Dataset

- 70,692 survey responses from the 2015 CDC Behavioral Risk Factor Survey.
- Pre-cleaned: no missing values, standardized formatting.
- Balanced: ~50% positive diagnoses, ~50% negative.
- 17 predictor features (mix of categorical and numeric, stored as floats).
- Some multicollinearity observed between features.
- Top 5 features most correlated with diabetes: **General Health, High Blood Pressure, BMI, High Cholesterol, Age**.

Models Implemented

- **Naive Bayes**
- **Decision Tree** (optimized with GridSearchCV)
- **Random Forest** (optimized with RandomizedSearchCV)

Results

- Accuracy:
    - Decision Tree → **0.738**
    - Random Forest → **0.745**
    - Naive Bayes → **0.719**

- Recall:
    - Decision Tree → **0.763**
    - Random Forest → **0.782**
    - Naive Bayes → **0.713**

Model Comparison

Naive Bayes

- Performed worst, with lowest accuracy and recall.
- Assumes feature independence, which was violated due to multicollinearity.
- This limitation prevents it from modeling interactions between related variables, leading to weaker probability estimates and predictions.

Decision Tree

- Naturally captured feature interactions by splitting on the most predictive variables at each level.
- More sensitive to outliers, but large dataset size minimized this issue.
- With `max_depth = 8`, overfitting was reduced, keeping splits statistically meaningful.

Random Forest

- Outperformed Decision Tree slightly, offering higher recall and accuracy.
- Improvement came from averaging multiple randomized trees, reducing variance and increasing robustness to outliers.
- The performance gap was smaller than expected, likely because the dataset size was large enough for a single decision tree to generalize effectively.

Conclusion

- ***Random Forest was the best overall model***, achieving the highest recall and accuracy, making it the most reliable predictor of diabetes in this dataset.
- The relatively small performance gap between Decision Trees and Random Forests highlights the strength of tree-based methods on large, balanced datasets.
- Naive Bayes underperformed due to its strong independence assumption