting algorithms were the most common approach used in the analysis.

Despite gaining insights from various sources, our project was designed and implemented to be unique, paying attention to certain weaknesses not addressed by other contributors such as carefully selecting algorithms that handled the peculiarities of our data better, such as cost-sensitive algorithms for imbalanced datasets and implementing SMOTE sampling techniques.

**1.3 What tools and programs are already available for the problem, or closely related ones?**

The tools most relevant to this project were those employed for developing and implementing machine learning algorithms:

* The project was conducted in Jupyter Notebook using the Python programming language, which provided a robust environment for data analysis and model development.
* Most of the functions and methodologies were sourced from Pandas, NumPy, Scikit-learn, Matplotlib, Imblearn, Seaborn Library, XG Boost, and Light GBM which offered a variety of tools for data preprocessing, model development, and evaluation.
* Other frameworks, such as TensorFlow/Keras and IBM Watson Health, are available to handle projects of similar nature, but these tools are more suited for projects involving larger datasets and more complex models, such as deep learning applications or advanced patient data analysis. While these tools are powerful, they were not prioritized in this project due to our scope and objectives.

The decision to utilize classical machine learning tools was deliberate. The focus of the project was on achieving clarity and interpretability in the results, ensuring that the models were accessible and understandable to a broad audience. This approach was designed to align the project's emphasis on actionable insights and clear practical applications.

**2 Overview of the architecture**

The project was organized into four major phases:

1. **Exploratory Data Analysis (EDA)**
2. **Data Cleaning and Preprocessing**
3. **Model Development and Training**
4. **Model Evaluation**
5. **Test Data Predictions**

**A. Exploratory Data Analysis (EDA):**

The focus of the EDA phase was to gain a comprehensive understanding of the dataset:

* Analyzing the structure of the dataset (rows and columns).
* Identifying unique values and average values across features.
* Examining the distribution of data to detect patterns or anomalies.
* Computing and visualizing correlations using a correlation matrix to identify relationships between the variables.

**B. Data Cleaning and Preprocessing:**

The purpose of this phase was to prepare a high-quality dataset for machine learning analysis:

* Handling Missing data: Missing data values were handled using imputation techniques.
* Column Merging: Complementary columns were combined for simplicity and improved relevance, for example:
* Work Pressure and Academic Pressure were merged into a new column Academic or Work Pressure.
* Study Satisfaction and Job Satisfaction were merged into a Study or Job Satisfaction column.
* Feature Selection: Features with inconsistent or highly varied data, such were dropped.
* Data Encoding: String-based categorical variables were converted into numerical formats to prepare them for the models.
* Data Scaling: The dataset was scaled to standardize feature values and ensure uniformity.

**C. Model Development and Training:**

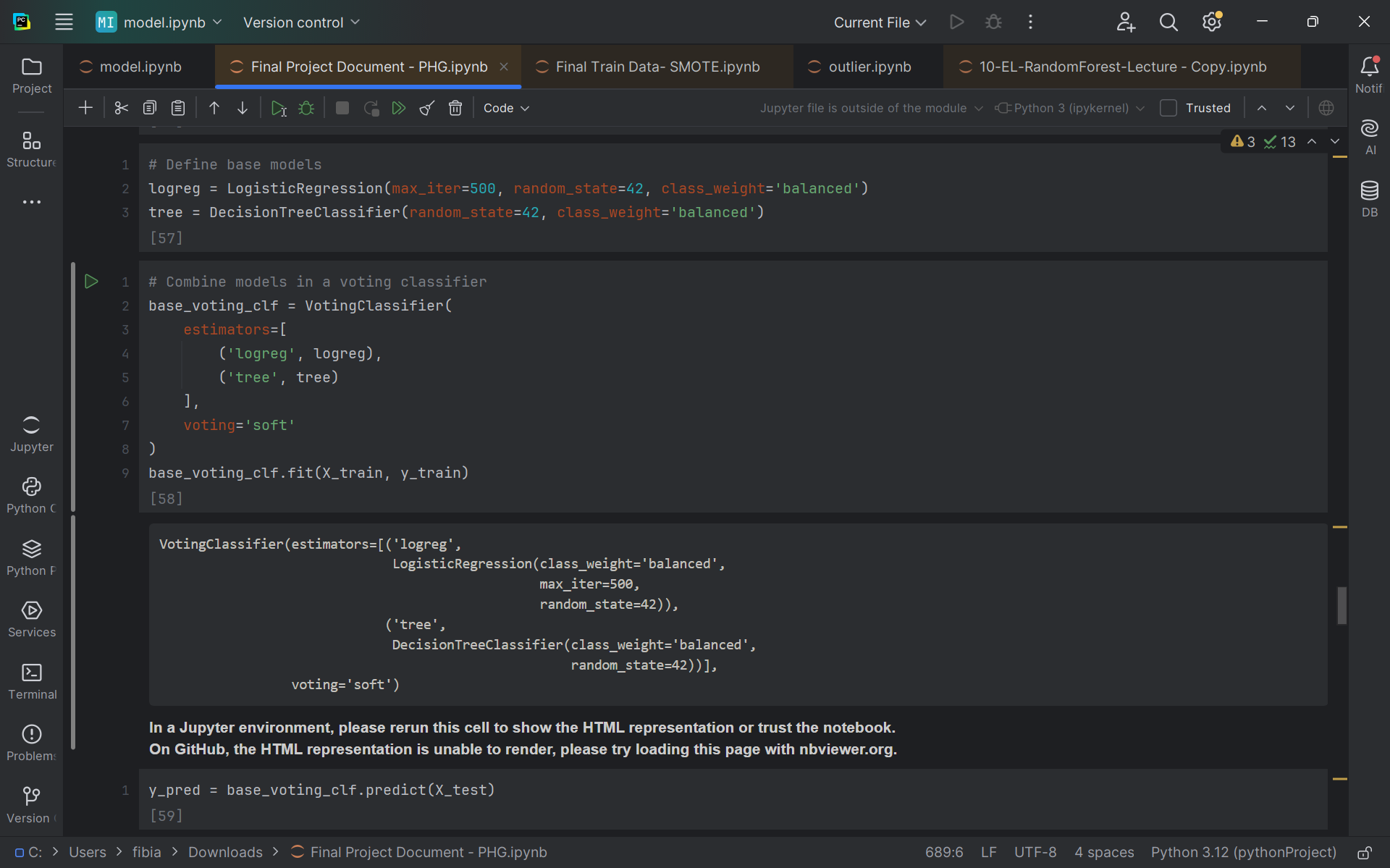
Models were trained following best practices to ensure optimal performance:

* Sampling Techniques: Techniques were applied to balance the dataset, addressing any class imbalances.
* Algorithm selection: The target variable necessitated that classification models be used for the project; hence classification algorithms were selected for the project. These machine learning algorithms were initially combined in an ensemble learning technique to leverage the collective strengths of the individual algorithms and advanced boosting models were implemented afterwards.
* Voting Classifier: A soft voting approach was chosen for the ensemble model to leverage probability outputs, which better suited the probabilistic nature of the analysis.
* Hyperparameter Tuning: The models were fine-tuned to improve accuracy and address potential weaknesses inherent in their default states.

**D. Model Evaluation:**

The models were evaluated to ensure their reliability and effectiveness. Key metrics used include:

* Accuracy, Precision, and Recall: To assess overall performance and balance between true positives and false positives/negatives.
* F1 score: Given the goal of accurately identifying cases of depression and existing class imbalance in the dataset, the F1 score was prioritized to balance precision and recall. Detecting depression cases (even at the risk of some false positives) was deemed more critical than failing to identify cases.

**2.1. Finished work: Running modules**

Logistic regression and Decision Tree Ensemble Models

Classification Report for Base Voting Classifier:

precision recall f1-score support

0 0.94 0.95 0.94 23027

1 0.74 0.71 0.73 5113

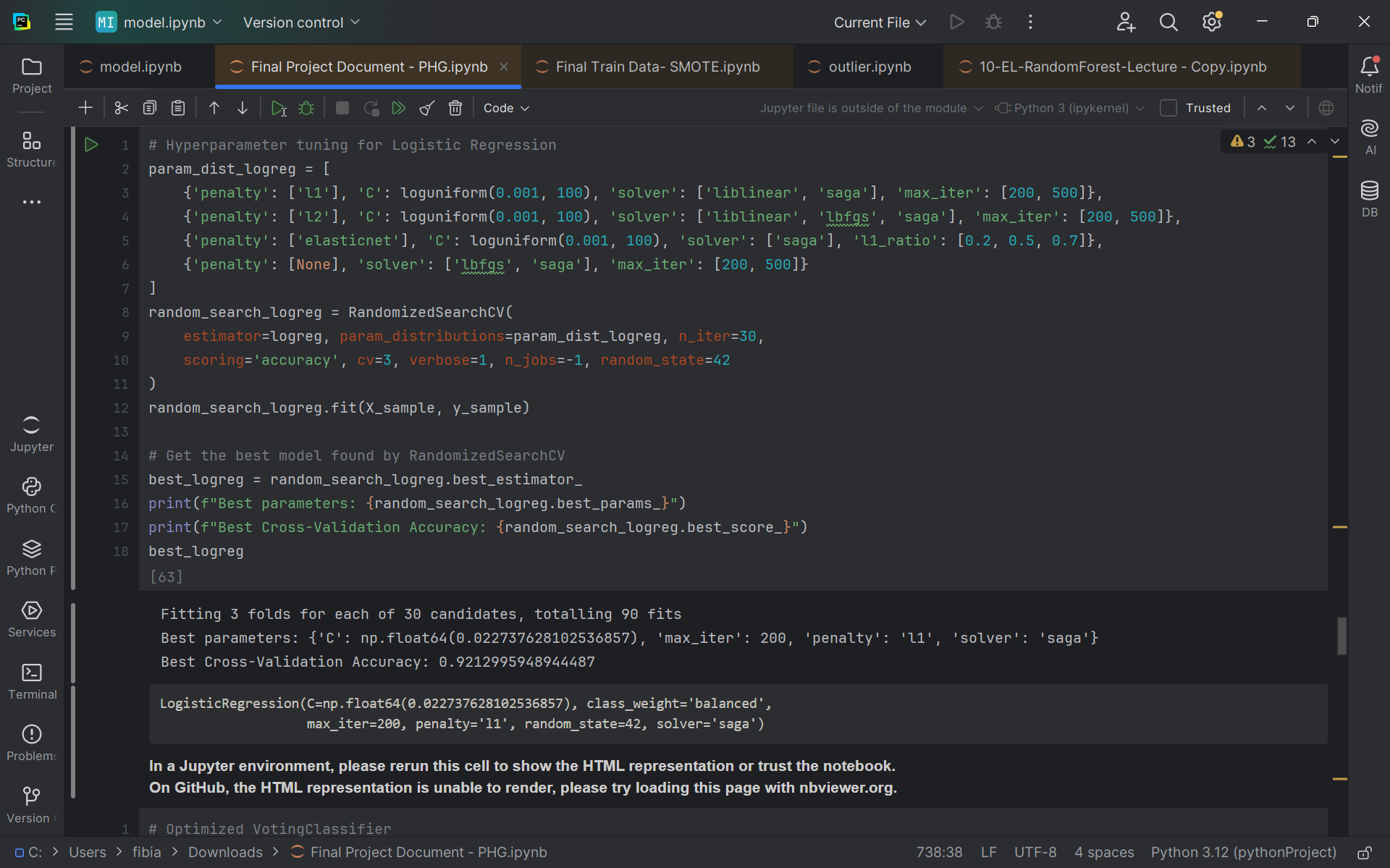
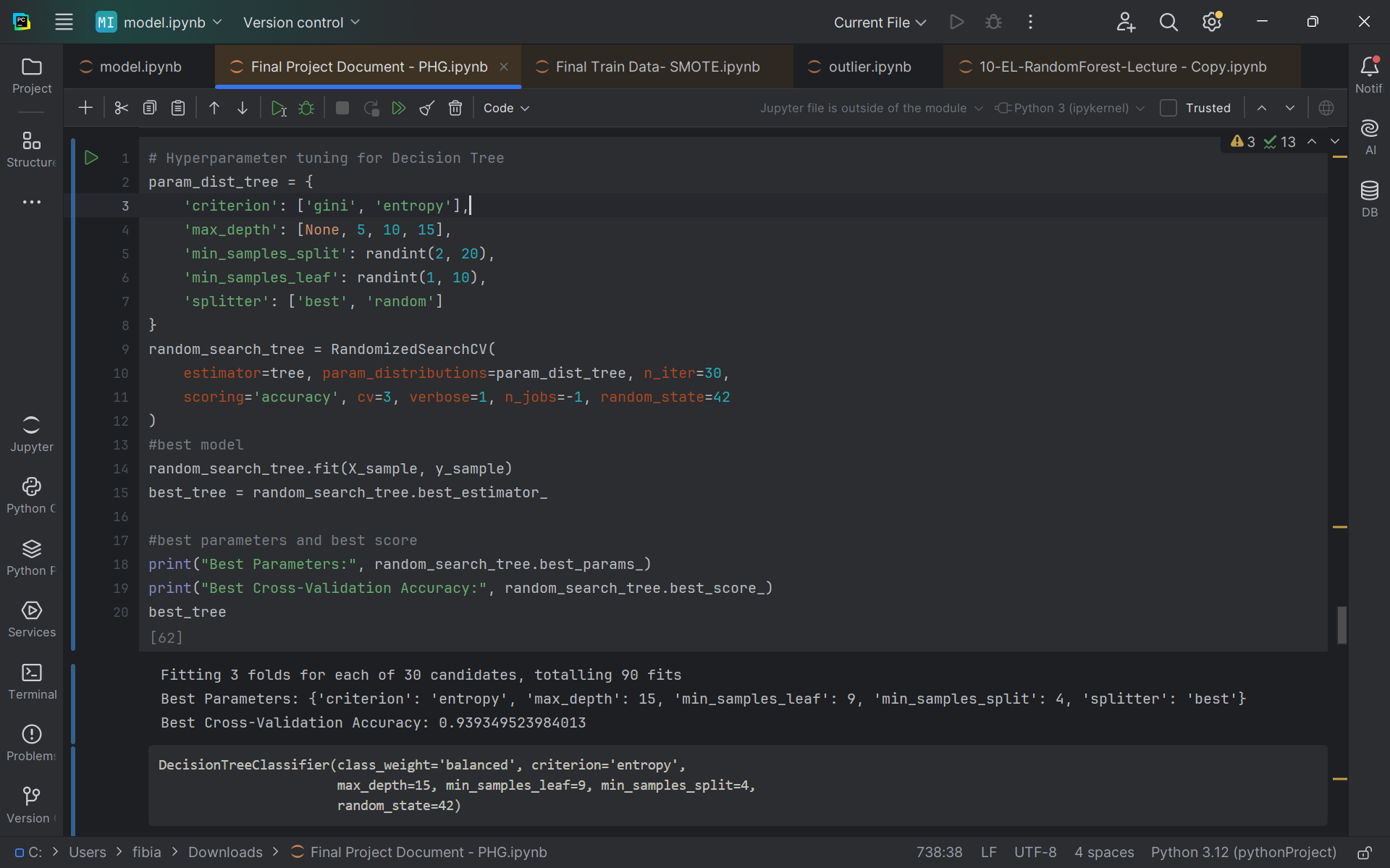
accuracy 0.90 28140

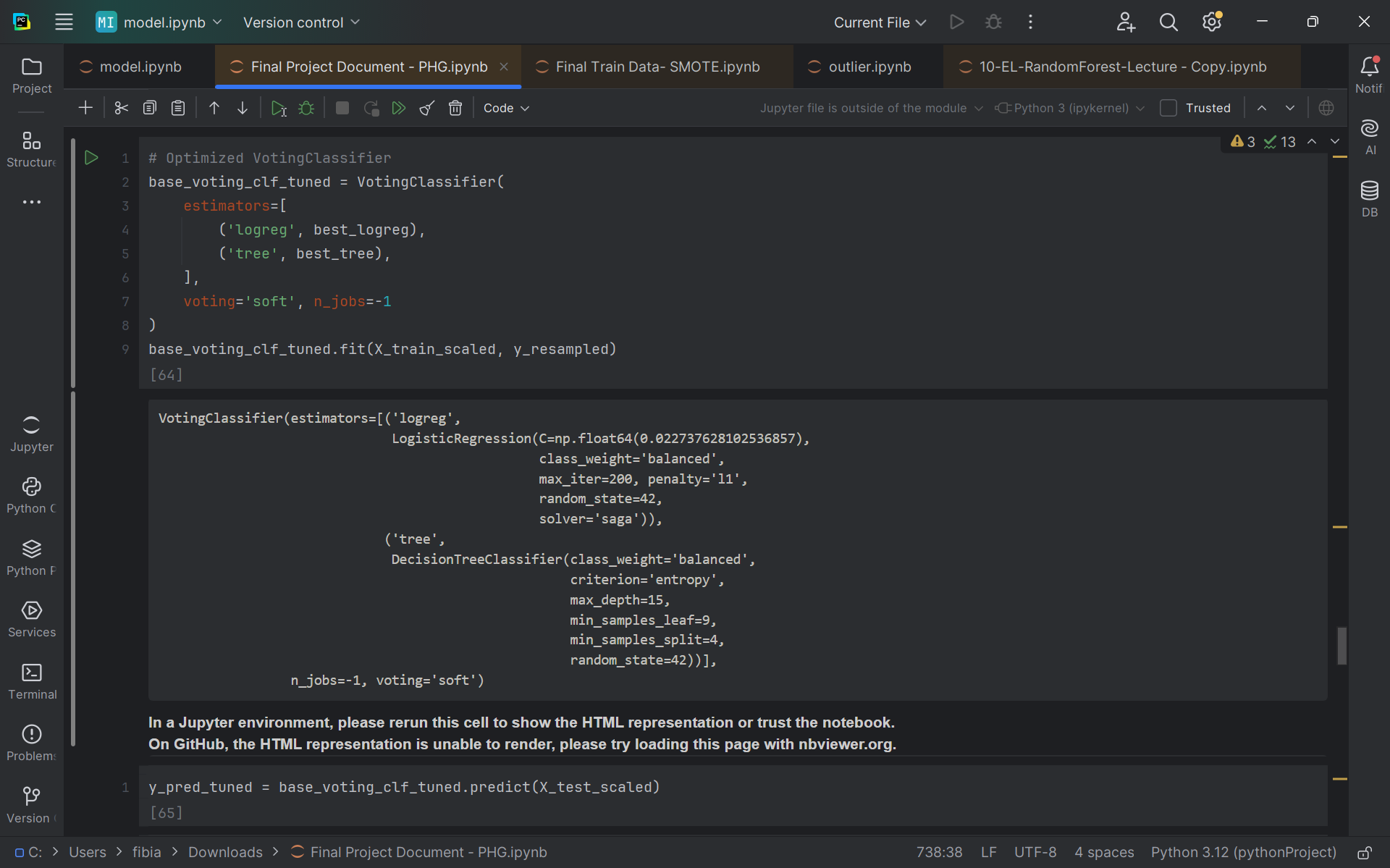
macro avg 0.84 0.83 0.83 28140

weighted avg 0.90 0.90 0.90 28140

Accuracy: **0.90** F1 Score: **0.73**

Tuned Logistic regression and Decision Tree Ensemble Models

****



Base Models Tuned Voting Classifier Report:

precision recall f1-score support

0 0.97 0.94 0.95 23027

1 0.76 0.87 0.81 5113

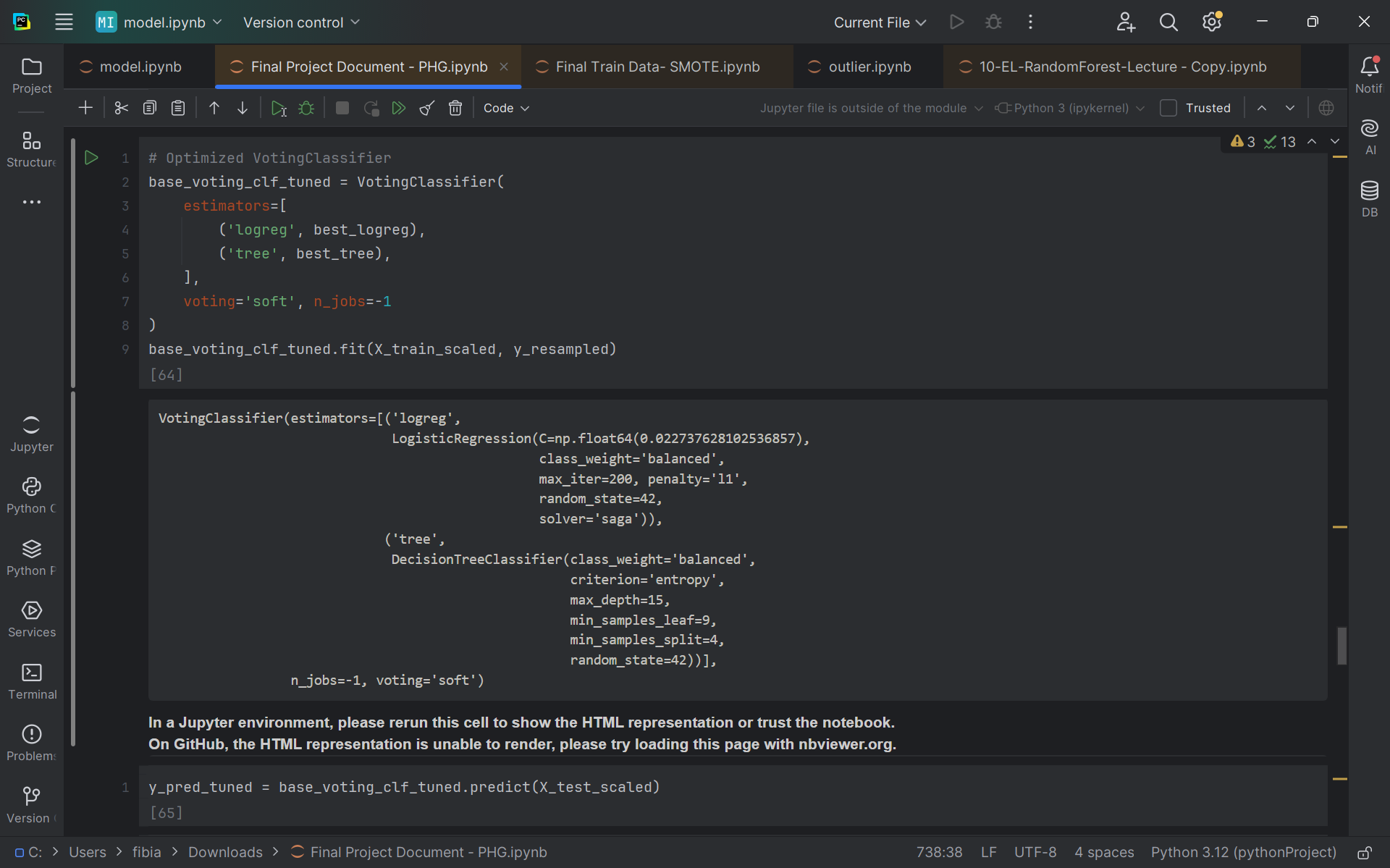
accuracy 0.93 28140

macro avg 0.87 0.90 0.88 28140

weighted avg 0.93 0.93 0.93 28140

Accuracy: **0.93** F1 Score: **0.81**

Gradient Boosting, LightGBM, and XGBoost Ensemble Models

Classification Report for Voting Classifier with All Main Models:

precision recall f1-score support

0 0.96 0.97 0.96 23027

1 0.84 0.81 0.83 5113

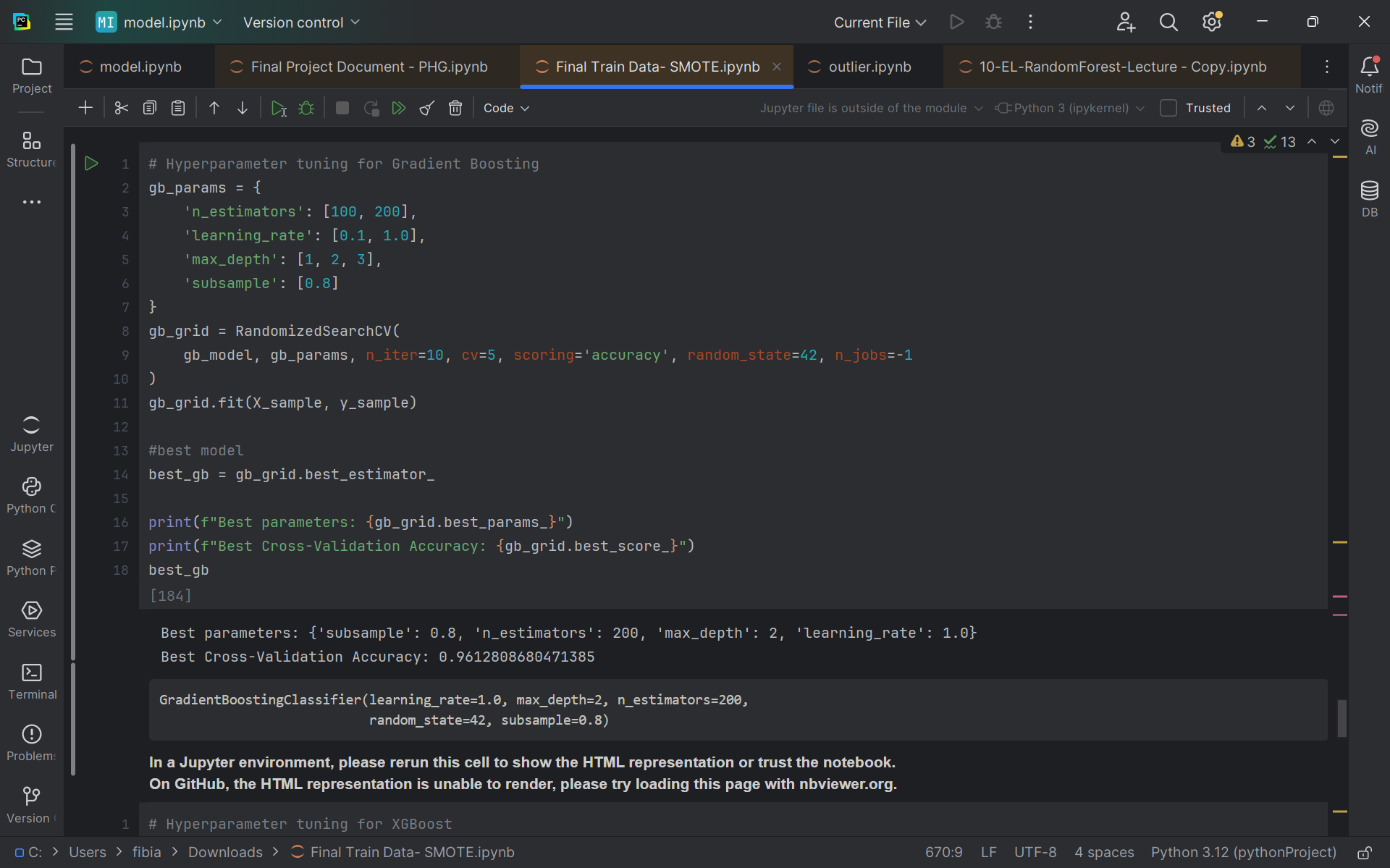
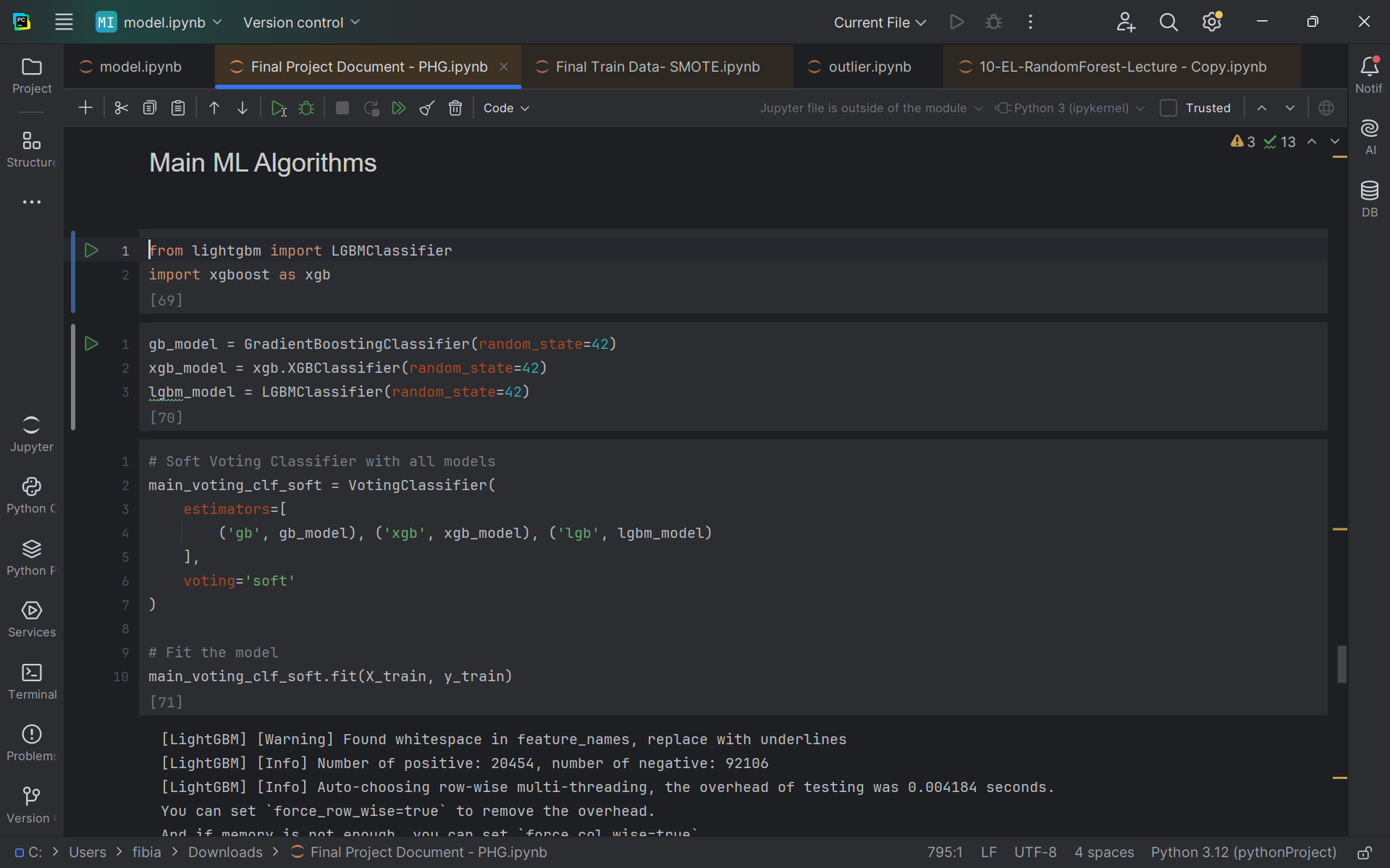
accuracy 0.94 28140

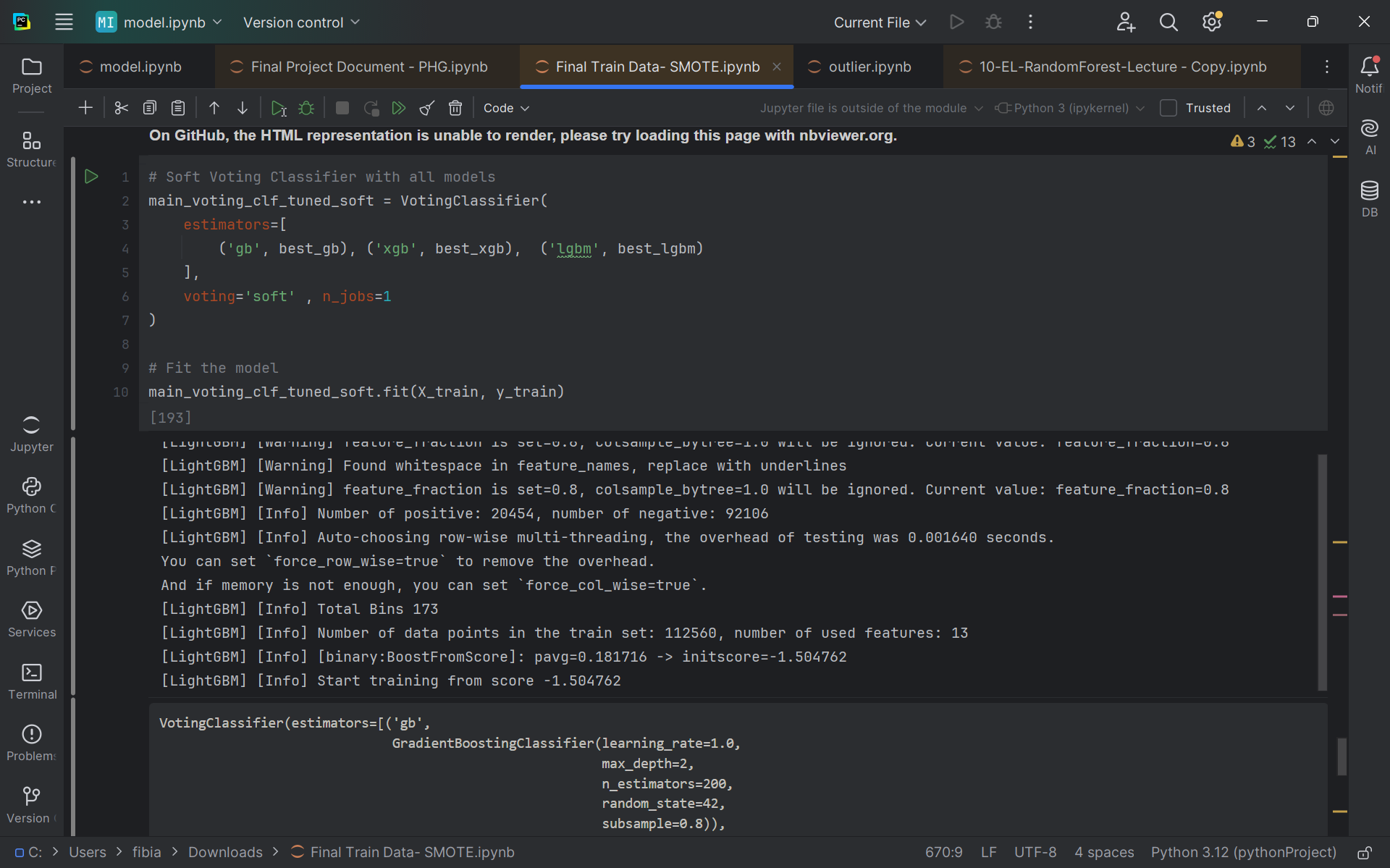
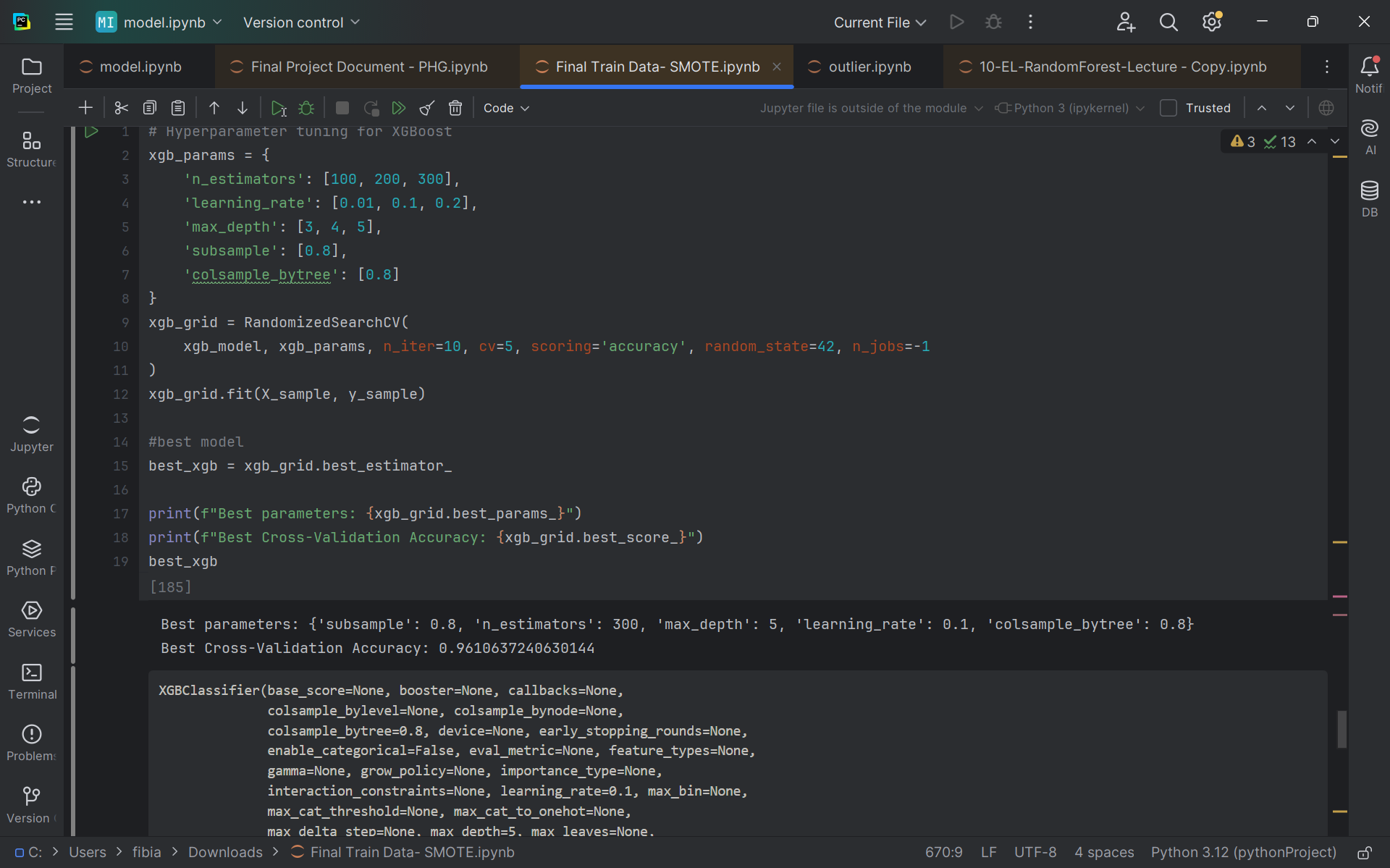
macro avg 0.90 0.89 0.89 28140

weighted avg 0.94 0.94 0.94 28140

Accuracy: **0.938**, F1 Score: **0.83**

Gradient Boosting, LightGBM, and XGBoost Ensemble Models

****

****

Classification Report for Voting Classifier with All Tuned Main ML Models:

precision recall f1-score support

0 0.96 0.97 0.96 23027

1 0.85 0.81 0.83 5113

accuracy 0.94 28140

macro avg 0.90 0.89 0.90 28140

weighted avg 0.94 0.94 0.94 28140

Accuracy: **0.94**, F1 Score: **0.83.**

**2.2 Work in progress: Modules designed but not implemented**

Support Vector Machine and Random Forest machine learning algorithms were designed but were not implemented. These models have an incredibly long runtime and put enormous demands on the computing resources of the machines used when combined with the rest of the models.

**2.3 Future work: Modules a future continuation may have**

While the project successfully developed machine learning models to predict depression, there is significant potential for future improvements. Some of these include:

* Implementation of Advanced Deep Learning Models
* A more sophisticated approach to handling data imbalances and biases
* Extensive study of individual risk factors
* Inclusion of personalized recommendations more suited to an individual's risk factors.

**3. Data Collection**

For this project, the "Exploring Mental Health Data" dataset from the Kaggle Playground Series (S4E11) was used. The target variable is “Depression”. Since this dataset was already available, it did not require raw data collection efforts. However, a series of preprocessing steps to ensure data quality and suitability for analysis were performed. Each feature offers a unique perspective, enabling us to analyze mental health from multiple dimensions like Academic Pressure, Work Pressure, Sleep Duration, Financial Stress, Study or Job Satisfaction, and Family History of Mental illness. The datasets included train and test datasets. Features provided valuable insights into various factors that could potentially be used to predict depression.

**4 Baseline and Proposed Methods**

For this project, we established a baseline to compare the performance of advanced models against simpler approaches.

**The baseline models:**

Voting Classifier - an ensemble learning algorithm that combined predictions from Logistic Regression and Decision Trees.

1. Logistic Regression
   * A fundamental algorithm for binary classification. It predicts the likelihood of depression based on weighted input features.
2. Decision Trees
   * This algorithm splits data into groups based on feature values

A soft voting approach was used, where probabilities from all models were averaged, providing more reliable predictions.

**Proposed Methods (Main Models):**

To improve predictive accuracy and robustness, we utilized advanced boosting machine learning algorithms:

1. Boosting technique that builds models sequentially, correcting the errors of prior models.

* Algorithms such as Gradient Boosting, XG Boost, and Light GBM were explored for their efficiency and high performance.

The baseline models provided a starting point for the analysis. However, the proposed methods excelled at:

* Improved predictive accuracy and performance metrics.
* Implementing regularization techniques to reduce overfitting.
* Handles imbalanced datasets effectively by focusing on minority classes.
* Capturing complex patterns in the data.

**5. Implementation**

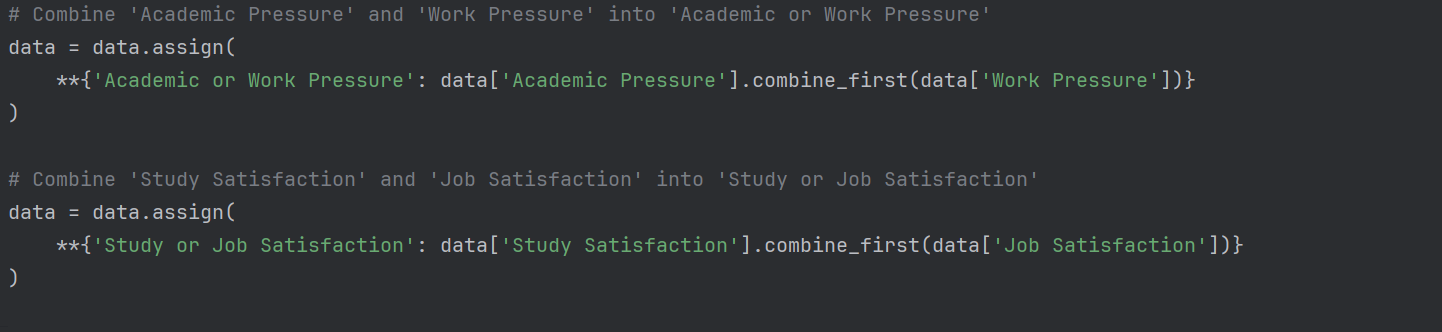
We performed exploratory data analysis (EDA) using Excel and Python to implement this project. From our findings, we identified key issues in the dataset and addressed them through comprehensive data preprocessing steps:

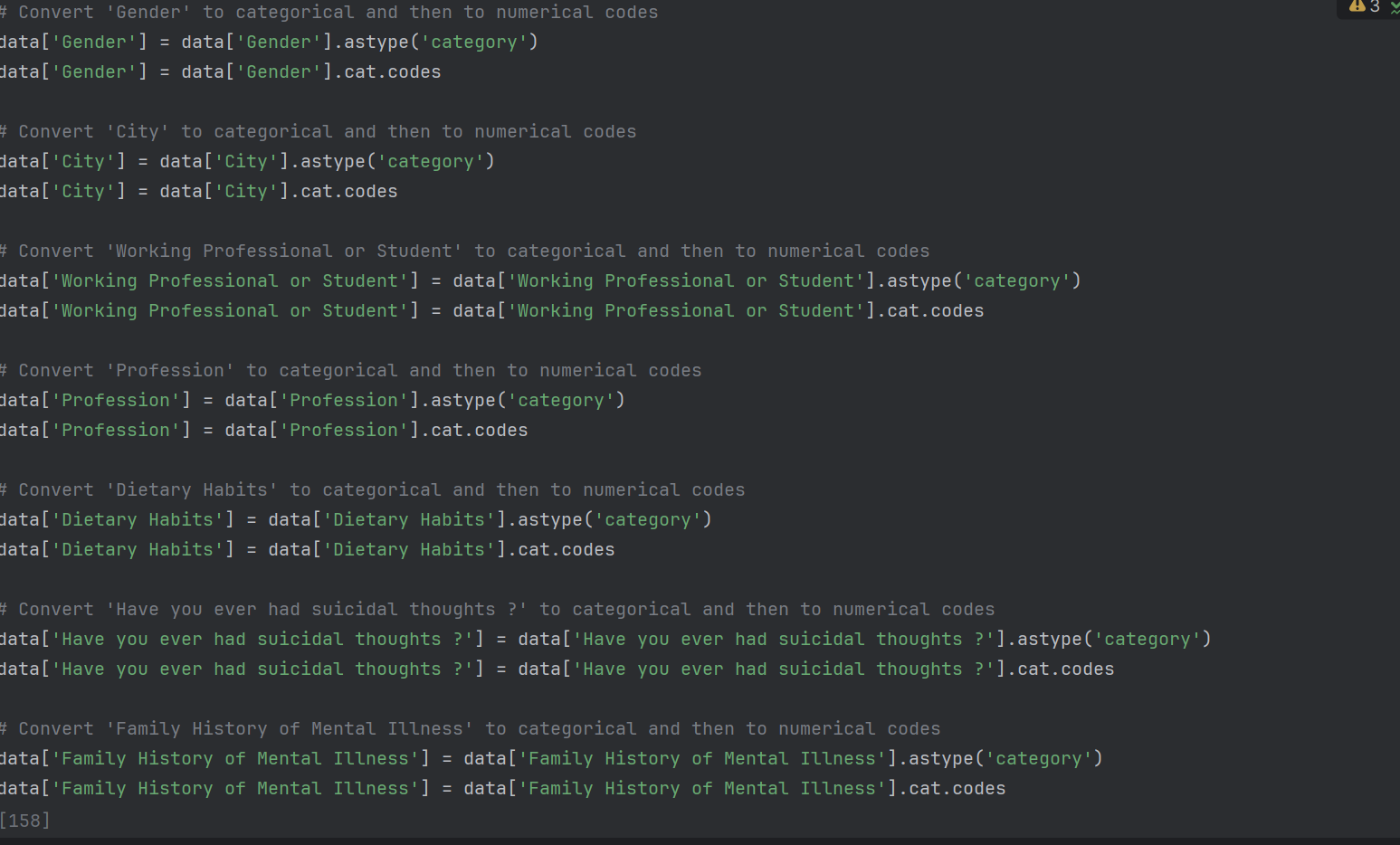
Data Preprocessing:

1. Handling Missing Data:

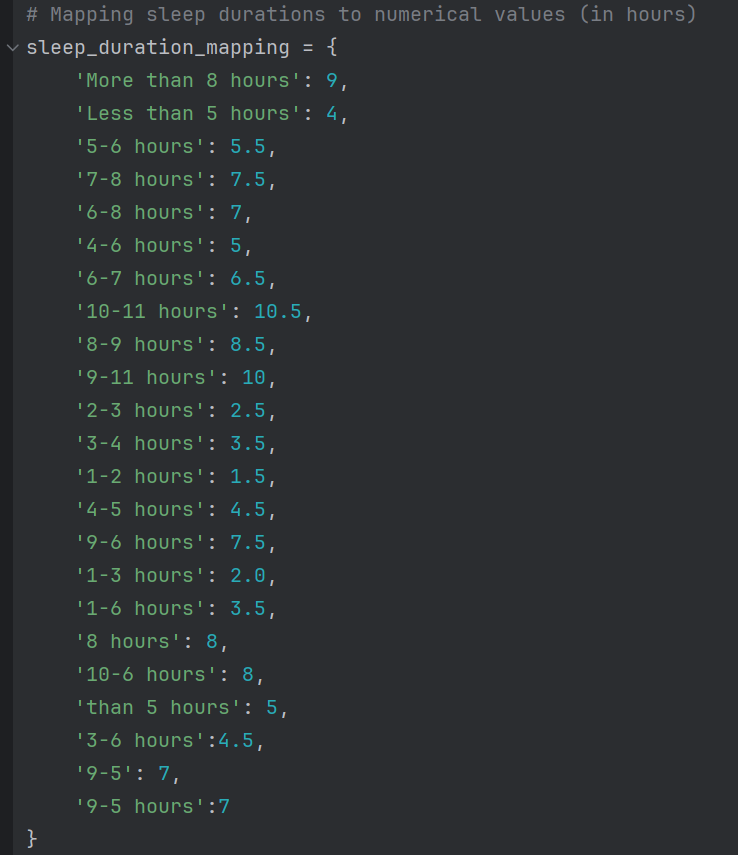
* The profession column contained ‘36630’ missing values. These were handled by filtering for the ‘student’ with no profession value and filling the profession as ‘student’, and for the rest, which was ‘working professionals’, these were filled using the mode profession value ‘Teacher’.
* The invalid and missing values for ‘cities’, dietary habits, sleep duration’, and ‘financial stress’ were replaced by the mode of valid city entries
* Ensured working professionals had no value for study satisfaction and academic pressure and vice versa for students with job satisfaction or work pressure. The empty values for these columns are filled with the mode value for the column where appropriate, i.e., student for study satisfaction or working professional for job satisfaction.

1. Feature Engineering:

* Related features such as ‘Academic & Work Pressure’ and ‘Study & Job Satisfaction’ were combined to create a new feature. 
* Categorical variables (Gender, City, Profession…) were encoded into numerical codes using target encoding (*astype('category').cat.codes*).

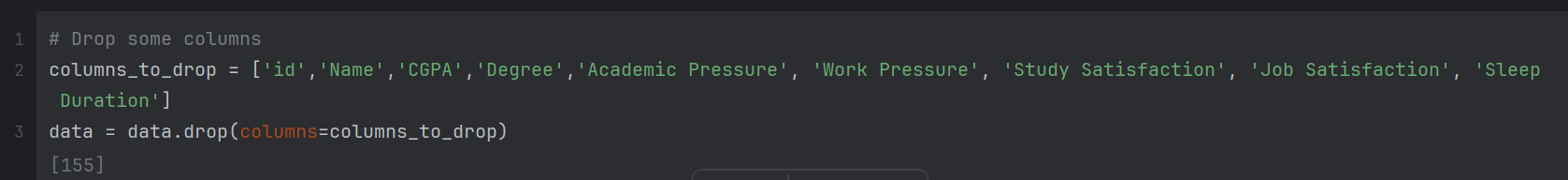


* The ‘Sleep duration’ column was mapped to numerical values for consistency and better modeling.



1. Data cleaning:

* Columns like ‘CGPA’ and ‘Degree’ were dropped due to many missing values that could not be filled logically and low correlation to the target variable. Columns like ‘name’ and ‘id’ were also removed due to irrelevance to prediction.
* Duplicate columns were removed.



1. Splitting data: The data was split into target ‘y’(depression’) and features ‘X’ (other columns).
2. Train Test Split: The ‘X’ and ‘y’ were further split into training and testing using ‘train\_test\_split ()’ with a test size of 20%.
3. SMOTE: SMOTE is applied to the split training data (‘X\_train’, ‘y\_train’) to handle imbalanced data in ‘y\_train’, ensuring that both depressed and non-depressed instances are equally represented.
4. Standardization: The resampled X\_train data and the X\_test were standardized using ‘StandardScaler()’ to a consistent scaling format for better modeling.



After the data preprocessing was done, the next steps were to build the base models, evaluate the performance, and then build the main models used for the prediction:

Model Training and Hyperparameter Tuning:

1. Baseline Model Training:

* The selected baseline model’s logistic regression and decision tree were trained on the resampled data and then combined using a soft voting classifier. The accuracy and f1 score of this model were gotten to be: 0.90 and 0.73 respectively.
* The models were tuned using ‘RandomSearchCV’ to tune hyperparameters and get the optimized models. The best models are combined in a soft voting classifier. The accuracy and f1 score of this tuned model are: 0.938 and 0.81, respectively

1. Main ML Model Training:

* The selected main model’s gradient boosting, LightGBM, and XGBoost were trained on the resampled data and then combined using the soft voting classifier. The accuracy and f1 score of this model were 0.938 and 0.83, respectively.
* The models were tuned using ‘RandomSearchCV’ to tune hyperparameters and get the optimized models. The best models are combined in a final soft-voting classifier. The accuracy and f1 score of this tuned model are: 0.94003 and 0.83, respectively

**5.1. Results and Evaluation**

The metrics we used to evaluate our models are as follows:

1. Accuracy Score: Measures the percentage of correct predictions.
2. F1 score: Particular to classification models for evaluating how well the models perform.
3. Confusion matrix: shows the true positives & negatives and false positives & negatives.

Model Comparison:

|  |  |  |
| --- | --- | --- |
| Models | Accuracy | F1- Score |
| Base voting classifier | 90% | 0.73 |
| Tuned voting classifier (baseline) | 93% | 0.81 |
| Main ML voting classifier | 93.83% | 0.83 |
| Tuned Voting classifier (Main ML) | 94.00% | 0.83 |

Base Models: From the tables above, we see a significant performance increase in the base-tuned voting classifier from the non-tuned base-voting classifier, indicating that tuning the models leads to a better performance.

Main ML Models: Applying these findings to the main, we see not much of a significant improvement but a slight improvement in the tuned voting classifier for the main models from the non-tuned voting classifier. This indicates that more hyperparameter tuning needs to be performed on the main algorithms for a better performance of the model.

Advantages of the Approach:

1. High Accuracy (94%): Indicates strong predictive power compared to simpler models.
2. Balanced F1-Score (0.83): Ensures both precision and recall are optimized, this is crucial in identifying the correct prediction for depressed individuals.
3. The use of SMOTE improved the recall significantly by addressing class imbalance.

Disadvantages of the Approach:

1. Complexity and Computation Time: Boosting models are more resource-intensive and slower to train compared to simpler models.

**A screenshot of a graph

Description automatically generated**

The above confusion matrix highlights the final model's strong overall performance with high accuracy, precision, and recall for the "Not Depressed" class. However, there is a slight trade-off in recall for the "Depressed" class, where the model could potentially improve in identifying all positive instances.

A graph with different colored bars

Description automatically generatedA screen shot of a graph

Description automatically generatedA screenshot of a computer

Description automatically generated

The analysis of feature importance across LightGBM, Gradient Boosting, and XGBoost models reveals consistent and model-specific insights. Age, Work/Study hours’, ‘Financial Stress’, and ‘Suicidal thoughts’ are common critical predictors. Each model also brings unique perspectives, such as the importance of ‘Academic or Work Pressure’ in LightGBM and Gradient Boosting and ‘Working Professional or Student’ in XGBoost.

**6. Achievements and Observations:**

|  |  |  |
| --- | --- | --- |
| Hephzibah Jayaraj | Organized and cleaned the dataset, addressing missing data and inconsistencies. Built and evaluated baseline models using Logistic Regression and Decision Tree classifiers. | Gained skills in data preprocessing, handling class imbalance with SMOTE, and model evaluation using cross-validation and ensemble methods. |
| Ginikachi Oyakhare | Cleaned and standardized the dataset, creating combined columns for pressure and satisfaction metrics. Developed optimized Voting Classifiers using Gradient Boosting, XGBoost, and LightGBM | Learned about advanced feature engineering, hyperparameter tuning with RandomizedSearchCV, and improving model performance through ensemble learning. |
| Enyojo Phoebe Alabi | Explored dataset structure and visualized data patterns. Analyzed feature importance and evaluated model predictions using confusion matrices and feature importance plots. | Enhanced understanding of data visualization, feature importance analysis, and interpreting model performance metrics to refine predictive models |

**7. Discussion and Conclusions**

This project aimed to predict depression using mental health survey data by building and evaluating multiple machine learning models. Through a systematic approach, we preprocessed the data, handled missing values, and engineered features to enhance predictive power. The models included baseline algorithms like Logistic Regression and Decision Tree and advanced ensemble models such as Gradient Boosting, XG Boost, and Light GBM. SMOTE was applied to the training data to address class imbalance, ensuring better recall for depression cases.

We achieved a final accuracy of **94%** and an **F1-Score of 0.83** using a tuned soft voting classifier, this showed that the ensemble models effectively balanced precision and recall. These results highlight the importance of hyperparameter tuning and ensemble methods in achieving high classification performance.

Key Achievements:

1. Model Performance Improvement: Tuned models significantly improved accuracy (from 90% to 94%) and F1-Score (from 0.73 to 0.83).
2. Effective Handling of Class Imbalance: The application of SMOTE helped in balancing the dataset, ensuring the model performed well on minority class predictions (Depression = 1).
3. Feature Engineering: Combining ‘Academic or Work Pressure’ and ‘Study or Job Satisfaction’ added valuable information, boosting model performance.

Challenges Faced:

1. Columns like CGPA and Degree were dropped due to high missing values, which may have resulted in the loss of potentially useful information.
2. While recall improved, maintaining high precision without sacrificing recall posed a challenge.
3. Ensemble models like Gradient boosting, XGBoost and LightGBM require a lot of computational resources and time for training and especially tuning. This limited the reruns of the model to get the optimal performance.
4. Time Constraints: more time was required to carry out more elaborate and detailed analysis.

Future Extensions and Improvements:

1. Advanced Hyperparameter Tuning: Exploration of a wider range of hyperparameters for boosting models, such as maximum depth, minimum child weight, and subsample ratios is recommended. This could potentially enhance model performance further.
2. Enhanced Handling of Missing Data: Instead of dropping columns, research alternative imputation techniques such as multiple imputation or using machine learning models to predict missing values can be explored.
3. Incorporation of more diverse datasets to improve the generalizability of the models. This could involve collecting data from different regions, age groups, or additional columns such as married/single, screen time, and much more.
4. The use of deep learning architectures, such as neural networks, could potentially capture more complex patterns in the data and improve predictive performance.
5. Extensive Research on Data Handling**:** Conduct extensive research to determine the optimal ways to handle missing data such as mental health-specific studies to identify the best practices for imputation or the use of additional data sources to fill gaps.

**8. References**

[1] L. C. López Steinmetz, M. Sison, R. Zhumagambetov, J. C. Godoy, and S. Haufe, “Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine,” *Frontiers in psychiatry*, vol. 15, no. 38690202, p. 1376784, Winter 2024, doi:<https://doi.org/10.3389/fpsyt.2024.1376784>.

This paper is about a study done in Argentina on using machine learning models to predict the emergence of depression among college kids. This study highlights the relevance of pre-existing depression and anxiety conditions in predicting depression and the potential of Machine Learning in the timely detection of at-risk students.

The most important reference to this paper in our document is in related work.

[2] Ali Akbar Jamali, C. Berger, and R. J. Spiteri, “Identification of Depression Predictors from Standard Health Surveys using Machine Learning,” *Current Research in Behavioral Sciences*, vol. 7, no. 100157, pp. 100157–100157, Jun. 2024, doi:<https://doi.org/10.1016/j.crbeha.2024.100157>.

This research was done to identify strongly correlated features with depression and on the ability of machine learning models to accurately identify predictors of depression. The study identified strongly correlated features of depression, demonstrated that ML algorithms can accurately identify depression predictors, and proposed that heterogeneous data improve ML models performance.

The most important reference to this paper in our document is in related work.

**Appendices**

A. Individual stories

Hephzibah Jayaraj:

I started by organizing and understanding the train and test datasets. I used Pandas for data manipulation, seaborn and Matplotlib for visualization, and Scikit-learn for machine learning. I also applied SMOTE to address the class imbalance, ensuring fair treatment of all data classes.

First, I explored the data to examine its structure, check for missing values, and summarize numerical features. This helped identify areas that needed cleaning and refinement. I addressed missing data in the "Profession" column by cross-referencing with the "Working Professional or Student" category and assigning "Student" or "Teacher" accordingly. I also fixed inconsistencies in the "City" column, replacing invalid entries with the most common valid city name. Additionally, I corrected misplaced values in the "Study Satisfaction" and "Job Satisfaction" columns based on whether the individuals were students or working professionals.

After ensuring the data was clean and consistent, I separated it into features and the target variable, with "Depression" as the target. I split the data into training and testing sets, maintaining an 80-20 split with stratification. I applied SMOTE to the training set to handle class imbalance and standardized the features using StandardScaler.

For baseline models, I used a Logistic Regression model with balanced class weights and a Decision Tree Classifier, both accounting for class imbalance. These models were combined into an ensemble soft-voting classifier and trained on the prepared data. I evaluated its performance on the test data using metrics like accuracy and F1 score.

To refine the models, I performed hyperparameter tuning using RandomizedSearchCV. I optimized the Decision Tree and Logistic Regression models by exploring various parameters and assessing their performance using cross-validation accuracy. The best-performing configurations were used to enhance the predictive system, ensuring robust performance for real-world applications.

Ginikachi Oyakhare:

While cleaning the dataset, I noticed inconsistencies in the "Academic Pressure" and "Work Pressure" columns. For working professionals, I moved values from "Academic Pressure" to "Work Pressure," and for students, I did the opposite. I then filled any remaining gaps with appropriate defaults: 3.0 for students and 2.0 for working professionals.

To simplify the data, I created two new columns: "Academic or Work Pressure" and "Study or Job Satisfaction," which combined relevant information. I also standardized the "Sleep Duration" column by converting various descriptions into consistent numerical values in a new column, "Sleep Duration (Hours)," and filled missing values with 5 hours.

For "Dietary Habits," I replaced invalid entries with the most common value, "Moderate," and ensured no gaps remained. I reviewed key columns related to mental health, filling any missing values with appropriate defaults, such as 2.0 for "Financial Stress." I also explored the relationship between "CGPA" and depression by calculating their correlation.

After cleaning and transforming the dataset, I focused on building predictive models. I created an optimized Voting Classifier by combining the best-tuned Logistic Regression and Decision Tree models using soft voting. I trained this ensemble on the resampled training data and evaluated its performance on unseen data.

Next, I built a new Voting Classifier combining Gradient Boosting, XGBoost, and LightGBM models. I trained and tested this ensemble, focusing on its ability to predict depression accurately.

To further refine the models, I performed hyperparameter tuning on Gradient Boosting, XGBoost, and LightGBM using RandomizedSearchCV. I optimized key parameters like the number of estimators and learning rate. The final tuned ensemble demonstrated improved performance, accurately identifying depression cases.

Enyojo Phoebe Alabi:

I began by exploring the dataset's structure and data types using `data.info()` and `data.dtypes`. I converted categorical columns like "Gender," "City," "Profession," and "Dietary Habits" into numerical codes.

Next, I visually examined the dataset by creating box plots to identify outliers and unusual patterns. I calculated a correlation matrix and visualized it as a heatmap to uncover relationships between features. To check for normality, I conducted a Shapiro-Wilk test on a subset of 200 rows. I also compared the distribution of each feature between individuals with and without depression using kernel density plots.

I then analyzed the skewness of numerical features, organizing the findings into a table from most to least skewed. This analysis helped guide the preparation of the dataset for machine learning.

To understand the model predictions we built, I printed the feature importances for tree-based models, LightGBM, and Gradient Boosting. I also generated and displayed the confusion matrix for the tuned soft-voting classifier, showing the true positives, true negatives, false positives, and false negatives.

I visualized feature importance for LightGBM, Gradient Boosting, and XGBoost using bar plots to understand each feature's contribution. For the next phase, I loaded the test data, separated the ID column, and used the tuned soft-voting classifier to predict depression. I saved these predictions in a new DataFrame, combining IDs and predicted values.

Finally, I checked the distribution of predicted classes and saved the DataFrame to a CSV file for future use.