## **Introduction**

Model comparison was a critical step in refining my approach to identifying Texas adults at high risk for diabetes. While logistic regression served as an effective baseline model due to its interpretability, I explored random forest to improve recall and enhance model robustness. This evaluation aimed to identify a model that could better support early intervention and public health outreach, especially given the real-world consequences of false negatives in chronic disease prediction.

## **Model Comparison Results:**

I conducted a comparative analysis between logistic regression and random forest models using the same preprocessed dataset and consistent feature set. This feature set included domain-specific variables such as Chronic Risk Load and the Healthcare Barrier Index, both of which are designed to capture complex behavioral and clinical patterns relevant to diabetes risk. The goal was to evaluate which model best supported the early identification of high-risk individuals in a public health context.

The models were assessed using five key performance metrics: accuracy, precision, recall, F1-score, and ROC AUC. These metrics were selected to ensure a balanced evaluation of overall performance, as well as the model's ability to correctly identify at-risk individuals (recall) and maintain reliability (AUC).

The results clearly favored the random forest model across all evaluated metrics. Accuracy improved from 0.82 with logistic regression to 0.85 with random forest. Precision rose from 0.69 to 0.72, while recall—a particularly important metric in health outreach—increased significantly from 0.56 to 0.66. F1-score, which balances precision and recall, improved from 0.62 to 0.69. Finally, the ROC AUC climbed from 0.79 to 0.84, demonstrating a stronger overall classification ability.

These performance gains, especially the improvement in recall, are critical. In a healthcare setting, failing to identify individuals at high risk for diabetes can mean missed opportunities for early intervention. The random forest model’s higher sensitivity makes it a more suitable candidate for supporting proactive public health strategies.

## **Evaluation and Interpret Results:**

The random forest model’s superior recall, F1-score, and ROC AUC support its deployment in a public health setting. Despite logistic regression’s transparency, its lower sensitivity is problematic when identifying high-risk individuals is the core goal. Random forest’s capacity to model nonlinear relationships between predictors (e.g., BMI and general health) made it better suited to the behavioral and clinical complexity of the BRFSS dataset.

## **Hyperparameter Tuning:**

I used randomized search with cross-validation to tune parameters such as "n\_estimators", "max\_depth", "min\_samples\_split", and "max\_features". This improved recall by approximately 3% and yielded a more stable model across folds. I intentionally avoided deep trees to maintain interpretability and reduce variance, finding a balance between complexity and generalization.

## **Generalization Assessment:**

Using 5-fold cross-validation and a holdout test set, I evaluated the model’s consistency. The recall’s standard deviation across folds remained below 0.03, and performance on the test set closely matched cross-validation results. Additionally, ROC curves showed consistent separation across folds, reinforcing the model’s reliability on unseen data.

## **AI Feedback:**

I used ChatGPT to audit my modeling approach and raise specific questions about metric prioritization and alternative models. Key takeaways included:

* **Prioritizing Recall:** Justified recall as the key metric in health outreach, where missing true positives undermines prevention goals.
* **Next Models:** XGBoost and LightGBM were suggested as promising future candidates.
* **Sampling Techniques:** SMOTE and other oversampling methods were flagged as options to revisit if performance plateaus.
* **Ensemble Learning:** Combining models might improve results while maintaining generalization.

In response, I emphasized recall in my final recommendation and chose to defer oversampling since class imbalance (~14%) was moderate and did not yet hinder sensitivity.

## **What I’d Like Feedback On:**

* Given the importance of sensitivity in chronic disease prevention, do you agree with prioritizing random forest over logistic regression despite its reduced interpretability?
* Would you recommend proceeding to XGBoost now, or should I first solidify deployment strategies for the current model?
* Do you see value in integrating ensemble or hybrid approaches at this stage, or should that be reserved for future iterations?