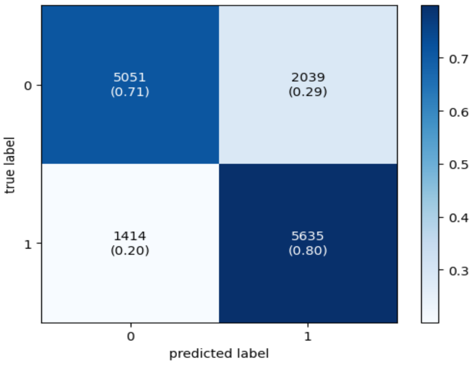
#Todo:

1. Integrate all parts into Latex.
   1. Add intro part
   2. Update the figure of xgb in Approach
2. Overall check and polish

Approach (add AdaBoost)

AdaBoost

We performed hyperparametric tuning of the AdaBoost classifier. This involved using grid search to optimize three key parameters. After finding the optimal combination of parameters, we trained the AdaBoost model using these optimal settings on training data. Finally, we evaluated the performance of the optimized model on a test set.

图表, 条形图

描述已自动生成

*Figure 11.* AdaBoost Confusion Matrix *Figure 12.* AdaBoost Feature Importance

Conclusion

Model Accuracy

图表, 条形图

描述已自动生成图表, 条形图

描述已自动生成

图表, 条形图

描述已自动生成图表, 条形图

描述已自动生成

The combined accuracy of all models ranges from 0.72558 to 0.75536. AdaBoost surpasses other models in terms of both accuracy (0.75536) and F1 score (0.75502), making it the most compelling model in estimating diabetes. Its high-rate precision (0.75694) and recall (0.75536) indicate a balanced performance, which is essential for practical applications in the medical field. Despite AdaBoost's immensely competitive success, Random Forest and XGBoost also show respectable results, with XGBoost reaching the second highest F1 score. This suggests that ensemble methods, in general, are more effective for this application due to their ability to handle complex patterns and data imbalances better than simpler models like Logistic Regression.

图表, 条形图

描述已自动生成

The Mean Squared Error (MSE) measurements indicate that although AdaBoost is the most accurate and has the highest F1 score, the overall model misclassification error does not vary greatly or is very similar across all models. This suggests that improvements in model performance may also be obtained from further feature engineering or data preprocessing to reduce these errors.

It is worth noting that two boosting methods performs better than other methods. The reason why boosting methods like AdaBoost and XGBoost often perform better, particularly in complex predictive tasks like diabetes prediction from survey data, can be attributed to several key characteristics of boosting algorithms.

First, many traditional models like decision trees tend to overfit the data, showing high variance but low bias, whereas models like logistic regression may underfit, showing high bias but low variance. Boosting methods manage to balance bias and variance effectively. They build on top of weak models (typically high bias) and by focusing iteratively on reducing errors, they also manage to keep variance in check. Also, Boosting is designed to sequentially correct the mistakes of prior models and combine many weak learners (models that perform only slightly better than random guessing) to create a strong learner (Chen & Guestrin, 2016). This approach effectively reduces both the bias (error due to erroneous assumptions in the learning algorithm) and variance (error from sensitivity to small fluctuations in the training set) of the model. In addition, according to Hastie et al. (2009) and Murphy (2012), boosting method can naturally process continuous and classified data, and can naturally process missing data.

Analysis Result

图表

描述已自动生成图表, 条形图

描述已自动生成*Figure 10.* XGBoost Feature Importance *Figure 12.* AdaBoost Feature Importance

Examining the outcome with the two models that have the highest accuracy scores, we utilized feature importance plots from both the AdaBoost and XGBoost models to derive multiple inferences about the importance of different predictors in determining the likelihood of someone being diagnosed with diabetes.

In the XGBoost model, high blood pressure (HighBP) is the most influential variable,

indicating a strong association with diabetes. This factor is still important in the AdaBoost model but ranks slightly lower.

The AdaBoost models, which place an emphasis on sex and body mass index as well as on their interactions, show the prediction that obesity and increased body mass index are strong predictors of increased diabetes risk. The AdaBoost model therefore revealed that these two factors markedly influence diabetes risk, and that by looking at sex as well as body mass index their combined effect on diabetes risk is certainly not simply additive.

The two models each individually and consistently identify Cholesterol Levels (HighChol) as well as General Health (GenHlth) as a factor for prediction. However, their orders of importance are different. This underscores the significance of lifestyle factors in diabetes risk, and Age shows up to be a significant factor in both models, as it should be since diabetes risk growth exponentially with age.

Reference

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowl- edge

Discovery and Data Mining, KDD 』16, pages 785–794, New York, NY, USA. AC

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Series in Statistics. Springer.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. The MIT Press.