

Using Predictive and Prescriptive Analytics to Predict Diabetes Risk and Improve Preventive Healthcare Resource Allocation

MIS-690 Capstone Project Thesis

Submitted to Grand Canyon University

Graduate Faculty of the Colangelo College of Business

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Business Analytics

by

Sara Seri

Phoenix, Arizona

June 23, 2026

Abstract

This capstone project developed predictive analytics models to identify individuals at elevated risk for diabetes using the CDC BRFSS 2015 Diabetes Health Indicators dataset. The business problem addressed in this study is the difficulty healthcare organizations face in identifying high-risk individuals early enough to support preventive care, screening, and resource allocation. The analysis used demographic, clinical, behavioral, healthcare-access, and self-reported health variables to predict binary diabetes status. Descriptive analysis showed that the full dataset contained 253,680 observations, with diabetes present in 13.93% of respondents. For model development, the balanced 50/50 dataset was used to reduce class imbalance and improve model comparison. Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting models were trained and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Gradient Boosting achieved the strongest ROC-AUC (0.8287), while Random Forest produced similar performance (ROC-AUC = 0.8261) and provided clear feature importance rankings. The most important predictors were general health, high blood pressure, BMI, high cholesterol, age, and difficulty walking. The findings suggest that predictive analytics can help healthcare organizations identify high-risk patients earlier and prioritize preventive interventions more effectively.

Keywords: diabetes prediction, BRFSS, predictive analytics, healthcare analytics, Random Forest, Gradient Boosting, preventive care

List of Tables

Table 1. Dataset Summary

Table 2. Variable Classification

Table 3. Descriptive Statistics

Table 4. Diabetes Distribution

Table 5. Model Performance Comparison

Table 6. Top Predictors of Diabetes

Table 7. Risk Segmentation Results

List of Figures

Figure 1. Diabetes Status Distribution

Figure 2. BMI Distribution by Diabetes Status

Figure 3. Diabetes Prevalence by Age Category

Figure 4. Correlation Heatmap

Figure 5. Random Forest Feature Importance

Figure 6. ROC Curve Comparison

Figure 7. Risk Segmentation by Predicted Probability

Business Problem Identification

Healthcare organizations and public-health agencies face ongoing challenges in identifying individuals at elevated risk for diabetes before serious complications occur. Diabetes is associated with increased healthcare utilization, preventable hospitalizations, reduced quality of life, and long-term costs. Many individuals are not identified as high risk until after symptoms become more serious, which limits opportunities for early intervention. This project uses predictive analytics to support earlier identification of high-risk individuals and improve preventive healthcare resource allocation.

Background

The Behavioral Risk Factor Surveillance System (BRFSS) is a large public-health survey used to collect information about chronic disease, health behaviors, healthcare access, and demographic characteristics. The diabetes health indicators dataset used in this project was derived from BRFSS data and contains health-related predictors that can support diabetes risk modeling. The project aligns with the healthcare need to move from reactive treatment to proactive prevention by using data to identify risk patterns before costly complications occur.

Business Problem Statement

Healthcare organizations need a data-driven method to identify individuals at elevated risk for diabetes using readily available demographic, behavioral, and health-status indicators. Without predictive analytics, preventive programs may be less targeted, limited resources may be inefficiently allocated, and high-risk individuals may be missed until complications develop.

Analytics Problem Statement

The analytics objective was to develop classification models that predict diabetes status using health indicators from the BRFSS diabetes dataset. The target variable was Diabetes_binary, where 1 indicates diabetes and 0 indicates no diabetes. Models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The results were used to identify the strongest predictors and recommend how healthcare organizations could use the model for preventive outreach and resource planning.

Data Understanding, Acquisition, and Preprocessing

The project used the Diabetes Health Indicators dataset from Kaggle, derived from CDC BRFSS 2015 survey data. The full binary dataset contains 253,680 observations and 22 variables, including one target variable and 21 predictors. A balanced 50/50 dataset containing 70,692 records was used for modeling to reduce class imbalance and provide a fairer comparison between diabetic and non-diabetic observations.

Table 1. Dataset Summary

Metric	Value
--------	-------

Full dataset records	253680
Balanced modeling records	70692
Total variables	22
Predictor variables	21
Target variable	Diabetes_binary
Full no diabetes count	218334
Full diabetes count	35346
Full no diabetes percent	86.07%
Full diabetes percent	13.93%
Missing values total	0

Table 2. Variable Classification

Variable	Description	Type
HighBP	High blood pressure indicator	Binary
HighChol	High cholesterol indicator	Binary
CholCheck	Cholesterol check within past five years	Binary
BMI	Body mass index	Numeric/Ordinal
Smoker	Smoked at least 100 cigarettes in lifetime	Binary
Stroke	History of stroke	Binary
HeartDiseaseorAttack	History of coronary heart disease or myocardial infarction	Binary
PhysActivity	Physical activity in past 30 days	Binary
Fruits	Consumes fruit at least once daily	Binary
Veggies	Consumes vegetables at least once daily	Binary
HvyAlcoholConsump	Heavy alcohol consumption indicator	Binary
AnyHealthcare	Has healthcare coverage	Binary
NoDocbcCost	Could not see doctor due to cost	Binary
GenHlth	Self-rated general health scale	Numeric/Ordinal
MentHlth	Poor mental health days in past 30 days	Numeric/Ordinal
PhysHlth	Poor physical health days in past 30 days	Numeric/Ordinal
DiffWalk	Difficulty walking or climbing stairs	Binary
Sex	Sex indicator	Binary
Age	Age category	Numeric/Ordinal

Education	Education level category	Numeric/Ordinal
Income	Income level category	Numeric/Ordinal

Exploratory Data Analysis and Descriptive Statistics

Exploratory analysis was performed to understand the distribution of diabetes status and evaluate relationships between diabetes and the predictor variables. The full dataset showed a class imbalance, with 13.93% of respondents coded as having diabetes. This imbalance supports the use of the balanced modeling file for model training and testing.

Table 3. Descriptive Statistics

Variable	count	mean	std	min	25%	50%	75%	max
BMI	253,680.00	28.38	6.61	12.00	24.00	27.00	31.00	98.00
GenHlth	253,680.00	2.51	1.07	1.00	2.00	2.00	3.00	5.00
MentHlth	253,680.00	3.18	7.41	0.0000	0.0000	0.0000	2.00	30.00
PhysHlth	253,680.00	4.24	8.72	0.0000	0.0000	0.0000	3.00	30.00
Age	253,680.00	8.03	3.05	1.00	6.00	8.00	10.00	13.00
Education	253,680.00	5.05	0.9858	1.00	4.00	5.00	6.00	6.00
Income	253,680.00	6.05	2.07	1.00	5.00	7.00	8.00	8.00

Table 4. Diabetes Distribution

Label	Count	Percentage
No diabetes	218334	86.07
Diabetes	35346	13.93

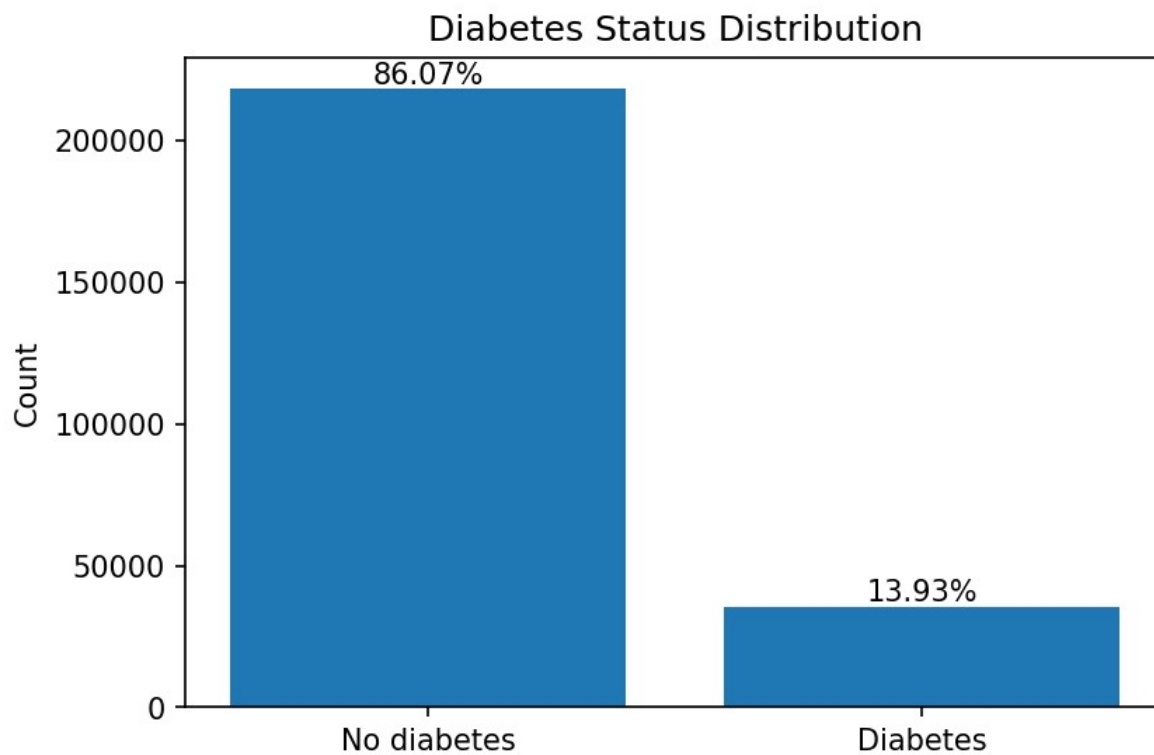


Figure 1. Diabetes Status Distribution

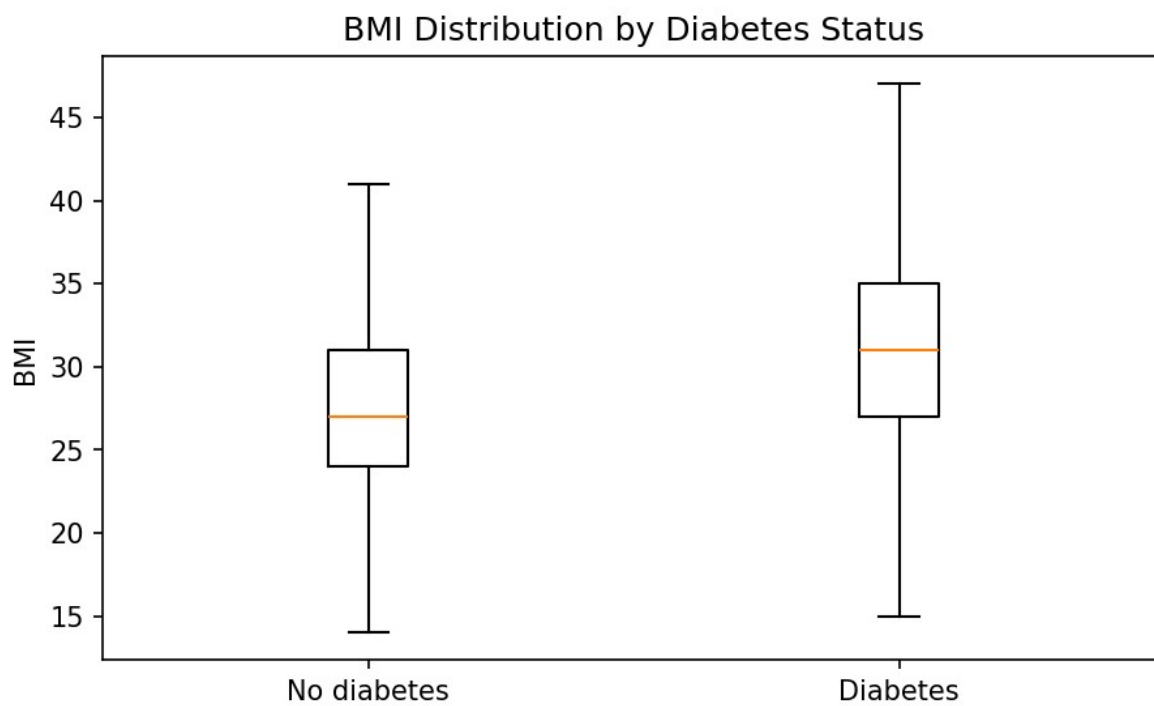


Figure 2. BMI Distribution by Diabetes Status

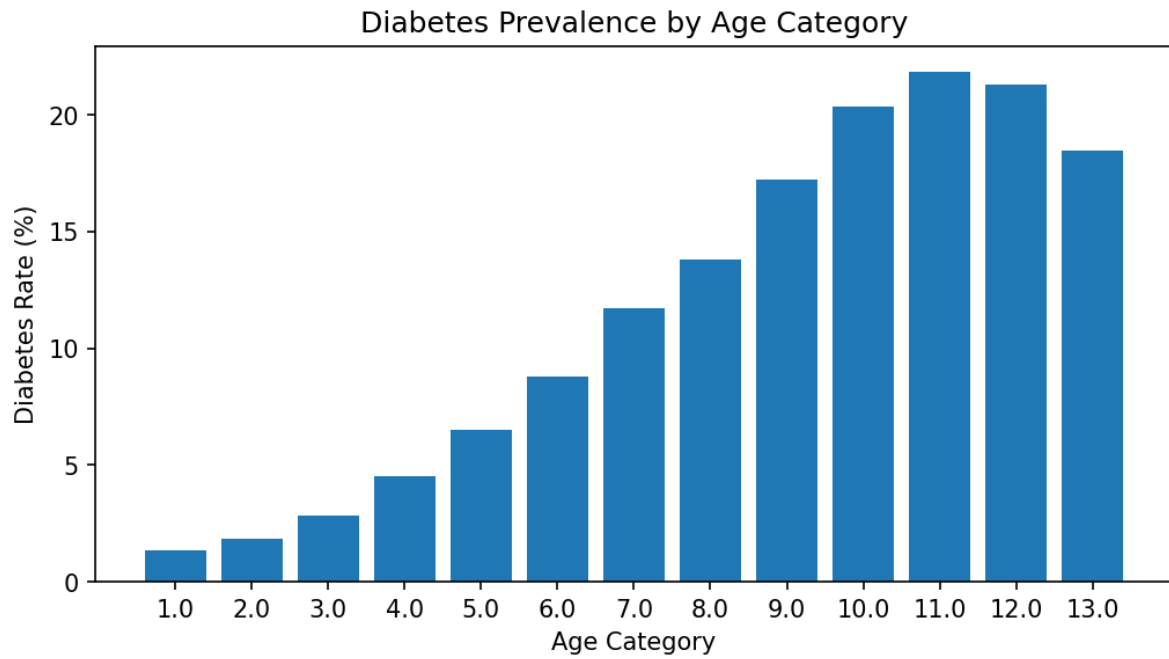


Figure 3. Diabetes Prevalence by Age Category

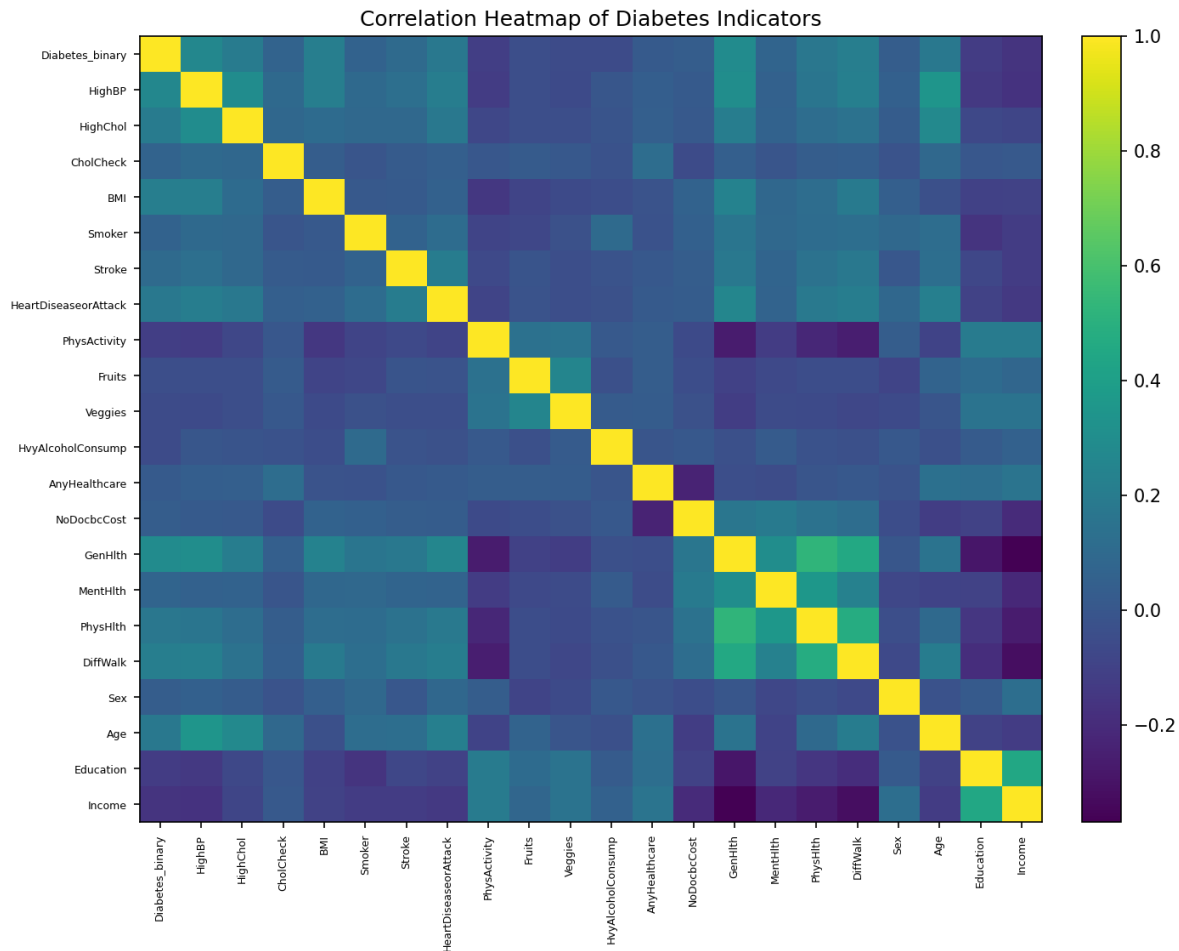


Figure 4. Correlation Heatmap of Diabetes Indicators

The descriptive results showed that diabetes prevalence increased across higher age categories and was higher among individuals with poorer general health, higher BMI, high blood pressure, high cholesterol, and difficulty walking. The correlation analysis identified general health, high blood pressure, BMI, difficulty walking, age, and high cholesterol as some of the strongest variables associated with diabetes status.

Table 5. Top Correlations with Diabetes

Feature	Correlation_with_Diabetes
GenHlth	0.2936
HighBP	0.2631
DiffWalk	0.2183
BMI	0.2168
HighChol	0.2003
Age	0.1774
HeartDiseaseorAttack	0.1773
PhysHlth	0.1713

Income	-0.1639
Education	-0.1245

Methodology Approach and Model Building

Four classification models were developed: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Logistic Regression was used as the interpretable baseline model. Decision Tree was included because it provides a transparent rule-based structure. Random Forest was used to capture nonlinear relationships and interactions among health indicators. Gradient Boosting was included as an advanced ensemble model that improves predictive performance by sequentially correcting errors from prior trees.

Test Design

The balanced dataset was split into training and testing sets using a 70/30 stratified split. Stratification preserved the diabetes and non-diabetes class distribution in both sets. Model performance was evaluated on the unseen testing data using accuracy, precision, recall, specificity, F1-score, and ROC-AUC. Because the project focuses on preventive healthcare, recall and ROC-AUC were emphasized because missing high-risk individuals would be more costly than generating additional screening referrals.

Model Assessment

Table 6. Model Performance Comparison

Model	Accur acy	Precis ion	Recall	Specif icity	F1 Score	ROC- AUC	True Negat ives	False Positi ves	False Negat ives	True Positi ves
Gradi ent Boosti ng	0.751 2	0.730 6	0.795 7	0.706 6	0.761 8	0.828 7	7493	3111	2166	8438
Rando m Forest	0.748 7	0.725 6	0.799 9	0.697 5	0.760 9	0.826 1	7396	3208	2122	8482
Logist ic Regre ssion	0.748 5	0.737 6	0.771 6	0.725 5	0.754 2	0.824 0	7693	2911	2422	8182
Decisi on Tree	0.740 8	0.720 6	0.786 5	0.695 1	0.752 1	0.815 5	7371	3233	2264	8340

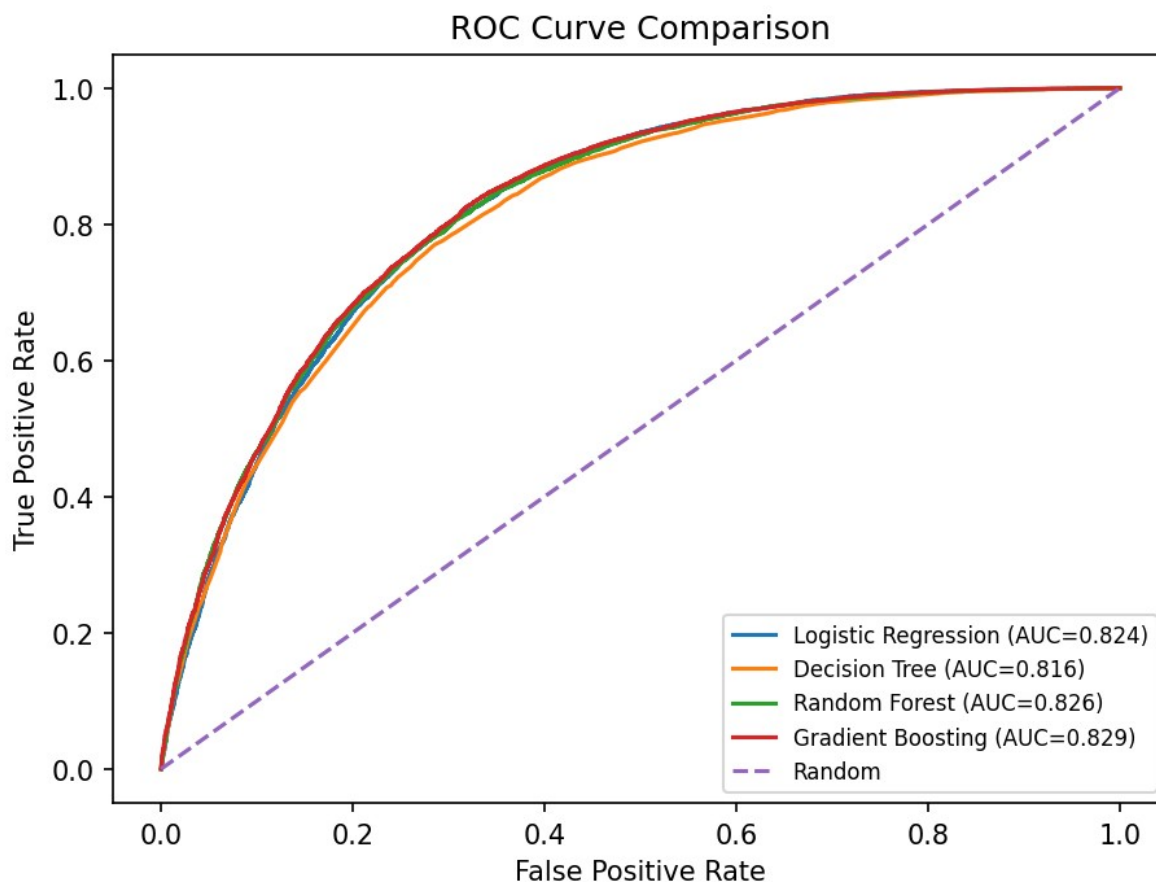


Figure 5. ROC Curve Comparison

Model Evaluation

Gradient Boosting produced the strongest overall ROC-AUC score at 0.8287, with an accuracy of 0.7512, precision of 0.7306, recall of 0.7957, and F1-score of 0.7618. Random Forest performed nearly as well, with a ROC-AUC of 0.8261, accuracy of 0.7487, recall of 0.7999, and F1-score of 0.7609. Logistic Regression also performed strongly with a ROC-AUC of 0.8240 and offered greater interpretability. Decision Tree had the lowest ROC-AUC at 0.8155 but remained useful for explaining high-level decision patterns.

The preferred model for deployment is Gradient Boosting because it achieved the strongest ROC-AUC and balanced recall with overall predictive accuracy. However, Random Forest is also recommended as a practical alternative because it produced similar predictive performance while offering clear feature-importance rankings that are easy to explain to healthcare stakeholders.

Feature Importance and Key Predictors

Table 7. Top Predictors of Diabetes

Feature	Importance
GenHlth	0.2279
HighBP	0.2219
BMI	0.1410
HighChol	0.1034
Age	0.0995
DiffWalk	0.0538
Income	0.0309
HeartDiseaseorAttack	0.0281
PhysHlth	0.0260
Education	0.0127

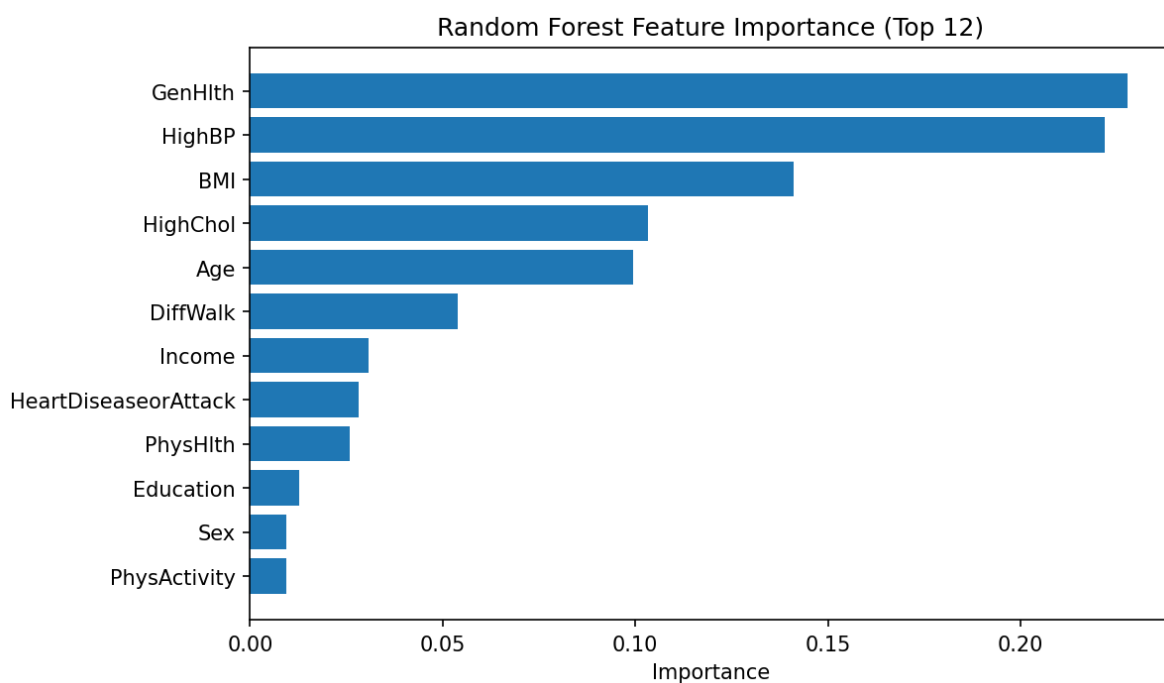


Figure 6. Random Forest Feature Importance

The most important predictors were general health, high blood pressure, BMI, high cholesterol, age, difficulty walking, income, heart disease or attack, physical health days, and education. These results are consistent with the business problem because they identify measurable health and demographic indicators that can be used to prioritize preventive interventions.

Risk Segmentation

Table 8. Risk Segmentation Results

Risk Segment	Records	Observed Diabetes Rate	Observed Diabetes Rate %
Very Low	4242	0.0736	7.36

Low	4241	0.2983	29.83
Moderate	4242	0.5514	55.14
High	4241	0.7253	72.53
Very High	4242	0.8515	85.15

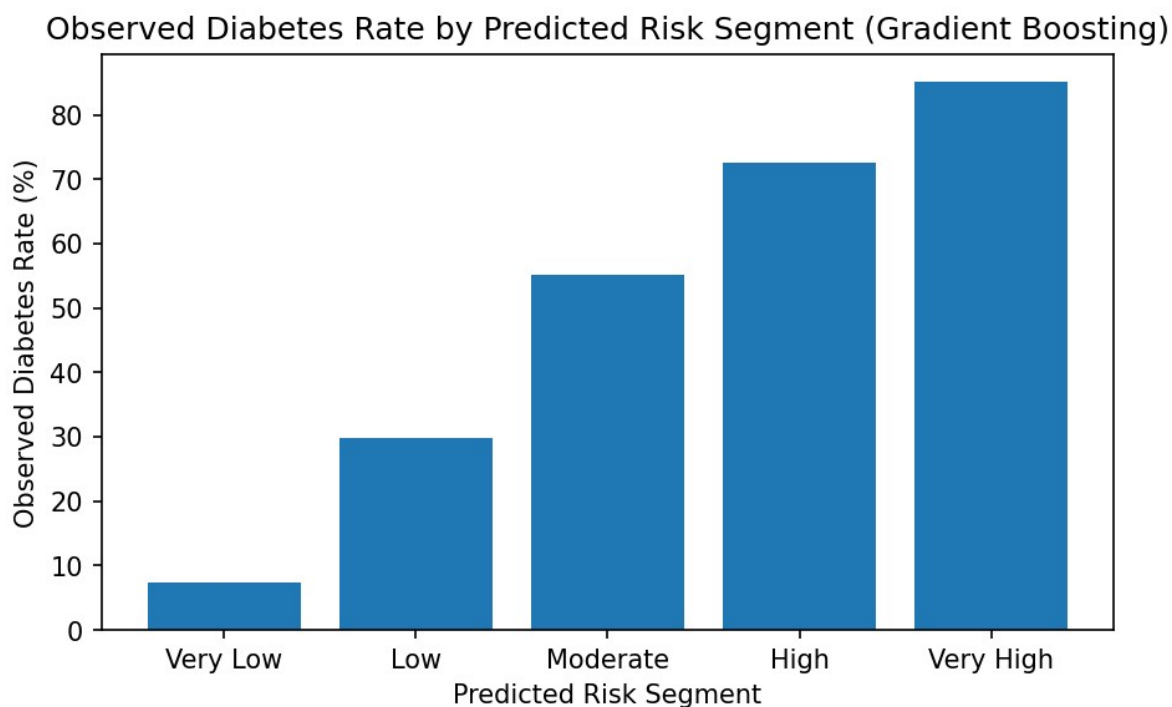


Figure 7. Observed Diabetes Rate by Predicted Risk Segment

The risk segmentation analysis shows a clear separation between low-risk and high-risk groups. The very-low-risk segment had an observed diabetes rate of 7.36%, while the very-high-risk segment had an observed diabetes rate of 85.15%. This indicates that the model can support prescriptive analytics by grouping individuals into actionable risk tiers for outreach, screening, and preventive care planning.

Evaluation of Risk

The primary implementation risk is misclassification. A false negative occurs when a high-risk individual is predicted as low risk, which may delay intervention. A false positive occurs when a lower-risk individual is incorrectly flagged, which may increase screening costs. For preventive healthcare, false negatives are more serious because they may lead to missed opportunities for early intervention. Therefore, recall and ROC-AUC should be monitored closely during model deployment.

Assess Results With Respect to Business Success Criteria

The model meets the business success criteria by identifying individuals who are more likely to have diabetes based on accessible health indicators. The results support preventive outreach by ranking individuals into risk groups. Healthcare organizations could use the high-risk and very-high-risk segments to prioritize diabetes screening, nutrition counseling, physical activity interventions, and follow-up appointments.

External Model Verification and Calibration

External validation should be performed using a more recent BRFSS dataset or another population-level healthcare dataset to determine whether the selected model generalizes beyond the 2015 data. Calibration should also be evaluated to ensure that predicted probabilities align with observed diabetes rates. If the model overestimates or underestimates risk, probability calibration methods such as Platt scaling or isotonic regression may be used.

Future Recommendations

Future work should evaluate the model using newer BRFSS data, compare performance across demographic subgroups, and assess fairness across age, income, sex, and education groups. Additional predictors such as laboratory results, medication history, family history, and clinical claims data could improve accuracy if available. The model should also be tested prospectively before being used in operational healthcare decision-making.

Model Deployment and Model Life Cycle

The recommended deployment strategy is to integrate the model into a healthcare analytics dashboard that scores individuals based on diabetes risk. The dashboard should display risk tier, key contributing factors, and recommended actions. High-risk patients could be referred for diabetes screening, nutrition education, physical activity programs, or follow-up care. The model should be retrained annually as new public-health data become available and monitored for drift in performance metrics.

Milestones, Training, Cost/Benefit, and Risks

Key milestones include final model approval, dashboard development, staff training, pilot testing, performance monitoring, and full deployment. Costs include analyst time, dashboard development, data storage, and staff training. Benefits include earlier risk identification, improved targeting of preventive services, better allocation of resources, and potential reductions in preventable complications. Main risks include model bias, data quality issues, privacy concerns, and overreliance on model outputs without clinical judgment.

Conclusions

This capstone project demonstrated that predictive analytics can support diabetes risk identification using BRFSS health indicators. Gradient Boosting achieved the strongest model performance, while Random Forest provided strong predictive power and clear feature interpretability. The most important predictors were general health, high blood pressure, BMI, high cholesterol, age, and difficulty walking. The risk segmentation results showed that predicted probabilities can be converted into actionable risk tiers, supporting preventive screening and healthcare resource allocation. Overall, the analysis supports the use of predictive and prescriptive analytics to improve early diabetes risk identification and strengthen preventive healthcare planning.

References

- Centers for Disease Control and Prevention. (2015). 2015 BRFSS survey data and documentation. https://www.cdc.gov/brfss/annual_data/annual_2015.html
- Centers for Disease Control and Prevention. (n.d.). Behavioral Risk Factor Surveillance System. <https://www.cdc.gov/brfss/>
- Teboul, A. (n.d.). Diabetes health indicators dataset. Kaggle. <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>
- UCI Machine Learning Repository. (2023). CDC diabetes health indicators. <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. <https://www.the-modeling-agency.com/crisp-dm.pdf>

Appendix A: Files Produced

dataset_summary.csv

descriptive_statistics.csv

diabetes_distribution.csv

model_performance.csv

random_forest_feature_importance.csv

risk_segments.csv

model_predictions.csv

Diabetes_Capstone_Results_Workbook.xlsx