Diabetes Risk Prediction

Empowering Early Detection and Management Within Children

Alex Gerwer

Overview

This groundbreaking project confronts the escalating global health crisis of diabetes by harnessing the power of advanced machine learning. We have developed sophisticated predictive models specifically designed for the early identification of diabetes risk, with a crucial focus on safeguarding the health of younger generations. By intelligently analyzing readily accessible health indicators—such as age, BMI, blood pressure, lifestyle choices, and relevant medical history—our approach provides a powerful tool for proactive healthcare. This isn't just about prediction; it's about empowerment. Our rigorously validated models enable timely, personalized interventions and foster effective preventative strategies long before the disease manifests fully. This initiative represents a significant leap forward, aiming to shift the paradigm from reactive treatment to proactive prevention, ultimately securing better long-term health outcomes and enhancing the well-being of our youth.

Methodology: From Vision to Validation

Our project progressed through a rigorous, iterative methodology, translating strategic planning into demonstrable results. Each phase built systematically on the previous, driven by empirical evidence and analytical refinement.

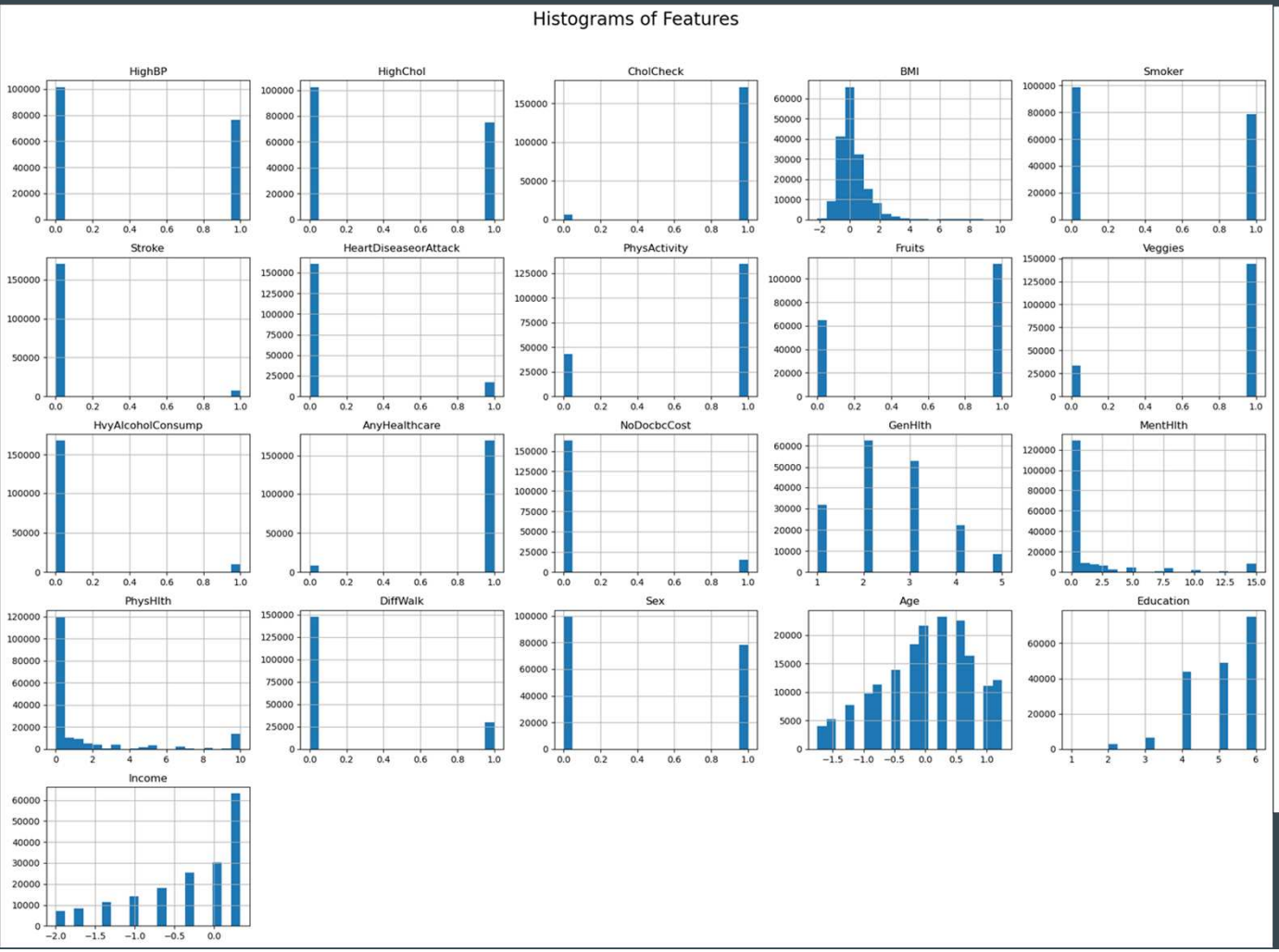
Laying a Robust Data Foundation: We began by evaluating three distinct dataset variations provided. After careful analysis comparing ternary vs. binary target encodings and dataset completeness, we strategically selected the most comprehensive file (binary target, full dataset) as the optimal foundation. Following this selection, meticulous data preparation prioritized integrity, employing sophisticated KNN imputation to address outliers without discarding valuable information, and establishing crucial performance baselines.

Harnessing EDA for Strategic Refinement: Exploratory Data Analysis was central, moving beyond visualization to actively guide strategy using the chosen dataset. Insights from EDA directly informed targeted feature engineering and a disciplined feature selection process, including rigorous VIF analysis to mitigate multicollinearity and enhance model efficiency.

Systematic Model Development & Optimization: Building on this solid groundwork, we systematically explored and evaluated a diverse range of machine learning algorithms. This comparative approach, coupled with iterative optimization, ensured the identification of models offering the highest predictive accuracy and reliability for identifying diabetes risk within the selected data context.

This disciplined, data-centric journey, starting with deliberate dataset selection, culminated in the development of demonstrably effective and impactful predictive models, showcasing our commitment to delivering robust, actionable healthcare insights.

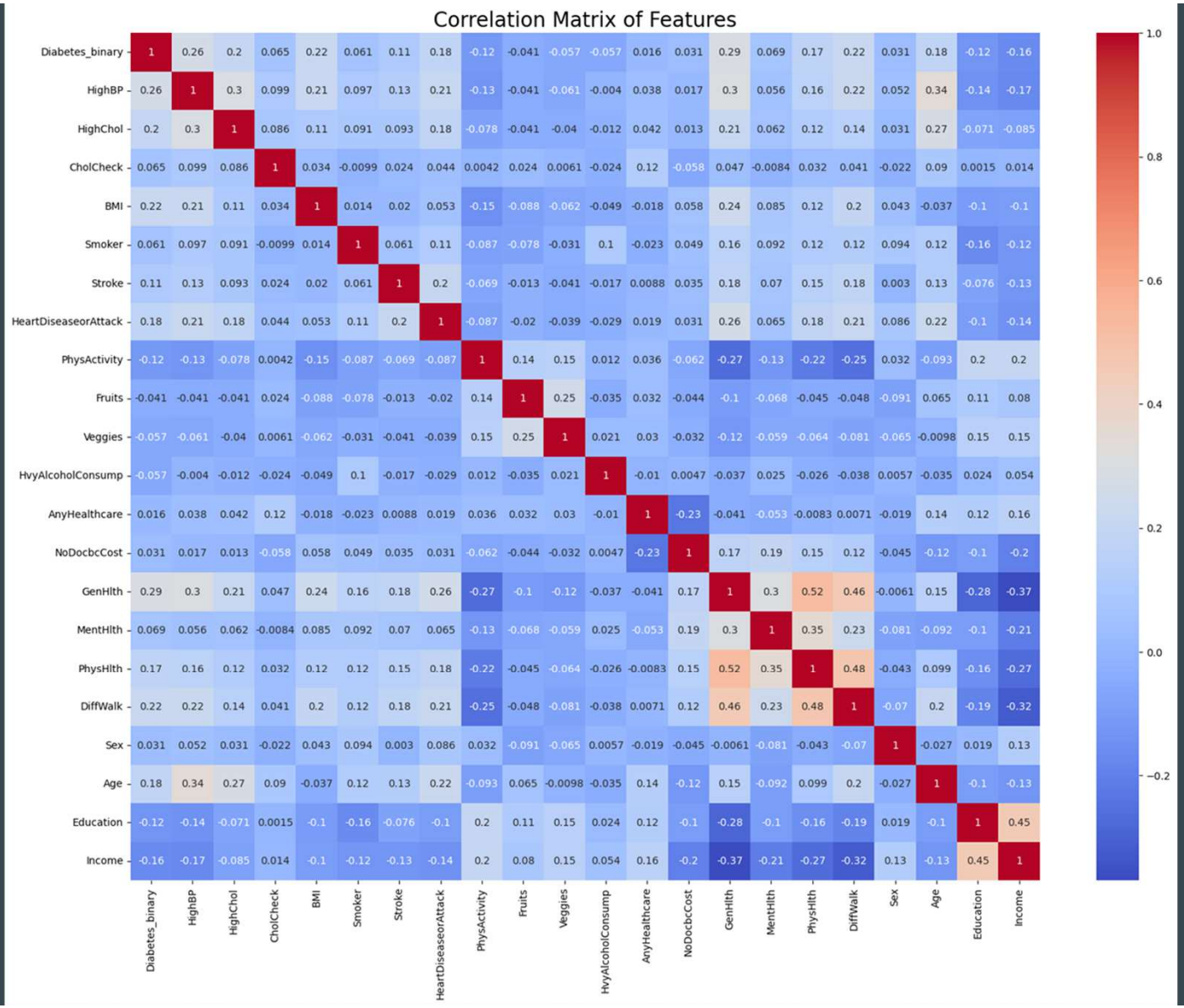
Raw Data Set:

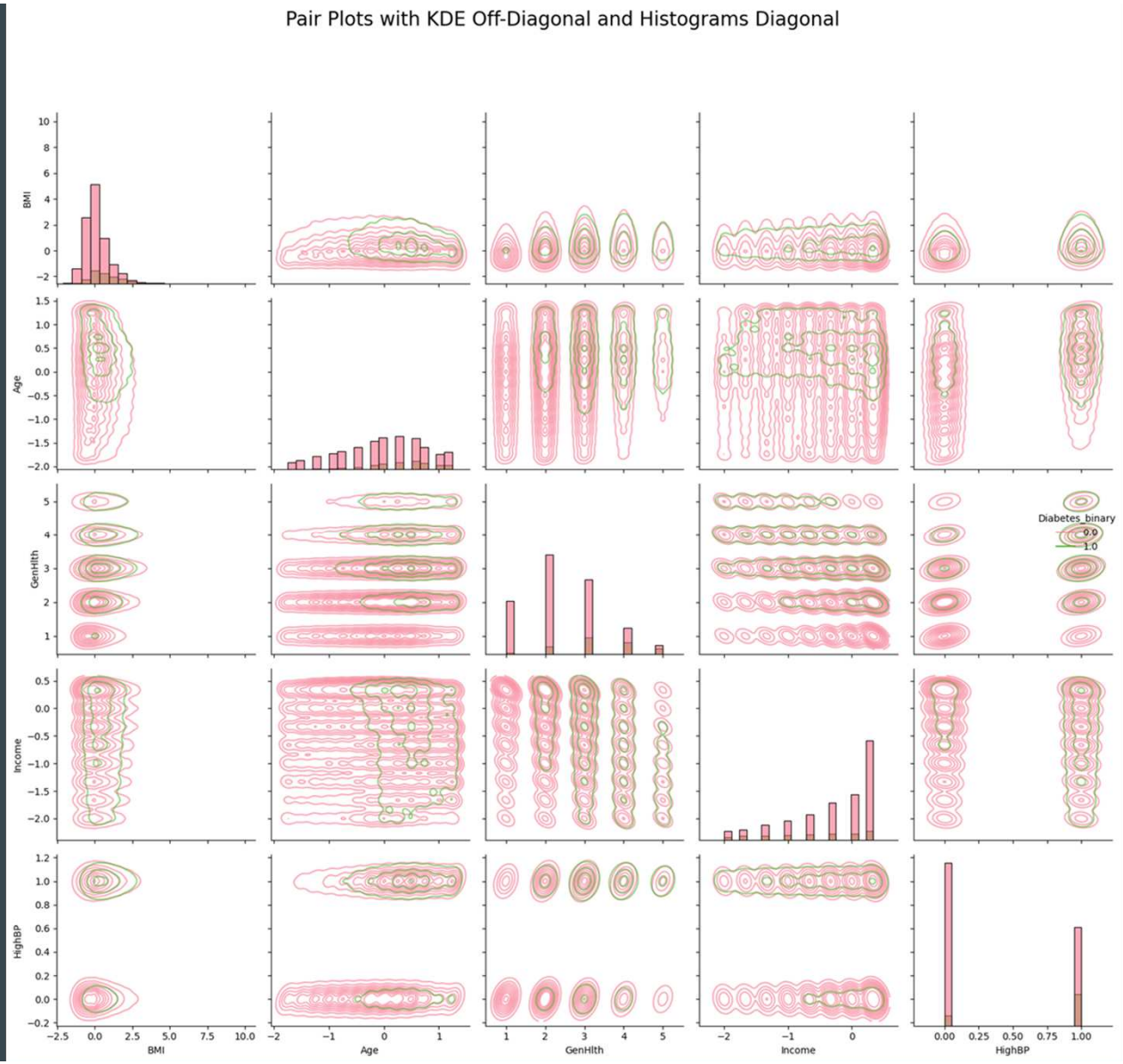


Features Name	Description
High Blood Pressure (BP)	0 = No High BP, 1 = High BP
High Cholesterol (Chol)	0 = No High Cholesterol, 1 = High Cholesterol
Cholesterol Check	0 = No, 1 = Yes
Body Mass Index (BMI)	Numerical value representing weight adjusted for height
Smoker	0 = No (has not smoked 100+ cigarettes in lifetime), 1 = Yes
Stroke	0 = No, 1 = Yes (has had a stroke)
Heart Disease or Attack	0 = No (no coronary heart disease or myocardial infarction), 1 = Yes
Physical Activity	0 = No physical activity in past 30 days (excluding work), 1 = Yes
Fruit Consumption	0 = No, 1 = Consumes fruits 1 or more times per day
Vegetable Consumption	0 = No, 1 = Consumes vegetables 1 or more times per day
Heavy Alcohol Consumption	0 = No, 1 = Meets criteria for heavy drinking
Health Insurance Coverage	0 = No, 1 = Has any health insurance/prepaid plan
Difficulty Affording Healthcare	0 = No, 1 = Faced difficulty affording healthcare in the past year
General Health Perception	Scale 1-5 (1 = Excellent, 5 = Poor)
Mental Health Days	Number of days in the past 30 with poor mental health (scale 1-30)
Physical Health Days	Number of days in the past 30 with poor physical health (scale 1-30)
Difficulty Walking	0 = No, 1 = Has serious difficulty walking/climbing stairs
Sex	0 = Female, 1 = Male
Age Category	13-level category (18-29 to 80+)

Data Preparation

- **Outlier Management (Focus: Data Preservation; Version 1):**
  - Evaluated various methods (Capping, Removal, Imputation).
  - Selected K-Nearest Neighbors (KNN) Imputation: Replace outliers with values based on the average of the nearest neighbors of the outlier data to address outliers in numerical features (like BMI, Age).
  - Contextual Handling: Ordinal features (e.g., GenHlth, Education) were excluded from numerical outlier techniques as IQR/KNN are inappropriate for their scale.
- **Feature Selection and Engineering (Version 2):**
  - Used pair plots with Kernel Density Estimate plots off-diagonal and histograms on diagonal for feature selection and engineering.
- **Feature Density Handling (Version 3):**
  - Used violin plots to further understand feature distributions (modality and denisties) and identify outliers.
  - Indicated need for the use of One Hot Encoding.
- **Feature Scaling (Focus: Consistency & Compatibility):**
  - Experimented with RobustScaler vs. StandardScaler (Version 4).
  - Adopted StandardScaler for feature normalization.
- **Multicollinearity Reduction (Focus: Model Stability):**
  - Initial correlation analysis considered (Version 5).
  - Primarily used Variance Inflation Factor (VIF) analysis for diagnosis.
  - Performed an iterative feature removal process based on high VIF scores (Versions 5, onwards).


Correlation Matrix of Features
Diabetes_binary 1 0.26 0.2 0.065 0.22 0.061 0.11 0.18 -0.12 -0.041 -0.057 -0.057 0.016 0.031 0.29 0.069 0.17 0.22 0.031 0.18 -0.12 -0.16
HighBP 0.26 1 0.3 0.099 0.21 0.097 0.13 0.21 -0.13 -0.041 -0.061 -0.004 0.038 0.017 0.3 0.056 0.16 0.22 0.052 0.34 -0.14 -0.17
HighChol 0.2 0.3 1 0.086 0.11 0.091 0.093 0.18 -0.078 -0.041 -0.04 -0.012 0.042 0.013 0.21 0.062 0.12 0.14 0.031 0.27 -0.071 -0.085
CholCheck 0.065 0.099 0.086 1 0.034 -0.0099 0.024 0.044 0.0042 0.024 0.0061 -0.024 0.12 -0.058 0.047 -0.0084 0.032 0.041 -0.022 0.09 0.0015 0.014
BMI 0.22 0.21 0.11 0.034 1 0.014 0.02 0.053 -0.15 -0.088 -0.062 -0.049 -0.018 0.058 0.24 0.085 0.12 0.2 0.043 -0.037 -0.1 -0.1
Smoker 0.061 0.097 0.091 -0.0099 0.014 1 0.061 0.11 -0.087 -0.078 -0.031 0.1 -0.023 0.049 0.16 0.092 0.12 0.12 0.094 0.12 -0.16 -0.12
Stroke 0.11 0.13 0.093 0.024 0.02 0.061 1 0.2 -0.069 -0.013 -0.041 -0.017 0.0088 0.035 0.18 0.07 0.15 0.18 0.003 0.13 -0.076 -0.13
HeartDiseaseorAttack 0.18 0.21 0.18 0.044 0.053 0.11 0.2 1 -0.087 -0.02 -0.039 -0.029 0.019 0.031 0.26 0.065 0.18 0.21 0.086 0.22 -0.1 -0.14
PhysActivity -0.12 -0.13 -0.078 0.0042 -0.15 -0.087 -0.069 -0.087 1 0.14 0.15 0.012 0.036 -0.062 -0.27 -0.13 -0.22 -0.25 0.032 -0.093 0.2 0.2
Fruits -0.041 -0.041 -0.041 0.024 -0.088 -0.078 -0.013 -0.02 0.14 1 0.25 -0.035 0.032 -0.044 -0.1 -0.068 -0.045 -0.048 -0.091 0.065 0.11 0.08
Veggies -0.057 -0.061 -0.04 0.0061 -0.062 -0.031 -0.041 -0.039 0.15 0.25 1 0.021 0.03 -0.032 -0.12 -0.059 -0.064 -0.081 -0.065 -0.0098 0.15 0.15
HvyAlcoholConsump -0.057 -0.004 -0.012 -0.024 -0.049 0.1 -0.017 -0.029 0.012 -0.035 0.021 1 -0.01 0.0047 -0.037 0.025 -0.026 -0.038 0.0057 -0.035 0.024 0.054
AnyHealthcare 0.016 0.038 0.042 0.12 -0.018 -0.023 0.0088 0.019 0.036 0.032 0.03 -0.01 1 -0.23 -0.041 -0.053 -0.0083 0.0071 -0.019 0.14 0.12 0.16
NoDocbcCost 0.031 0.017 0.013 -0.058 0.058 0.049 0.035 0.031 -0.062 -0.044 -0.032 0.0047 -0.23 1 0.17 0.19 0.15 0.12 -0.045 -0.12 -0.1 -0.2
GenHlth 0.29 0.3 0.21 0.047 0.24 0.16 0.18 0.26 -0.27 -0.1 -0.12 -0.037 -0.041 0.17 1 0.3 0.52 0.46 -0.0061 0.15 -0.28 -0.37
MentHlth 0.069 0.056 0.062 -0.0084 0.085 0.092 0.07 0.065 -0.13 -0.068 -0.059 0.025 -0.053 0.19 0.3 1 0.35 0.23 -0.081 -0.092 -0.1 -0.21
PhysHlth 0.17 0.16 0.12 0.032 0.12 0.12 0.15 0.18 -0.22 -0.045 -0.064 -0.026 -0.0083 0.15 0.52 0.35 1 0.48 -0.043 0.099 -0.16 -0.27
DiffWalk 0.22 0.22 0.14 0.041 0.2 0.12 0.18 0.21 -0.25 -0.048 -0.081 -0.038 0.0071 0.12 0.46 0.23 0.48 1 -0.07 0.2 -0.19 -0.32
Sex 0.031 0.052 0.031 -0.022 0.043 0.094 0.003 0.086 0.032 -0.091 -0.065 0.0057 -0.019 -0.045 -0.0061 -0.081 -0.043 -0.07 1 -0.027 0.019 0.13
Age 0.18 0.34 0.27 0.09 -0.037 0.12 0.13 0.22 -0.093 0.065 -0.0098 -0.035 0.14 -0.12 0.15 -0.092 0.099 0.2 -0.027 1 -0.1 -0.13
Education -0.12 -0.14 -0.071 0.0015 -0.1 -0.16 -0.076 -0.1 0.2 0.11 0.15 0.024 0.12 -0.1 -0.28 -0.1 -0.16 -0.19 0.019 -0.1 1 0.45
Income -0.16 -0.17 -0.085 0.014 -0.1 -0.12 -0.13 -0.14 0.2 0.08 0.15 0.054 0.16 -0.2 -0.37 -0.21 -0.27 -0.32 0.13 -0.13 0.45 1
Diabetes_binary
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
HeartDiseaseorAttack
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Pair Plots with KDE Off-Diagonal and Histograms Diagonal
BMI
Age
GenHlth
Income
HighBP
Diabetes_binary
0.0
1.0

Feature Selection and Feature Engineering Using Pair Plots

Diagonal:

- **BMI:** Shows a right-skewed distribution indicating higher BMI values are more common in diabetic individuals. There's a clear visual separation.
- **Age:** Also shows a right-skewed distribution the separation isn't as pronounced as BMI, but shows distinct age groups.
- **GenHlth:** The non-diabetic group is heavily concentrated in the "Excellent" and "Very Good" categories (lower numerical values), while the diabetic group has a much higher proportion in "Fair" and "Poor" categories (higher numerical values). This feature shows strong discriminatory power.
- **Income:** Similar to GenHlth, Non-diabetic individuals are more prevalent in higher income brackets, and diabetic individuals are more represented in lower income brackets. This feature also seems informative.
- **HighBP:** Non-diabetic group is heavily skewed towards HighBP=0 (no high blood pressure), while the diabetic group has a significant proportion with HighBP=1 (high blood pressure). This is a strong indicator.

Off-Diagonal:

- **BMI vs. Age:** The KDE plot shows that the highest density for non-diabetic individuals is at lower BMI and younger ages (bottom left), indicating that the combination of higher BMI and older age significantly increases diabetes likelihood.
- **BMI vs. GenHlth:** Non-diabetic individuals are concentrated at lower BMI and better general health (lower GenHlth numerical values), confirming the combined influence of these factors.
- **GenHlth vs. HighBP:** Strong separation. Excellent/Very Good health is strongly associated with no HighBP, while Fair/Poor health is strongly associated with HighBP, especially for the diabetic group.

Logistic Regression (Versions 6-9):

The Logistic Regression model served as an interpretable baseline. Initial optimization via GridSearchCV (Versions 6-7) showed limited gains, suggesting inherent limitations of the linear model for this dataset. Feature importances were directly extracted from coefficients (Version 8), and further analysis using Calibration Curves and Odds Ratios (Version 9) provided deeper insights into its probabilistic predictions and feature influence, though its predictive ceiling was apparent (AUC-ROC ~0.824).

To make the GridSearchCV optimization for Logistic Regression as thorough as practically possible within the code, we expanded the param_grid_lr to explore a wider and finer range of hyperparameter values. The focus was on the C parameter (regularization strength) and also included different penalty options for LogisticRegression with the 'liblinear' solver.

Hyperparameter tuning was performed using GridSearchCV and, while thorough, GridSearchCV didn't significantly boost the metrics for Logistic Regression. This suggests that further optimization of the Logistic Regression model itself might be limited, and the issue might lie elsewhere.

Random Forest (Version 10):

The Random Forest model was introduced to capture non-linearities. Optimization using RandomizedSearchCV (balancing thoroughness and efficiency) provided improved performance over the baseline Logistic Regression (AUC-ROC ~0.803), demonstrating the value of tree-based ensembles.

Efficient Random Forest Optimization: The core change is using RandomizedSearchCV instead of GridSearchCV for the Random Forest. Instead of trying every combination of hyperparameters (like GridSearchCV), RandomizedSearchCV samples a specified number of combinations (n_iter) randomly from the provided distributions. This is much faster, especially with large search spaces.

- Balanced n_iter: The n_iter parameter is the primary control over the trade-off between search thoroughness and computation time. 20 iterations is a good starting point; you can adjust it based on your available time and the specific problem.
- Strategic Parameter Grid: Even with randomized search, a well-chosen param_grid_rf is important. The ranges provided cover the most important hyperparameters and typical values that work well.
- Reduced CV Folds: Using cv=5 for the Random Forest further speeds up the process without sacrificing too much accuracy in the performance estimate.

Gradient Boosting (Versions 9 [initial], 11-12):

This is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. This model emerged as a top performer. Initial exploration (Version 9 baseline) showed promise. Systematic optimization using GridSearchCV (Version 11) and further refinement incorporating subsample, max_features, and crucially, early stopping (Version 12, detailed in Notes), proved highly effective. Version 11 achieved the highest AUC-ROC (0.828541) and F1-Score (0.837105), indicating a strong ability to discriminate and balance precision/recall. Version 12 maintained strong performance while potentially improving generalization and efficiency through early stopping. The optimization process took around two hours, with very little improvement.

Base Model

```
Accuracy: 0.8675102491327656
Confusion Matrix:
[[42795   944]
 [ 5778  1219]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     43739
         1.0       0.56      0.17      0.27      6997

    accuracy                           0.87     50736
   macro avg       0.72      0.58      0.60     50736
weighted avg       0.84      0.87      0.84     50736
```

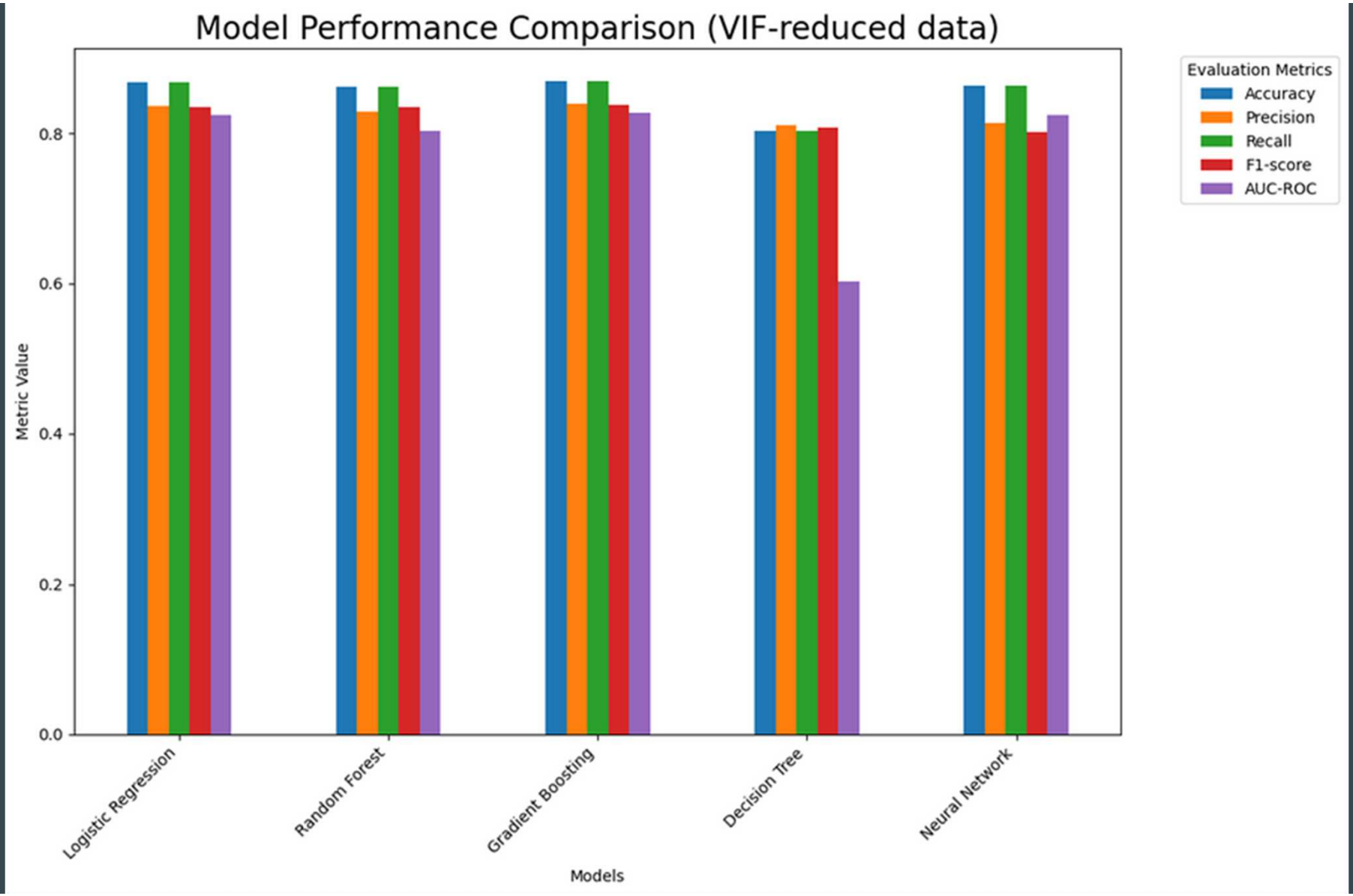
Optimized Model

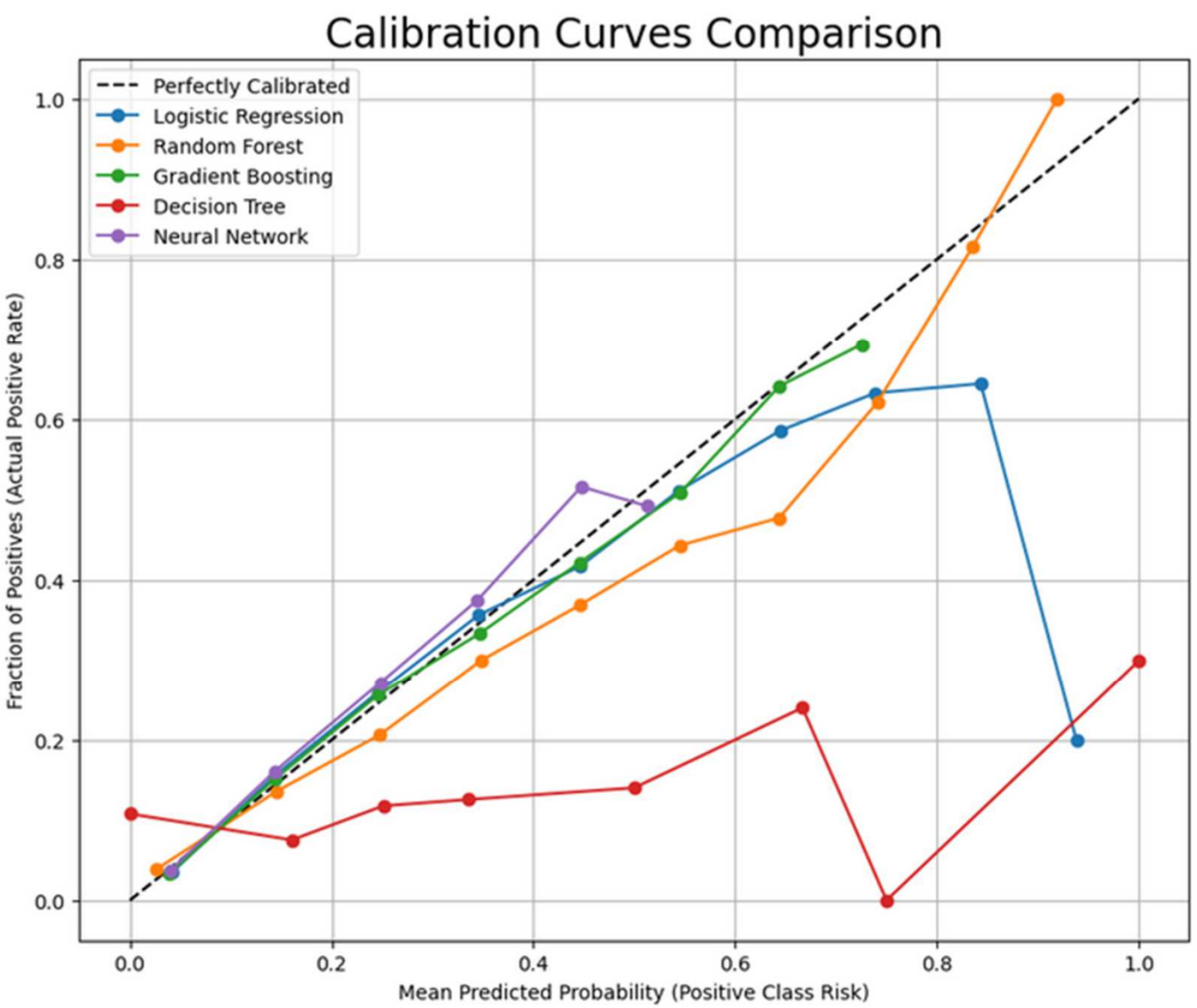
```
Accuracy: 0.8678650268054242
Confusion Matrix:
[[42817   922]
 [ 5782  1215]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     43739
         1.0       0.57      0.17      0.27      6997

    accuracy                           0.87     50736
   macro avg       0.72      0.58      0.60     50736
weighted avg       0.84      0.87      0.84     50736
```

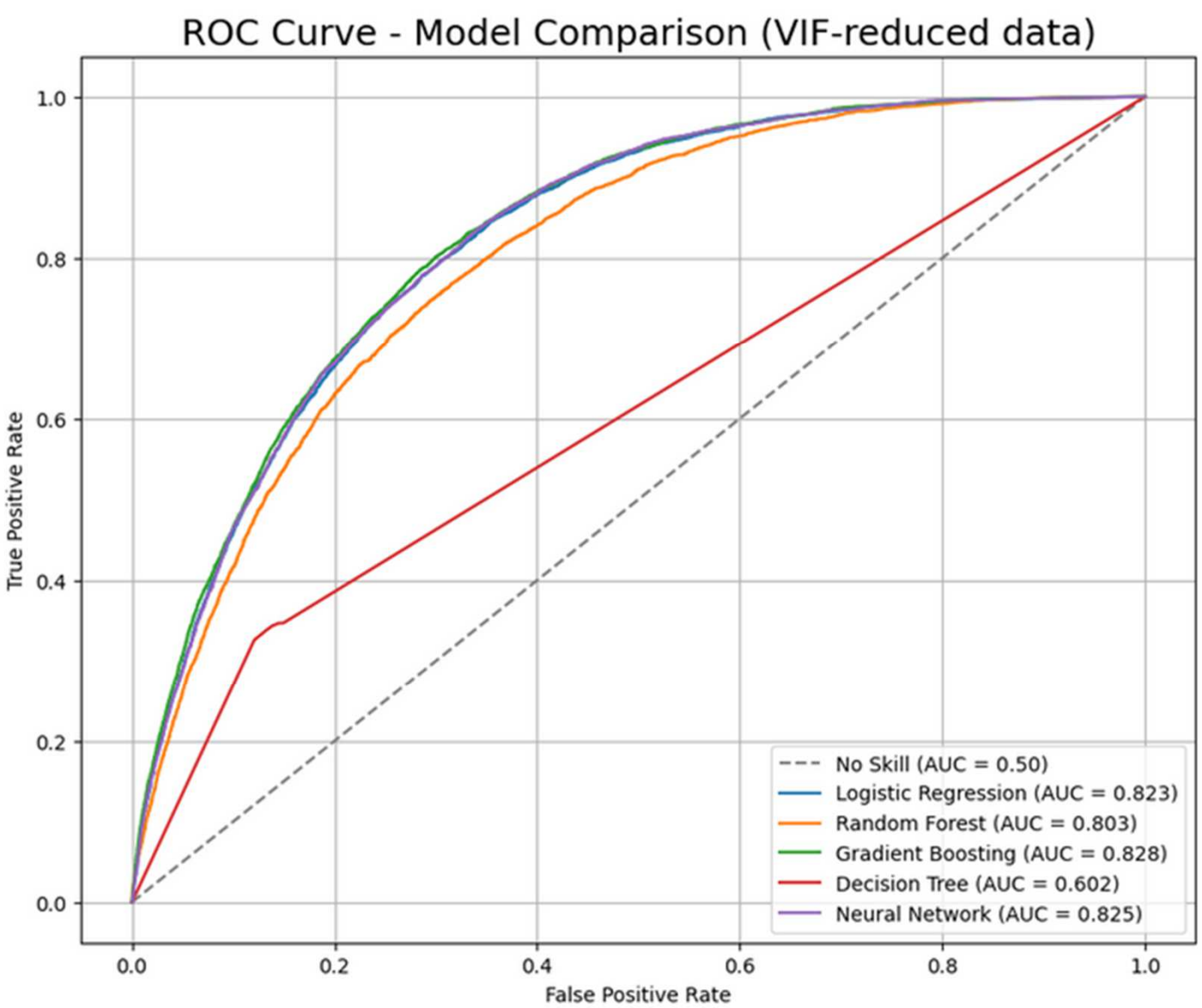
Neural Network (Versions 12 [initial], 13):

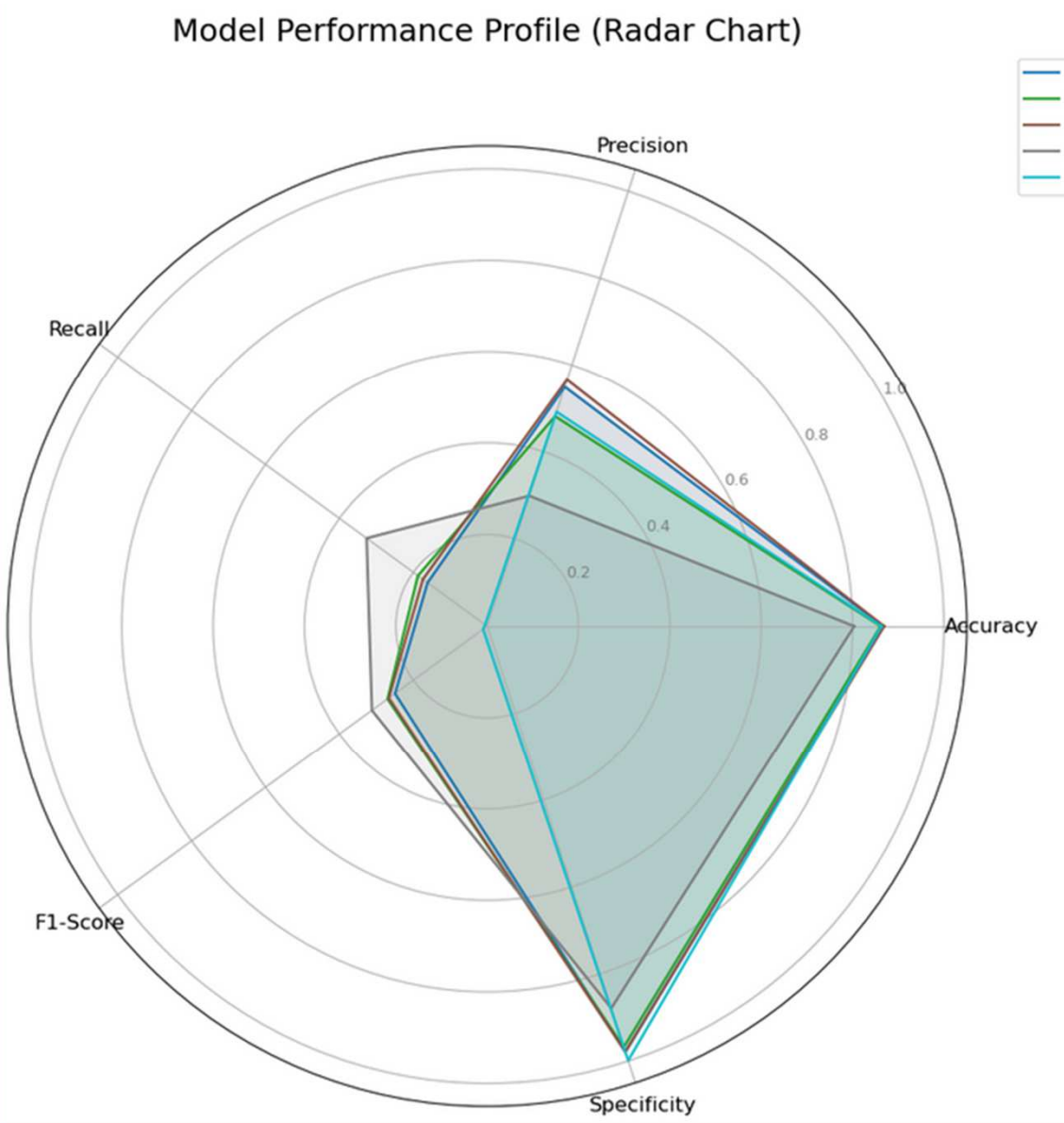
The Neural Network model allowed for the exploration of the potential of deep learning. Optimization using Keras Tuner (RandomSearch) with L2 regularization and early stopping (Version 13, detailed in Notes) was implemented. While achieving respectable results (AUC-ROC ~0.8246), it did not surpass the optimized Gradient Boosting models in this instance, potentially requiring further architectural tuning or larger data scales to unlock its full potential relative to the highly effective boosting methods.

- **keras-tuner Integration:** The core addition is the use of keras-tuner's RandomSearch to optimize the Neural Network's hyperparameters. This is much more efficient than manually trying different combinations.
- **build_nn_model(hp) Function:** This function defines the model architecture, but parameterized by the hp object (hyperparameter space) provided by keras-tuner. This is how keras-tuner explores different configurations.
- **L2 Regularization:** Added kernel_regularizer=l2(0.01) to the Dense layers to help prevent overfitting.
- **Output Layer:** activation='softmax' is used for multi-class classification. The number of units is y_train.shape[1], which corresponds to the number of classes
- **Compilation:** loss='categorical_crossentropy' is used for multi-class classification.
- **RandomSearch**

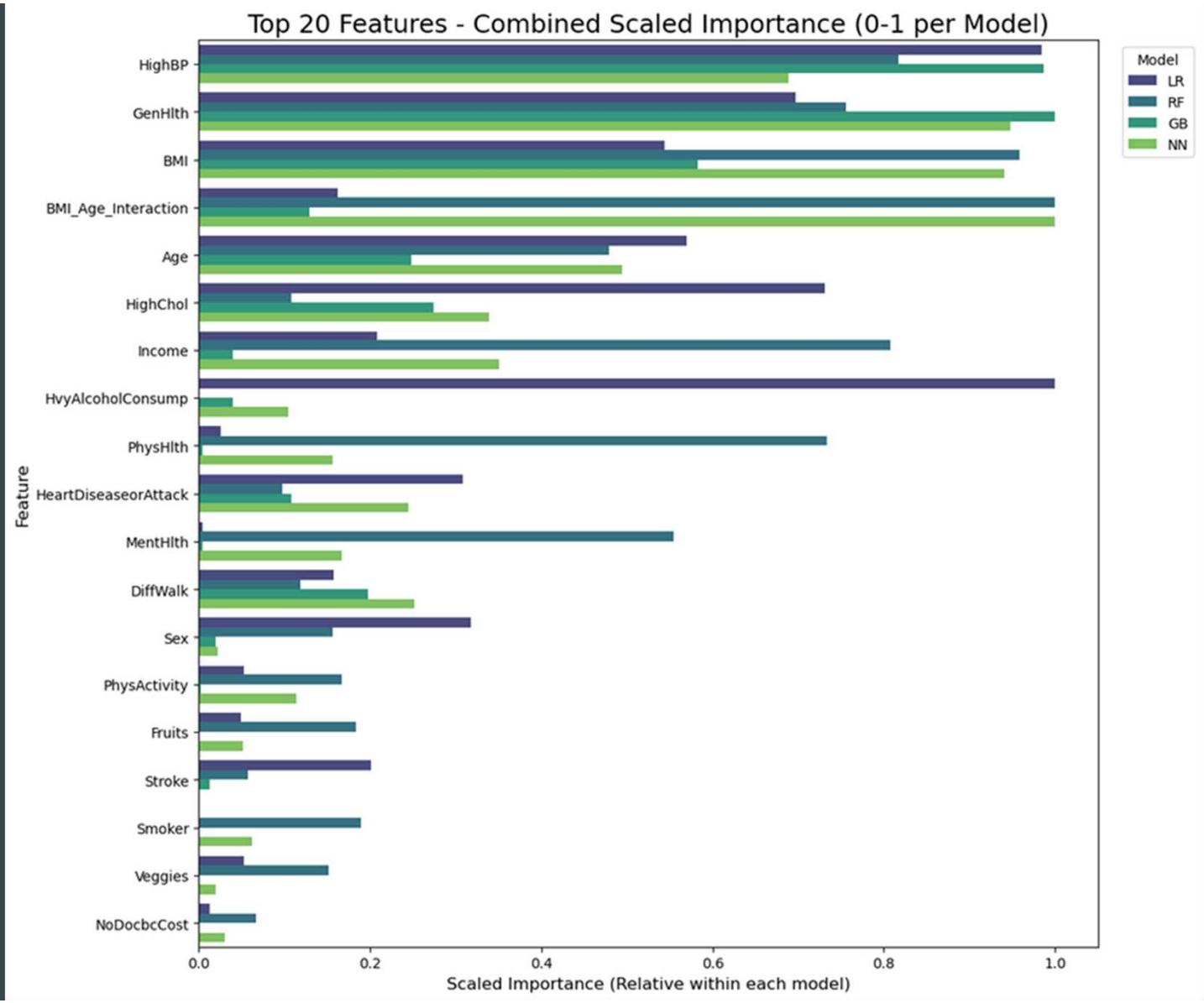

Model Performance Comparison (VIF-reduced data)
Evaluation Metrics
Accuracy
Precision
Recall
F1-score
AUC-ROC
Metric Value
0.0
0.2
0.4
0.6
0.8
Logistic Regression
Random Forest
Gradient Boosting
Decision Tree
Neural Network
Models


Calibration Curves Comparison
Perfectly Calibrated
Logistic Regression
Random Forest
Gradient Boosting
Decision Tree
Neural Network
Fraction of Positives (Actual Positive Rate)
Mean Predicted Probability (Positive Class Risk)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve - Model Comparison (VIF-reduced data)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
No Skill (AUC = 0.50)
Logistic Regression (AUC = 0.823)
Random Forest (AUC = 0.803)
Gradient Boosting (AUC = 0.828)
Decision Tree (AUC = 0.602)
Neural Network (AUC = 0.825)



	Accuracy	Precision	Recall	F1-score	AUC-ROC
Logistic Regression	0.867445	0.835481	0.867445	0.834772	0.823493
Random Forest	0.861505	0.828025	0.861505	0.834254	0.802718
Gradient Boosting	0.869021	0.839101	0.869021	0.837746	0.827785
Decision Tree	0.803532	0.810982	0.803532	0.807134	0.602289
Neural Network	0.863476	0.813205	0.863476	0.801695	0.824584


Top 20 Features - Combined Scaled Importance (0-1 per Model)
HighBP
GenHlth
BMI
BMI_Age_Interaction
Age
HighChol
Income
HvyAlcoholConsump
PhysHlth
HeartDiseaseorAttack
MentHlth
DiffWalk
Sex
PhysActivity
Fruits
Stroke
Smoker
Veggies
NoDocbcCost
Feature
0.0
0.2
0.4
0.6
0.8
1.0
Scaled Importance (Relative within each model)
Model
LR
RF
GB
NN

Conclusions

Top Performer

- Gradient Boosting Was the Top Performer: Across various evaluation metrics (AUC-ROC, F1-Score, Accuracy, Precision, Recall), the optimized Gradient Boosting model consistently demonstrated the strongest predictive performance (AUC = 0.828, F1 = 0.838).

Other Model Performances

- Logistic Regression provided a solid, interpretable baseline (AUC = 0.823) but was slightly outperformed by non-linear models.
- Neural Networks achieved competitive performance (AUC = 0.825), comparable to Logistic Regression, but didn't surpass Gradient Boosting, potentially needing more data or complex architecture.
- Random Forest showed decent performance but lagged behind GB, NN, and LR in AUC (= 0.803).
- A single Decision Tree performed poorly (AUC = 0.602), highlighting the benefit of ensemble methods like RF and GB.

Model Calibration

- Gradient Boosting, Logistic Regression, and Neural Network models demonstrated relatively good calibration, meaning their predicted probabilities align reasonably well with actual observed risk.

Key Risk Factors Identified

- Feature importance analysis consistently highlighted High Blood Pressure (HighBP), General Health perception (GenHlth), BMI, Age, and High Cholesterol (HighChol) as significant predictors across multiple models. Income and potentially engineered features (like BMI_Age_Interaction) also showed relevance.

Questions?