Final Report

After evaluating all the models and ensemble techniques, the Bayesian Model Averaging (BMA) method emerged as the most effective performer. BMA surpasses the other methods as it does not presume that all models possess equal quality. Rather, it assigns greater importance to the models that best fit the validation data, resulting in more precise and well-calibrated predictions. Consequently, BMA generally attained the highest ROC AUC scores and the most balanced precision, recall, and F1 values across various age-group targets.

The simple averaging ensemble followed as the second-best performer. Although it grants equal influence to each of the top three models, averaging still aids in minimizing noise and smoothing out extreme predictions. This characteristic rendered it more stable and often superior to relying on individual models, yet it was not as robust as BMA since it cannot account for one model being significantly better or worse than the others. The soft VotingClassifier typically ranked third, as it treats all selected models equally, which is not optimal when model performance differs.

In terms of individual models, Random Forest and SVC generally exhibited the best performance, with Logistic Regression also performing surprisingly well due to its stability. Nevertheless, no single model consistently outperformed the ensemble methods. In summary, the findings indicate that model combination results in more dependable predictions, with BMA being the most effective due to its intelligent weighting of each model based on its actual generalization capability.

Contributions

Albert-For my part of the project, I focused on preprocessing the dataset and preparing it for modeling. Since the original cleaning steps were already completed, my main task was to finalize the formatting, ensure that all features were correctly encoded, and verify that the data was ready for machine learning models. After confirming that missing values, categorical variables, and scaling had been handled properly, I completed the dataset split. Following the assignment requirements, I divided the data into a 70% training set, a 15% validation set, and a 15% test set. This split ensures that the models can be trained effectively while still providing fair and unbiased assessments during validation and testing.

Hoang- In the project I trained all of the required classification models using the training set. I implemented six different models: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting (using either XGBoost or LightGBM), K-Nearest Neighbors, and a Support Vector Classifier. Each model was fit using the preprocessed training data, and default hyperparameters were used unless otherwise specified. By training all six algorithms, I helped create a broad

comparison across linear, tree-based, ensemble, and distance-based classifiers. After training each model, I saved them so they could be used by the next person for evaluation and comparison.

Alexis-I evaluated and compared all six trained models using the validation dataset. I calculated key performance metrics including accuracy, precision, recall, and F1-score for each classifier. When applicable, I also computed ROC–AUC scores to measure how well the models separated the classes. After gathering these metrics, I organized them into a comparison table to make performance differences clear. From this analysis, I identified the top three models overall, based on balanced scores across accuracy and F1. This provided the foundation for the ensemble stage and helped us understand which models generalized the best on unseen data.

David- My contribution to the project involved building and evaluating ensemble models using the top three classifiers selected during validation. First, I created an ensemble using either a Voting Classifier or an average of prediction probabilities. I tested this ensemble on both the validation and test sets to measure how much performance improved compared to the individual models. In addition to the traditional ensemble, I also implemented a Bayesian ensemble model to compare probabilistic weighting methods against simple voting or averaging. After evaluating both approaches, I compared their accuracy, precision, recall, and F1-scores. This allowed us to see whether Bayesian averaging provided more stable and reliable predictions than the standard ensemble technique.