Diabetes Risk Prediction by Patient Health Indicators

A supervised classification analysis using CDC / UCI health indicator data

---

Joshua Ermert

MIS 401: Business Intelligence & Analytics

Dr. Xialu Liu

San Diego State University

7 May 2026

Table of Contents

Executive Summary

This project evaluates whether common health, lifestyle, and demographic indicators can predict three-class diabetes status: NoDiabetes, Prediabetes, or Diabetes, using the CDC Diabetes Health Indicators dataset. The dataset contains 253,680 observations, 21 predictors, and no missing values. The analysis compares three supervised classification methods: multinomial logistic regression, k-nearest neighbors (KNN), and a tuned decision tree. Each model is trained and evaluated on the same 70/30 random split with seed(401), allowing the model comparison to reflect differences in modeling behavior rather than differences in sampled data.

The headline accuracy result is clear but must be interpreted carefully. The tuned Decision Tree achieves the highest raw test accuracy at 0.8475, followed closely by multinomial logistic regression at 0.8467, KNN with best k = 9 at 0.8382, and KNN with k = 5 at 0.8301. On a simple leaderboard, the Decision Tree wins. However, the dataset is severely imbalanced: 84.24% of observations are NoDiabetes, only 1.83% are Prediabetes, and 13.93% are Diabetes. Because the majority-class baseline is already roughly 84.2%, overall accuracy alone can exaggerate performance and hide poor minority-class detection.

The deeper finding is therefore not just identifying which algorithm has the highest accuracy but instead how the models behave under class imbalance. Logistic regression and the tuned decision tree never predict Prediabetes. KNN with k = 5 predicts Prediabetes 88 times but is correct only twice. Diabetes sensitivity is also low overall: about 9.1% for the tuned tree, 17.3% for logistic regression, and 22.0% for KNN k = 5. This means the models are better at identifying the broad majority class than at screening all true Diabetes or Prediabetes cases.

Despite that limitation, the analysis produces a useful risk-factor story. The decision tree variable-importance output is led by HighBP, GenHlth, HighChol, DiffWalk, Age, PhysHlth, and BMI. These predictors align with the exploratory analysis: diabetes risk rises with high blood pressure, worse general health, high cholesterol, higher BMI, older age category, and mobility/physical-health limitations. The tuned tree translates this relationship into an interpretable prioritization structure: a large low-risk branch where HighBP = 0 has only about a 6% Diabetes probability, while a small high-risk leaf defined by HighBP = 1, GenHlth >= 4, BMI >= 28, HighChol = 1, and BMI >= 35 has about a 60% Diabetes probability.

The practical conclusion consists of two parts. If the only criterion is raw accuracy, the tuned Decision Tree is the best model. If the goal is a defensible screening model that catches more true Diabetes cases while remaining interpretable, multinomial logistic regression is a stronger candidate despite slightly lower accuracy. Future work should prioritize class-weighted training, oversampling, cost-sensitive thresholds, and possibly a binary reframing of the task before treating the model as a real clinical screening tool.

Executive summary model comparison.

Model	Test Accuracy	Interpretive Summary
Decision Tree	0.8475	Highest raw accuracy; strong interpretability; low Diabetes sensitivity.
Multinomial Logistic Regression	0.8467	Nearly tied accuracy; more balanced Diabetes detection than the tree.
KNN (best k = 9)	0.8382	Best tuned KNN result; lower accuracy and computationally heavier.
KNN (k = 5)	0.8301	Only model to predict Prediabetes, but with very low precision.

1. Introduction and Research Objective

Diabetes is a major public-health condition because it affects long-term health, healthcare cost, patient quality of life, and clinical resource allocation. Earlier identification of individuals at elevated diabetes risk can support prevention, screening, education, and intervention planning. In a business-intelligence context, the question is not only whether a model can classify patients, but also whether the output can be converted into usable decision support.

This project focuses on a three-class diabetes prediction task using health-indicator data. The response variable, Diabetes_012, distinguishes NoDiabetes, Prediabetes, and Diabetes. The predictors include blood pressure, cholesterol, BMI, smoking, physical activity, general health, mental and physical health days, difficulty walking, sex, age, education, and income. These variables are useful because they are routinely collected in health surveys and represent information that could plausibly inform population-level risk stratification.

The research objective is to compare three classification methods - multinomial logistic regression, KNN, and a decision tree, all on the same dataset and held-out test split. The project asks four related questions:

- Which model achieves the highest overall test accuracy?
- How does severe class imbalance affect the interpretation of that accuracy?
- Which predictors appear most important in the decision tree?
- Do the modeling results reinforce the preliminary EDA findings?

My personal motivation for choosing this topic is that Type II diabetes has affected my father, making the subject personally meaningful. Second, prior experience in biopharmaceutical commercial operations and business intelligence made the project relevant to real analytics

work, where model outputs must be translated into understandable guidance for nontechnical decision-makers.

2. Data Source, Scope, and Variable Design

The analysis uses the CDC Diabetes Health Indicators dataset distributed through the UCI Machine Learning Repository and the cleaned Kaggle version of the BRFSS 2015 data. The working file is diabetes_012_health_indicators_BRFSS2015_PRO.csv. It contains 253,680 observations and 22 variables: one response variable and 21 predictors. The dataset has no missing values, which simplifies preprocessing and allows all observations to be used in the modeling pipeline.

Dataset scope and modeling split.

Dataset Property	Value
Working file	diabetes_012_health_indicators_BRFSS2015_PRO.csv
Observations	253,680
Variables	22 total: 21 predictors + Diabetes_012 response
Missing values	None
Train/test split	70/30 random split with set.seed(401)
Training rows	177,576
Testing rows	76,104

2.1 Response Variable

The response variable is Diabetes_012, a three-level categorical variable. The classes are NoDiabetes, Prediabetes, and Diabetes. This is harder than a binary diabetes/no-diabetes problem because the model must distinguish not only diagnosed diabetes from no diabetes, but also the rare intermediate Prediabetes class.

Response variable coding.

Code	Class Label	Meaning
0	NoDiabetes	Respondent does not report diabetes.
1	Prediabetes	Respondent reports prediabetes or borderline diabetes.
2	Diabetes	Respondent reports diabetes.

2.2 Predictors

The predictors include a combination of binary/categorical indicators and continuous or ordinal measures. Binary and categorical predictors were factor-coded for logistic regression and the decision tree. KNN used numeric encoding with scaling because distance-based methods are sensitive to feature scale.

Predictor groups used in the models.

Type	Variables
Binary / categorical	HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, Sex
Continuous / ordinal	BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income

3. Data Preparation and Modeling Design

The modeling design was intentionally kept consistent across all three classifiers. The same full dataset, same response definition, same 21 predictors, and same train/test split were used throughout. This consistency matters because it makes the model comparison fair: observed differences in accuracy and confusion-matrix behavior come from the algorithms and their preprocessing needs, not from different data partitions.

3.1 Factor Encoding

The response variable was converted from numeric labels into readable factor labels: NoDiabetes, Prediabetes, and Diabetes. Fourteen binary/categorical predictors were also converted to factors. This is appropriate for multinomial logistic regression and decision trees because those methods can treat categorical values as category indicators rather than continuous quantities.

3.2 Train/Test Split

The split used set.seed(401), with 70% of observations assigned to training and 30% assigned to testing. The resulting sample sizes were 177,576 training rows and 76,104 testing rows. The class proportions in the training and testing sets are nearly identical, which means the random split preserved the class structure of the full dataset.

Class proportions in the train/test split.

Split	NoDiabetes	Prediabetes	Diabetes
Training set	0.8421	0.0181	0.1397
Testing set	0.8430	0.0186	0.1384

3.3 KNN Scaling

KNN required an additional preprocessing step. Because KNN classifies observations based on distance, predictors with larger numeric ranges can dominate the distance calculation if the data are left unscaled. To avoid this, the training predictors were z-score scaled, and the same training-set centering/scaling parameters were applied to the test predictors. This avoids test-set leakage while preserving comparability between train and test rows.

3.4 Decision Tree Tuning Rationale

The default rpart decision tree did not split. This was not a coding failure; it was a meaningful modeling signal. The default complexity parameter requires a split to improve the relative error enough to justify increasing tree complexity. Because 84.2% of the dataset is NoDiabetes, predicting the majority class is already highly accurate. Under the default settings, no split cleared the threshold enough to produce a useful tree.

To create an interpretable tree for the report, the rpart control parameters were relaxed: cp = 0.001, minsplit = 2000, minbucket = 1000, and maxdepth = 5. This tuning was intentionally conservative. It allowed a shallow tree to form while preventing tiny, unstable terminal nodes. The tuning should be understood as a class-imbalance adjustment for interpretability, not as an exhaustive hyperparameter optimization process.

4. Preliminary Data Analysis and EDA

Because Step 2 is not being submitted as a separate file, this section preserves the preliminary analysis evidence inside the final report. The purpose of the EDA was to inspect variable types, review the response distribution, compare continuous predictors across Diabetes_012 classes, and examine categorical predictors using frequency-style comparisons before any modeling was performed.

4.1 Variable Types Reviewed Before Modeling

The response variable is categorical, so the modeling plan required classification rather than regression. The predictors were separated into binary/categorical variables and continuous or ordinal variables. This distinction determined both the graphing strategy and the preprocessing pipeline: binary predictors were factor-coded for logistic regression and the decision tree, while KNN used the numeric version of the predictors with z-score scaling.

Step 2 variable-type review and graphing plan.

Variable Group	Variables	EDA Treatment
Response	Diabetes_012	Frequency table and class distribution chart.
Categorical / Binary predictors	HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, Sex	Frequency counts, bar charts, and response-class comparisons.
Continuous / Ordinal predictors	BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income	Summary statistics, histograms, and boxplots by Diabetes_012.

4.2 Response Distribution and Class Imbalance

The most important EDA finding is the response distribution itself. The dataset is dominated by NoDiabetes observations, while Prediabetes represents less than 2% of the data. This imbalance shapes both training and evaluation because a model can achieve high overall accuracy by predicting the majority class often, even if it performs poorly on Prediabetes or Diabetes.

Full dataset class distribution.

Class	Count	Share
NoDiabetes	213,703	84.24%
Prediabetes	4,631	1.83%
Diabetes	35,346	13.93%

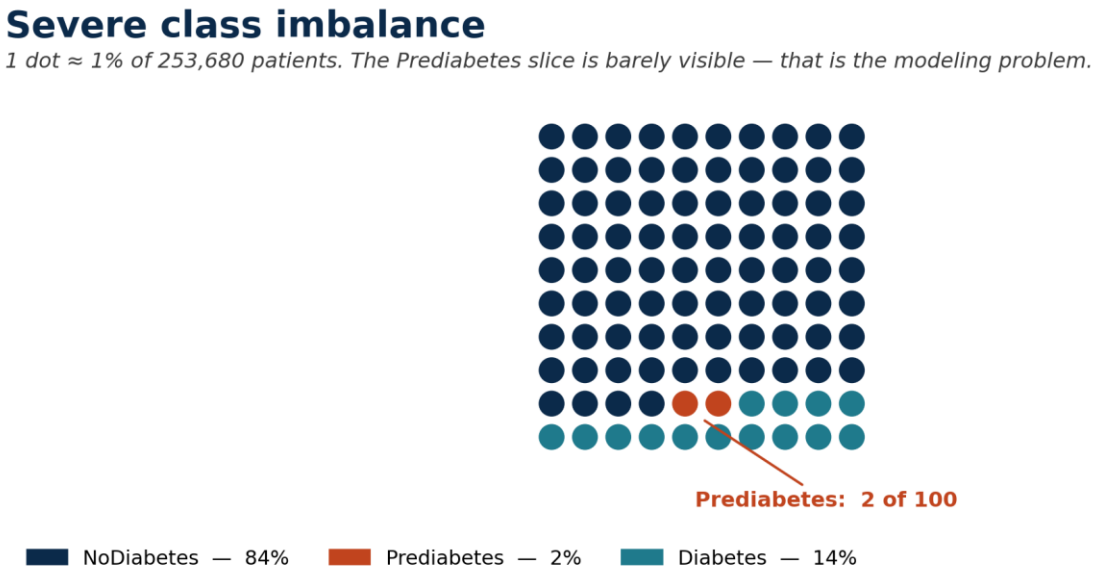

Figure 1. Class imbalance in the Diabetes_012 response. NoDiabetes dominates the dataset, which makes overall accuracy an incomplete evaluation metric.

4.3 Continuous and Ordinal Predictor Distributions

Histograms were used to review the distributions of continuous and ordinal predictors before modeling. BMI is right-skewed, while Age and GenHlth are ordinal measures rather than truly continuous clinical measurements. MentHlth and PhysHlth are also concentrated near zero, which indicates that many respondents reported no poor mental-health or physical-health days in the past month. These distributions matter because KNN relies on distance and therefore required scaling, while logistic regression and the tree were less directly affected by raw numeric ranges.

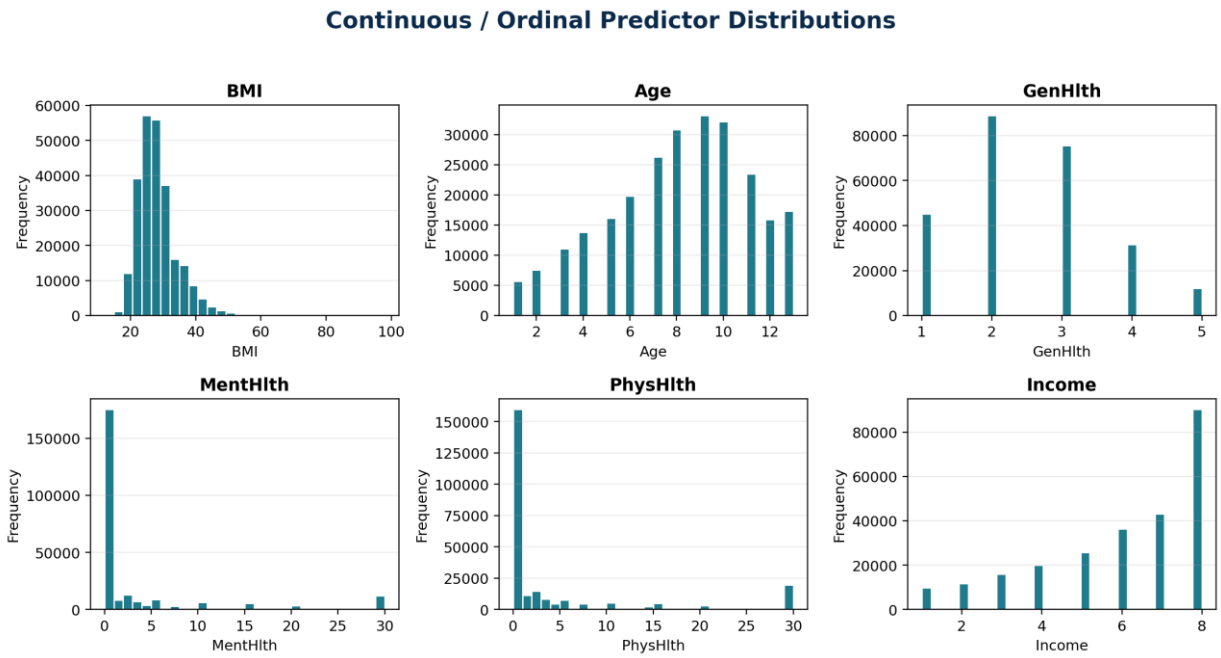

Figure 2. Step 2 histogram review for selected continuous and ordinal predictors. These distributions were reviewed before model fitting to understand scale, skew, and ordinal structure.

4.4 Continuous Predictor Relationships with Diabetes_012

Because the response variable is categorical, boxplots were used to compare continuous and ordinal predictors across NoDiabetes, Prediabetes, and Diabetes groups. BMI increases across the response classes, with the Diabetes group having the highest median BMI. Age category follows the same general pattern, with older respondents more likely to be in the Diabetes group. Self-rated general health worsens sharply in the Diabetes class, which is consistent with diabetes being associated with broader health burden. Physical-health days also tend to be higher among Diabetes observations, though the distribution remains concentrated for many respondents.

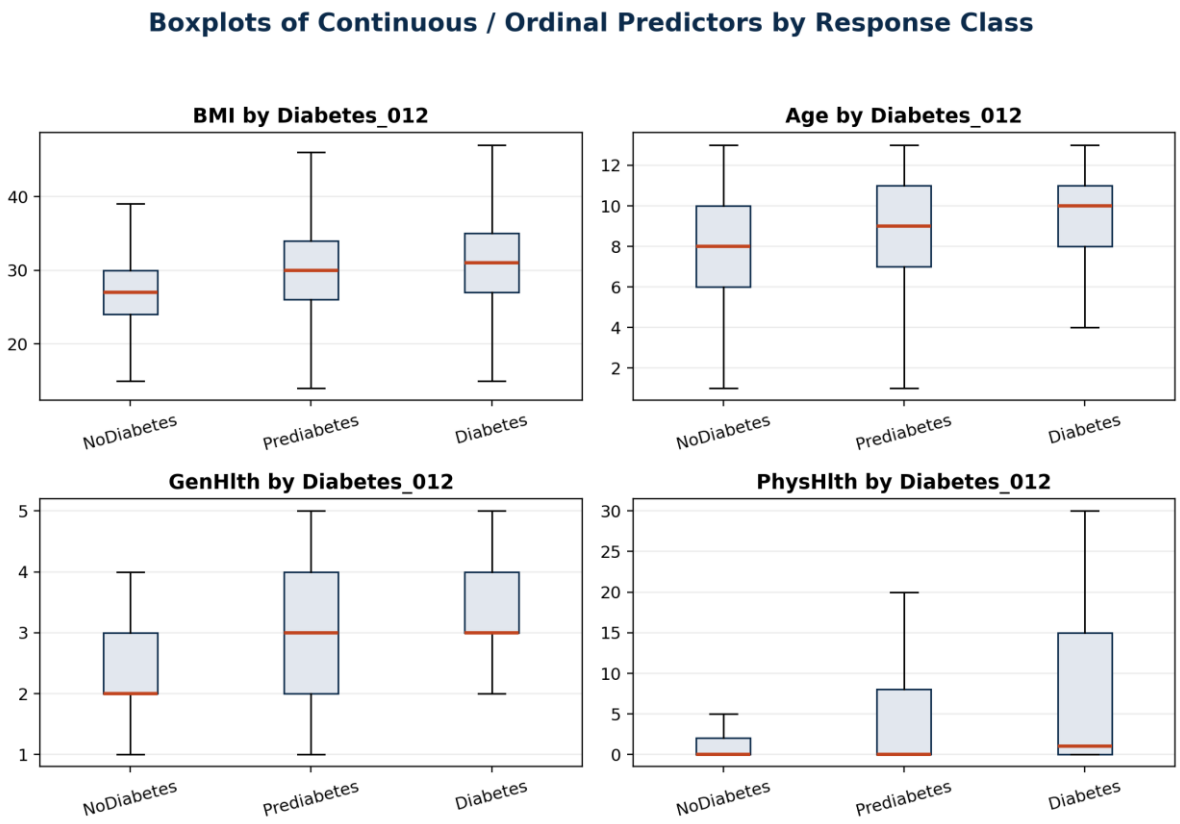

Figure 3. Step 2 boxplot review for selected continuous / ordinal predictors by Diabetes_012 class. BMI, Age, GenHlth, and PhysHlth show visible separation across response groups.

4.5 Categorical Predictor Relationships with Diabetes_012

Categorical predictors were reviewed using frequency and bar-chart comparisons against the Diabetes_012 response. HighBP shows the strongest visual separation: diabetes is much more common among respondents with high blood pressure than among those without it. HighChol and HeartDiseaseorAttack show the same risk direction, while DiffWalk is also associated with higher diabetes concentration. PhysicalActivity moves in the opposite direction: respondents reporting physical activity have a lower observed Diabetes rate than those reporting no physical activity. Sex is present in the data, but its effect is weaker than the biomedical and health-status indicators.

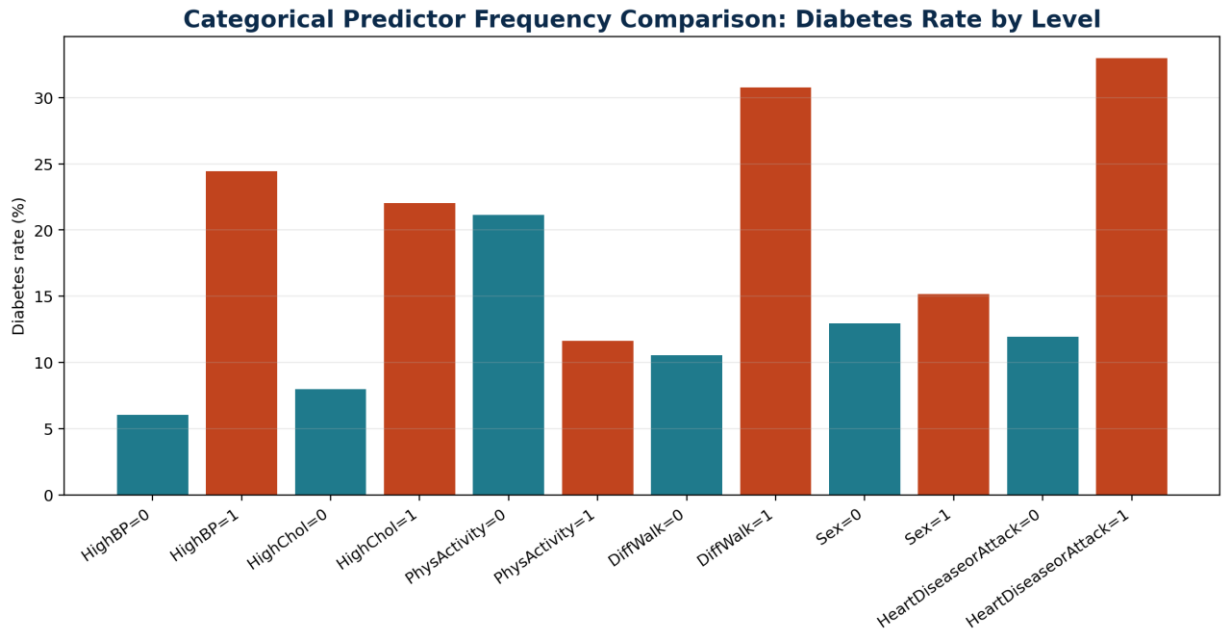

Figure 4. Step 2 categorical predictor comparison. Diabetes rate is visibly higher for HighBP=1, HighChol=1, DiffWalk=1, and HeartDiseaseorAttack=1, and lower among respondents with PhysActivity=1.

4.6 EDA-to-Modeling Bridge

The preliminary analysis created clear expectations before the models were fitted. The strongest early signals were HighBP, GenHlth, HighChol, BMI, and Age. The final decision-tree variable-importance output later confirms that these indicators carry most of the predictive structure. This matters because the model results are not isolated numerical outputs; they reinforce patterns visible during Step 2 exploratory analysis.

5. Model Development

5.1 Multinomial Logistic Regression

Multinomial logistic regression was selected because the response variable has three categories. The method estimates the log-odds of Prediabetes and Diabetes relative to NoDiabetes, using all 21 predictors. It is useful as a baseline because it is interpretable, statistically familiar, and efficient on a dataset of this size.

The fitted model produced a residual deviance of 142,377.2 and AIC of 142,465.2. The coefficient patterns were directionally consistent with EDA: higher HighBP, HighChol, BMI, GenHlth, and Age were associated with higher odds of Diabetes relative to NoDiabetes. While individual coefficients are not the main focus of this report, the logistic model provides an important contrast to the decision tree: it is less visually interpretable than a tree, but it predicts more true Diabetes cases than the tuned tree.

5.2 K-Nearest Neighbors

KNN was selected as an instance-based, nonparametric classifier. Unlike logistic regression, KNN does not estimate coefficients. It classifies each test observation based on the classes of nearby training observations. This makes scaling essential and makes runtime more expensive on a large dataset.

The initial KNN model used k = 5. A short tuning sweep tested k = 3, 5, 7, and 9. Accuracy increased across this range, with k = 9 giving the best KNN result at 0.8382. The tuning range was intentionally simple and course-appropriate rather than exhaustive. Because the dataset is large, KNN also creates a computational burden; this is important when considering practical deployment.

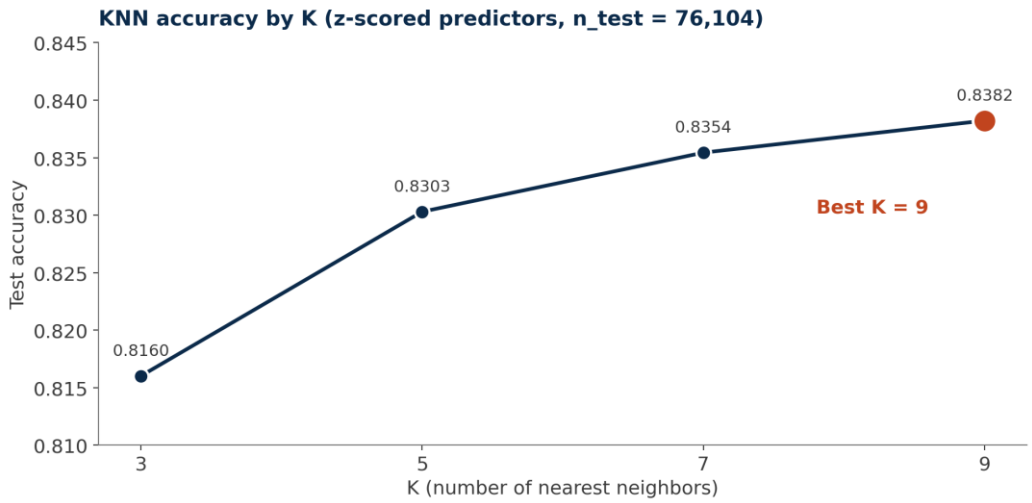

Figure 2. KNN tuning over k = 3, 5, 7, and 9. Accuracy improved as k increased within the tested range.

5.3 Tuned Decision Tree

The decision tree was selected because it can convert predictive patterns into human-readable rules. This is especially valuable in healthcare analytics, where stakeholders often need to understand why a group has been identified as high risk. The tuned tree used conservative complexity settings to form a shallow tree with interpretable branches while avoiding over-fragmentation.

The resulting tree is dominated by splits on HighBP, GenHlth, BMI, and HighChol. The first split is HighBP, which creates a large low-risk branch when HighBP = 0. The highest-risk terminal node requires a combination of high blood pressure, poor self-rated general health, elevated BMI, high cholesterol, and severe obesity. This path is clinically intuitive and easy to communicate.

6. Results and Model Comparison

The model-comparison table shows that the tuned Decision Tree has the highest raw test accuracy, but the margin over logistic regression is extremely small. The difference between the two leading models is only 0.0008, or 0.08 percentage points. This is not enough to claim that the tree is categorically superior in a practical screening context. Instead, the better conclusion is that the tree and logistic regression are essentially tied on accuracy, while differing in interpretability and per-class behavior.

Final model comparison on held-out test set.

Rank	Model	Test Accuracy	Interpretation
1	Decision Tree (cp = 0.001, maxdepth = 5)	0.8475	Highest raw accuracy; most interpretable rule structure.
2	Multinomial Logistic Regression	0.8467	Near-tied accuracy; better Diabetes sensitivity than the tree.
3	KNN (best K = 9)	0.8382	Best KNN result within the tuning sweep.
4	KNN (k = 5)	0.8301	Initial KNN baseline; only model to predict Prediabetes.

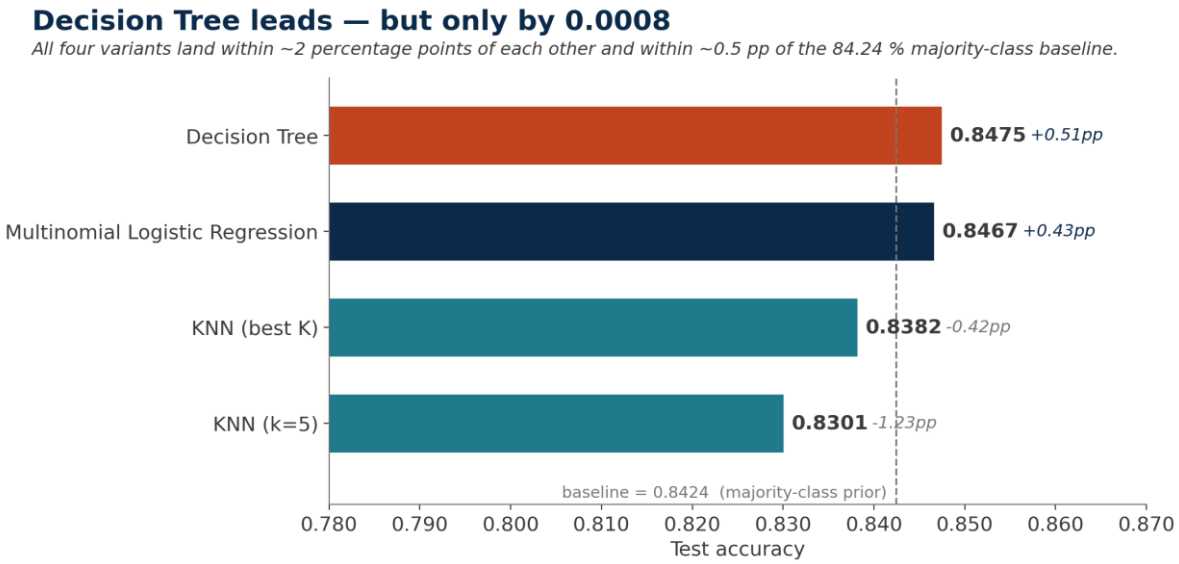

Figure 3. Test accuracy by model. The Decision Tree narrowly leads, but all results are close to the majority-class baseline.

6.1 Confusion Matrix Interpretation

The confusion matrices show why accuracy must be interpreted with caution. All models correctly classify many NoDiabetes observations because that class dominates the dataset. The more challenging question is whether the models detect Diabetes and Prediabetes. Here, performance is much weaker.

Diabetes detection behavior from the test confusion matrices.

Model	Predicted NoDiabetes on true Diabetes	Predicted Diabetes on true Diabetes	Diabetes Sensitivity
Decision Tree	9,572	959	9.1%
Multinomial Logistic Regression	8,713	1,818	17.3%
KNN (k = 5)	8,183	2,312	22.0%

The decision tree achieves high accuracy partly by predicting Diabetes only when the evidence is very concentrated. This makes its Diabetes predictions more selective but causes it to miss many true Diabetes cases. Logistic regression predicts more Diabetes cases and therefore captures more true Diabetes observations. KNN k = 5 has the highest Diabetes sensitivity among the displayed models, but it pays for that with lower overall accuracy and more false Diabetes predictions.

6.2 Prediabetes Detection

Prediabetes is the weakest class for every model. Logistic regression and the decision tree never predict Prediabetes at all. KNN k = 5 predicts Prediabetes 88 times, but only 2 of those predictions are correct. This result suggests that the Prediabetes class is both rare and difficult to separate using the available predictors. It may sit between NoDiabetes and Diabetes in feature space, but the model is not given enough balanced examples to learn it reliably.

6.3 Decision Tree Structure

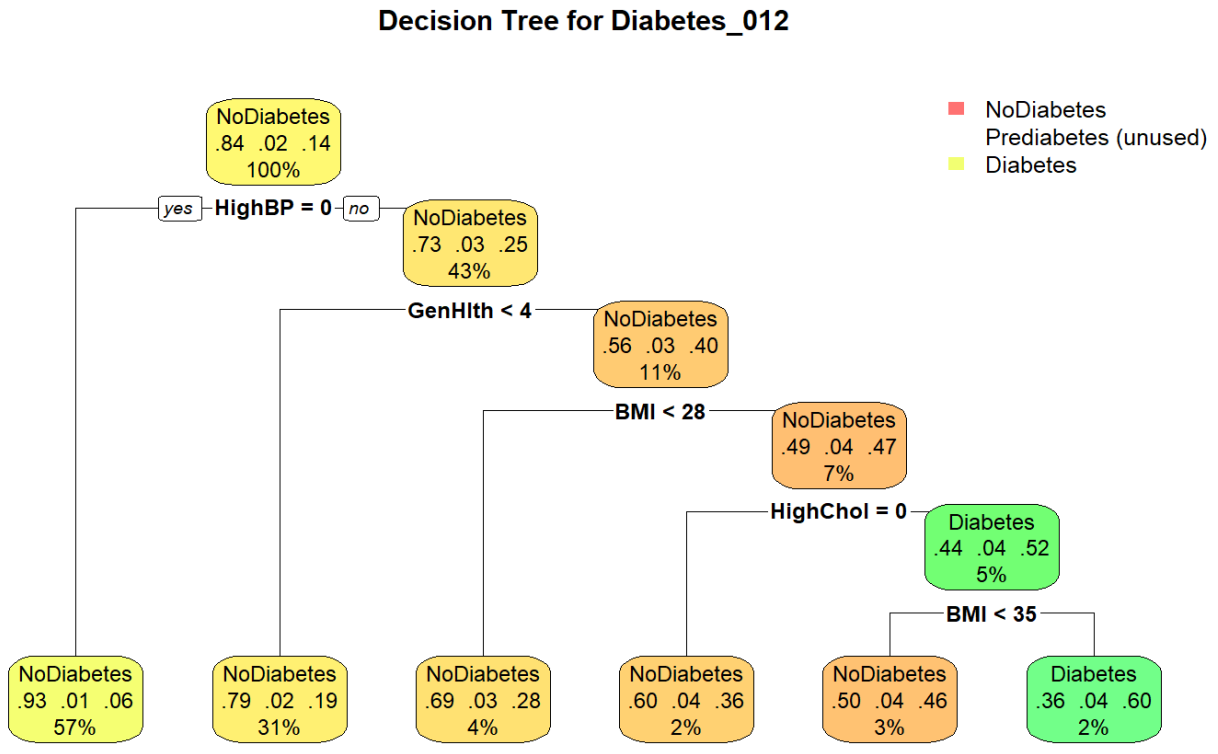

Figure 4. Tuned decision tree for Diabetes_012. The tree separates a large low-risk HighBP = 0 branch from a small high-risk Diabetes-predicting leaf.

The tuned tree has eight terminal nodes and one Diabetes-predicting leaf. The structure is valuable because it makes the risk logic visible. The first split, HighBP = 0, creates the largest low-risk branch. This branch contains 101,437 observations and has only about a 6% Diabetes probability. From a prioritization perspective, this is a rule-out branch: the model can de-prioritize this large group from intensive screening relative to higher-risk groups.

The high-risk leaf has the opposite profile. Respondents with HighBP = 1, GenHlth >= 4, BMI >= 28, HighChol = 1, and BMI >= 35 form a much smaller group of 3,961 observations, but the Diabetes probability is about 60%. This leaf is practically meaningful because it identifies a manageable subgroup where targeted outreach, screening, or prevention resources may be more efficient.

Selected terminal-node structure from the tuned decision tree.

Tree Path	Predicted Class	n	Diabetes Probability
HighBP = 0	NoDiabetes	101,437	0.06
HighBP = 1 and GenHlth < 4	NoDiabetes	55,854	0.19
HighBP = 1 and GenHlth >= 4 and BMI	NoDiabetes	7,056	0.28

< 28			
HighBP = 1 and GenHlth >= 4 and BMI >= 28 and HighChol = 0	NoDiabetes	4,086	0.36
HighBP = 1 and GenHlth >= 4 and BMI >= 28 and HighChol = 1 and BMI < 35	NoDiabetes	5,182	0.46
HighBP = 1 and GenHlth >= 4 and BMI >= 28 and HighChol = 1 and BMI >= 35	Diabetes	3,961	0.60

6.4 Variable Importance

The decision tree variable-importance output reinforces the exploratory analysis. HighBP is the dominant predictor, followed by GenHlth, HighChol, DiffWalk, Age, PhysHlth, and BMI. The highest-ranked variables are not surprising: they describe cardiovascular/metabolic risk, self-rated health burden, mobility limitations, and age-related risk. Their alignment with EDA increases confidence that the model is using meaningful health indicators rather than arbitrary noise.

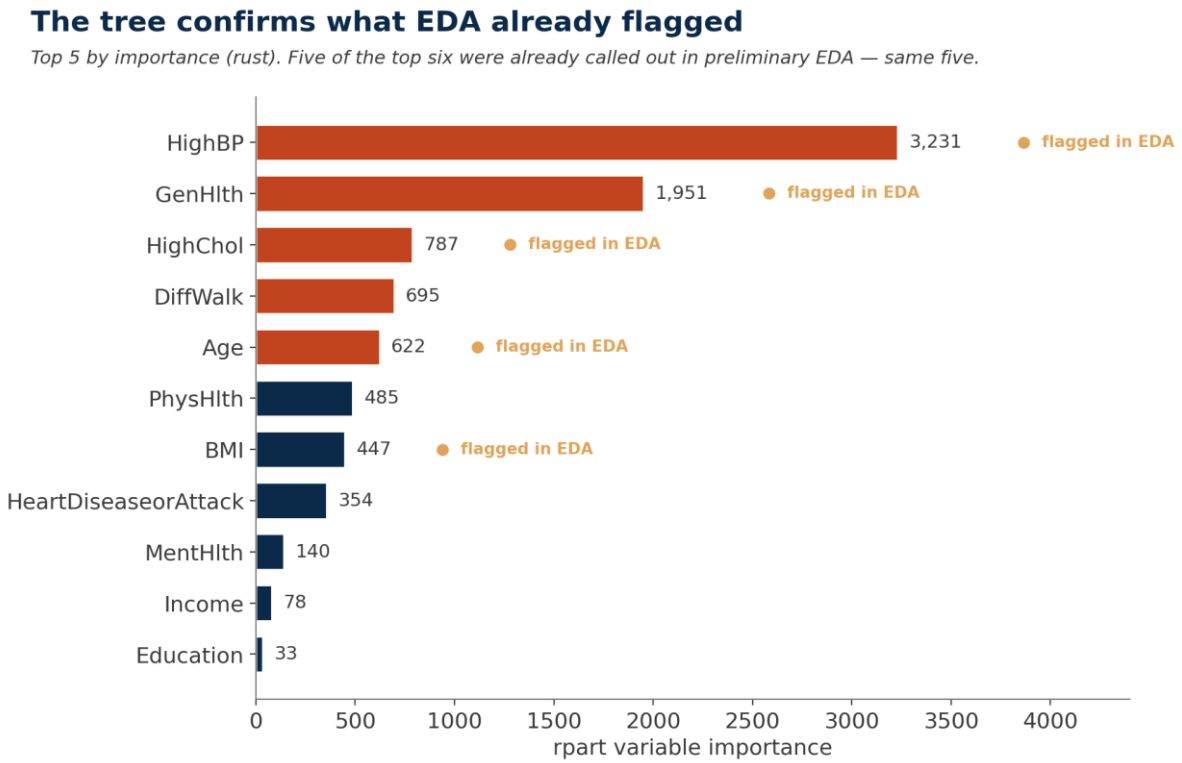

Figure 5. Decision tree variable importance. HighBP, GenHlth, HighChol, DiffWalk, Age, PhysHlth, and BMI carry the strongest tree signal.

Decision tree variable-importance output.

Rank	Variable	Importance
1	HighBP	3,230.6
2	GenHlth	1,950.6
3	HighChol	786.7
4	DiffWalk	694.5
5	Age	621.7
6	PhysHlth	485.5
7	BMI	446.9
8	HeartDiseaseorAttack	354.3
9	MentHlth	139.8
10	Income	77.8
11	Education	32.8

7. Discussion and Interpretation

7.1 Accuracy Is Real, but Incomplete

The Decision Tree accuracy of 0.8475 is a real result from the held-out test set. However, it should not be oversold. The majority-class baseline is approximately 0.8424 because NoDiabetes dominates the dataset. This means the tuned tree improves over a naive majority-class classifier by only about 0.51 percentage points. Logistic regression improves by about 0.43 percentage points, while KNN at k = 9 sits below that baseline.

This does not weaken the project. It instead reveals the main analytical lesson: model evaluation must match the structure of the problem. In imbalanced healthcare classification, overall accuracy is often too blunt. A model can appear strong by correctly classifying the majority class while failing to detect clinically important minority classes. This is exactly what happens here with Prediabetes and, to a lesser extent, Diabetes.

7.2 Why the Decision Tree Wins on Accuracy

The decision tree wins on accuracy because it is a conservative outcome. It predicts NoDiabetes for nearly everyone and reserves Diabetes predictions for a narrow subgroup with a very high concentration of diabetes-related indicators. This behavior protects accuracy because false positives are limited, but it misses many true Diabetes cases. The high-risk leaf is meaningful, but the tree is not broad enough to serve as a complete detection model.

This trade-off is important. If the objective is to create an interpretable prioritization tool, the tree is useful because it identifies high- and low-risk branches clearly. If the objective is to screen as many true Diabetes cases as possible, the tree is less attractive because its Diabetes sensitivity is only about 9.1%.

7.3 Why Logistic Regression May Be More Deployable

Logistic regression is slightly less accurate than the tree but more balanced for Diabetes detection. It predicts more observations as Diabetes and captures 1,818 true Diabetes cases in the test set, compared with 959 for the tree. The cost is more false Diabetes predictions, but in many screening contexts, that may be acceptable. Screening models often prioritize sensitivity because a false positive can be followed by additional testing, while a false negative may delay care.

For this reason, the analysis should not treat the Decision Tree as the only winner. This distinction matters because raw leaderboard performance and usefulness in practice are not the same thing. The tree is the best raw-accuracy and interpretability model, while logistic regression may be the more practical first-choice screening model if detecting more Diabetes cases matters.

7.4 Why KNN Underperforms

KNN underperforms the other methods on overall accuracy and is computationally expensive for this dataset. The method compares test observations against the training set, so 177,576 training rows and 76,104 testing rows create a heavy scoring task. KNN also struggles when rare classes

are sparsely distributed in feature space. The improvement from k = 3 to k = 9 suggests that a slightly smoother neighborhood helps, but the method still does not surpass logistic regression or the tree.

KNN does have one interesting behavior: k = 5 is the only model that predicts Prediabetes at all. That does not make it a strong Prediabetes model because only 2 of 88 Prediabetes predictions were correct, but it suggests that local-similarity methods may identify small pockets of rare-class behavior that more global models ignore. With resampling, class weighting, or binary reframing, this behavior could be worth exploring further.

7.5 Practical Healthcare Analytics Value

The most practical outcome is not a perfect classifier. The value is a transparent risk-stratification story. A healthcare organization could use the tree to identify a large low-risk branch and a small high-risk branch. The HighBP = 0 branch contains many patients with low Diabetes probability. The high-risk leaf contains far fewer patients but a dramatically higher Diabetes concentration. This kind of segmentation supports targeted outreach, triage, and education, even if the model is not sufficient for diagnosis.

The report therefore positions the model as a decision-support tool, not a clinical decision-maker. It can help prioritize where screening resources might be directed, but it should be paired with clinical judgment and improved modeling methods before real-world deployment.

8. Limitations and Future Improvements

8.1 Class Imbalance

The largest limitation is class imbalance. Prediabetes is only 1.83% of the dataset, and Diabetes is 13.93%. Without class weighting, oversampling, undersampling, or threshold adjustment, the algorithms are rewarded for predicting the majority class. This explains why Prediabetes is almost completely ignored and why Diabetes sensitivity remains low.

8.2 Metric Scope

The assignment focuses on confusion matrices and overall accuracy, but a healthcare screening problem should also evaluate sensitivity, specificity, precision, F1-score, balanced accuracy, and possibly area under the ROC curve for binary or one-vs-rest formulations. These metrics would better reflect the cost of missing Diabetes or Prediabetes cases.

8.3 Model Scope

The models are appropriate for the course and are intentionally interpretable. However, the analysis does not include more advanced methods such as random forest, gradient boosting, penalized logistic regression, or calibrated probability thresholds. These methods could improve minority-class performance if paired with appropriate imbalance handling.

8.4 Survey and Causality Limits

The data are self-reported survey responses, not clinical measurements. The analysis cannot prove that any predictor causes diabetes. Variables such as GenHlth and DiffWalk may partly reflect consequences of diabetes rather than purely antecedent risk factors. The appropriate interpretation is predictive association and risk stratification, not causal inference.

8.5 Future Improvement Strategies

- Use class weights or oversampling to improve minority-class learning.
- Reframe the task as binary diabetes screening: Diabetes vs. NoDiabetes/Prediabetes, or AnyDiabetesRisk vs. NoDiabetes.
- Evaluate sensitivity, specificity, precision, F1-score, balanced accuracy, and ROC-AUC where appropriate.
- Test a probability-threshold strategy that prioritizes recall for screening use cases.
- Compare the interpretable baseline models against random forest or gradient boosting while preserving explainability through feature importance or SHAP-style explanations.

9. Conclusion

This project compared multinomial logistic regression, KNN, and a tuned decision tree for predicting diabetes status using the CDC Diabetes Health Indicators dataset. The tuned Decision Tree produced the highest raw test accuracy at 0.8475, followed extremely closely by multinomial logistic regression at 0.8467. KNN with k = 9 reached 0.8382, while the initial KNN k = 5 model reached 0.8301.

The strongest analytical conclusion is that the raw accuracy leaderboard is not enough. Because 84.24% of the observations are NoDiabetes, all models sit close to the majority-class baseline. The confusion matrices show that Prediabetes is essentially not detected and that Diabetes sensitivity remains low. These findings make the project more realistic, not less valuable: they show the importance of choosing evaluation metrics that match the health decision being supported.

The most useful practical output is the risk-factor and segmentation story. The decision tree identifies HighBP, GenHlth, HighChol, DiffWalk, Age, PhysHlth, and BMI as the strongest predictors. It also produces a large low-risk branch and a small high-risk leaf with about 60% Diabetes probability. These interpretable segments could support prioritization for screening and outreach. However, a real deployment would require class-imbalance handling, threshold tuning, and broader validation before being used operationally.

Overall, the project demonstrates how business-intelligence methods can convert health survey data into usable insight: not by producing a perfect classifier, but by revealing the constraints, trade-offs, and risk signals that matter when analytics is applied to healthcare decision support.

10. AI Use Disclosure

Generative AI was used to help organize the report, polish wording, and translate validated R outputs into clearer written explanations. The dataset selection, modeling direction, preprocessing choices, model execution, decision-tree tuning rationale, interpretation of results, and final analytical claims are supported by my R code and completed runtime output. No numerical results were invented. The completed R output and saved CSV artifacts are the source of truth for the analysis.

Appendix A. Reproducibility Details

Reproducibility checklist.

Item	Detail
Final code file	diabetes_modeling_script.R
Data file	diabetes_012_health_indicators_BRFSS2015_PRO.csv
Seed	set.seed(401)
Split	70% training / 30% testing
Training rows	177,576
Testing rows	76,104
R version	4.5.2
Packages	nnet, class, rpart, rpart.plot
Tree control	cp = 0.001, minsplit = 2000, minbucket = 1000, maxdepth = 5
Key outputs	runtime_output_full.md, model_comparison_results.csv, confusion matrices, knn_k_tuning_results.csv, tree_variable_importance.csv, decision_tree_plot.png

Appendix B. Supporting Output Tables

KNN tuning output.

K	Accuracy
3	0.8160
5	0.8303
7	0.8354
9	0.8382

Decision Tree confusion matrix.

Predicted / Actual	NoDiabetes	Prediabetes	Diabetes
NoDiabetes	63,541	1,358	9,572
Prediabetes	0	0	0
Diabetes	618	56	959

Multinomial logistic regression confusion matrix.

Predicted / Actual	NoDiabetes	Prediabetes	Diabetes
NoDiabetes	62,617	1,291	8,713
Prediabetes	0	0	0
Diabetes	1,542	123	1,818

KNN k = 5 confusion matrix.

Predicted / Actual	NoDiabetes	Prediabetes	Diabetes
NoDiabetes	60,860	1,201	8,183
Prediabetes	50	2	36
Diabetes	3,249	211	2,312

Works Cited

"CDC Diabetes Health Indicators." UCI Machine Learning Repository, University of California Irvine, https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators.

Teboul, Alex. "Diabetes Health Indicators Dataset." Kaggle, https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.