Diabetes Prediction

Alexander Wen, Raymond Wong, Michael Eirikson

Table of contents

Summary

In this project we attempt to build a model to predict diabetes disease. We compared a decision tree model and naive bayes model and found the decision tree is stronger in this context. We used f2-score as our scoring function because detecting diabetes is the priority: a false negative could be much worse then a false positive.

In the test dataset: the decision tree model correctly detected 8283 of 10604 positive cases (recall rate is about 78%). This result does come at a fairly significant cost in terms of false positives (precision rate is about 30%) with 19650 false positives. Depending on the actual cost of false positive this may need significant improvement to be a viable screening model.

Introduction

In Canada and the USA approximately 10% of people are living with diabetes. In Canada in 2023 approximately 3.7 million people were living with diabetes and in the USA in 2021 approzimately 38.4 million people were living with diabetes. ("Snapshot of Diabetes in Canada, 2023" (2023)) In the USA it is the 8th leading cause of death. (Rios et al. (2017)) Globally an estimated 44% of people living with diabetes are undiagnosed. (Stafford et al. (2025))

In this project we try to predicted diabetes disease based on common health factors. A reliable model could help to prescreen people and recommend following up with a physician for people who are at risk. Given the large number of people living with undiagnosed diabetes this could potentially have a significant positive impact of world health.

The analysis uses the American CDC Behavioural Risk Factor Surveillance System (BRFSS) 2015 Diabetes Health Indicators dataset (UCI ID 891), containing 253,680 survey responses with 21 health-related features and a binary diabetes outcome (0 = no diabetes/pre-diabetes, 1 = diabetes). (Dane and Teboul (2021))

No missing values were present and all features were already encoded numerically. The target classes is imbalanced (86% non-diabetic, 14% diabetic).

Methods

This analysis was performed in Python 3.11.6 (*Python 3.11.6 Documentation* 2021-2025). Additionally, here is a list of the Python packages used within the analysis with brief explanation:

Table 1: Table of Python packages used

Package	Version	Use case	Reference
`numpy`	`1.26.4`	General analysis use	*NumPy Documentation* (2008-2022)
`pandas`	`2.1.2`	Data management/processing	team (2020), McKinney (2010)
`pandera`	`0.27.0`	Data valiadion	Bantilan (2020)
`altair`	`5.1.2`	Generating plots	VanderPlas et al. (2018), Satyanarayan et al. (2017)
`scikit-learn`	`1.3.2`	Model creation and evaluation	Pedregosa et al. (2011)
`ucimlrepo`	`0.0.7`	Data extraction	Kelly, Longjohn, and Nottingham (2021)

Package	Version	Use case	Reference
`deepchecks`	`0.18.1`	Data validation	Chorev et al. (2022)
`click`	`8.3.1`	Script tool	Pallets (2020)
`quarto`	`1.8.26`	Report creation	Allaire et al. (2025)
`tabulate`	`0.9.0`	Table formatting	Astanin (2025)

Modeling Approach

The data were split 70/30 into training and test sets with stratification on the target. Two classifiers were trained and tuned using 5-fold cross-validated grid search with **f2-score** as the scoring metric. We chose to use f2-score because it is most important to not miss true positives.

1. **Decision Tree** (class_weight='balanced')
   Hyperparameters: max_depth: {6,8,10,12,14}, min_samples_leaf: {175, 200, 225, 250}
   Best parameters: max_depth=10, min_samples_leaf=225
   Best CV f2-score = 0.587
2. **Bernoulli Naive Bayes** (with StandardScaler preprocessing)
   Hyperparameters: alpha: {1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4}
   Best parameters: alpha=0.001
   Best CV f2-score = 0.46

Modeling Results

Table 2: Scores for models.

Unnamed: 0	Model	Test Accuracy	Test f2-score	Test recall	Test precision
0	Decision Tree	0.706	0.587	0.784	0.293
1	Naive Bayes	0.814	0.46	0.489	0.373

EDA

Data Summary

First, here is a sample of the training data showing the first few entries and last few entries in the dataset in Table 3 and Table 4.

Table 3: First few rows of the training data.

	Unnamed: 0	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income	Diabetes
0	0	0	1	1	23	0	0	0	1	1	1	0	1	0	1	0	0	0	0	10	6	8	0
1	1	0	0	1	25	0	0	0	1	1	1	0	1	0	3	0	30	0	0	12	6	7	0
2	2	1	1	1	28	0	0	0	1	0	1	0	1	0	2	15	2	0	1	6	5	6	0
3	3	0	0	1	25	0	0	0	0	1	1	0	1	0	2	0	0	0	0	8	5	7	0
4	4	1	0	1	30	1	0	0	1	1	1	0	1	1	4	30	15	0	0	8	4	4	0

Table 4: Last few rows of the training data.

	Unnamed: 0	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income	Diabetes
0	177571	1	1	1	29	0	0	0	1	1	1	0	1	0	2	0	2	0	0	4	6	5	0
1	177572	1	0	1	22	0	0	0	0	1	1	0	1	0	3	0	0	1	0	13	4	1	0
2	177573	0	1	1	25	1	0	0	1	1	1	0	1	0	2	0	0	0	1	6	5	7	0
3	177574	0	1	1	24	1	0	0	1	1	1	0	1	0	3	2	1	0	0	4	4	8	0
4	177575	1	1	1	31	1	0	0	0	1	1	0	1	0	2	0	2	0	0	8	4	4	0

All features of the dataset are numeric, and further EDA shows there are no null values in the dataset.

Table 5: Description of the training data.

	Unnamed: 0	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income	Diabetes
0	count	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576	177576
1	mean	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.139332
2	std	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.346294
3	min	0	0	0	12	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0
4	25%	0	0	1	24	0	0	0	1	0	1	0	1	0	2	0	0	0	0	6	4	5	0
5	50%	0	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0	0	0	8	5	7	0
6	75%	1	1	1	31	1	0	0	1	1	1	0	1	0	3	2	3	0	1	10	6	8	0
7	max	1	1	1	98	1	1	1	1	1	1	1	1	1	5	30	30	1	1	13	6	8	1

Table 5 displays a numerical distribution of the dataset. Most features are binary in nature, with the exceptions of To better visualize we then plotted frequency counts of the target labels:

Visualizations

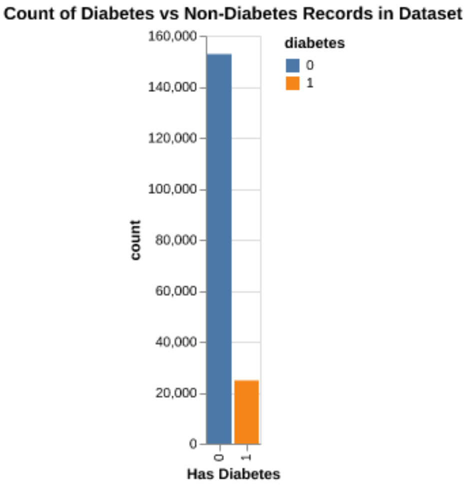

Figure 1: Frequency bar graph of the labels.

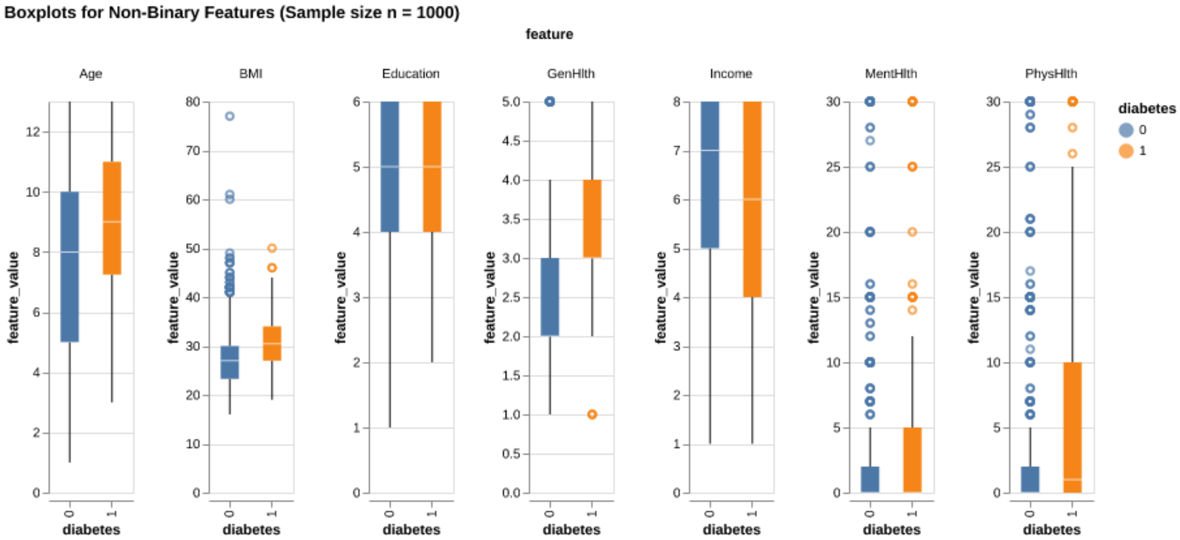

Figure 2: Boxplots of the non-binary numeric features.

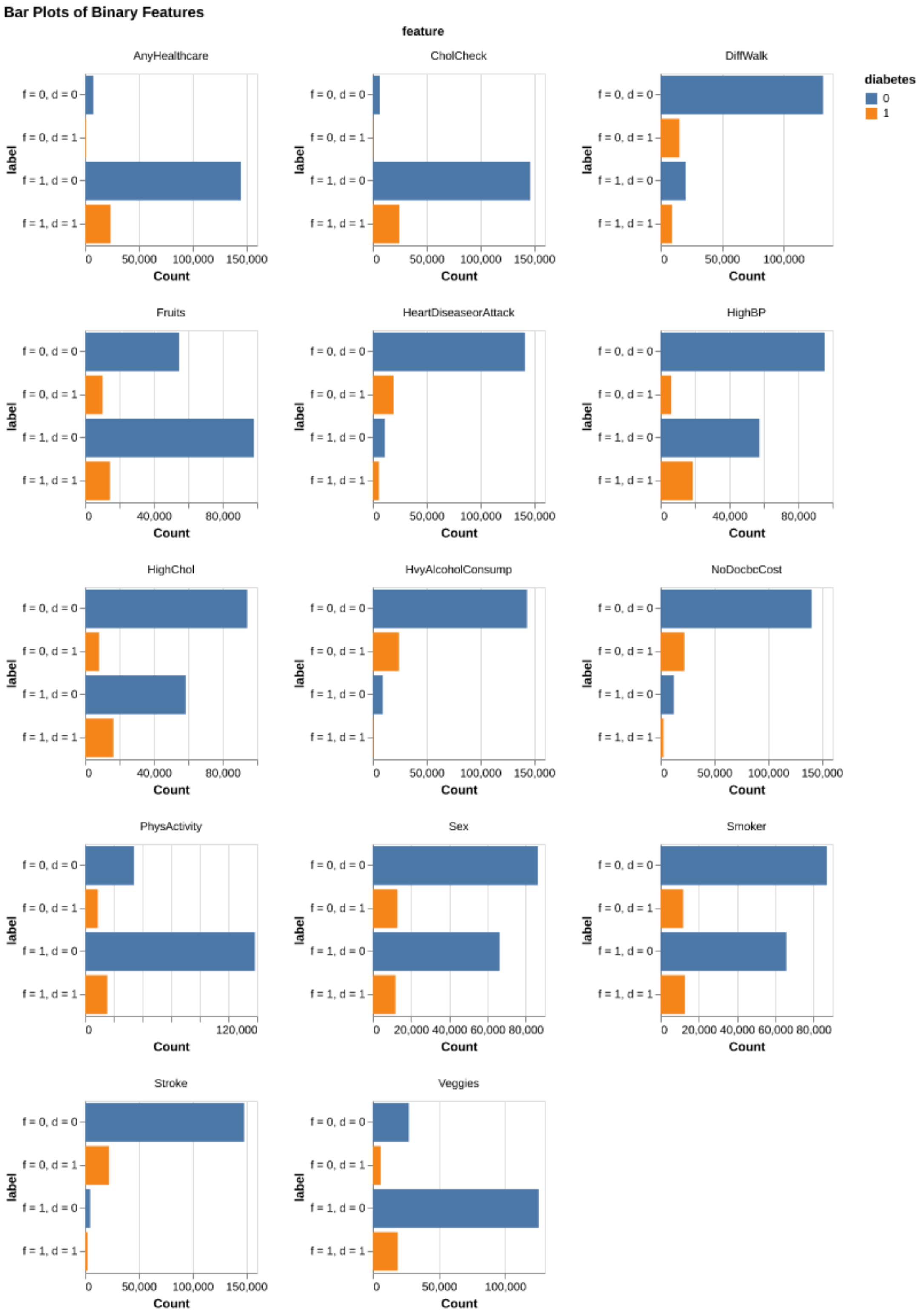

Figure 3: Frequency bar graphs of the binary features.

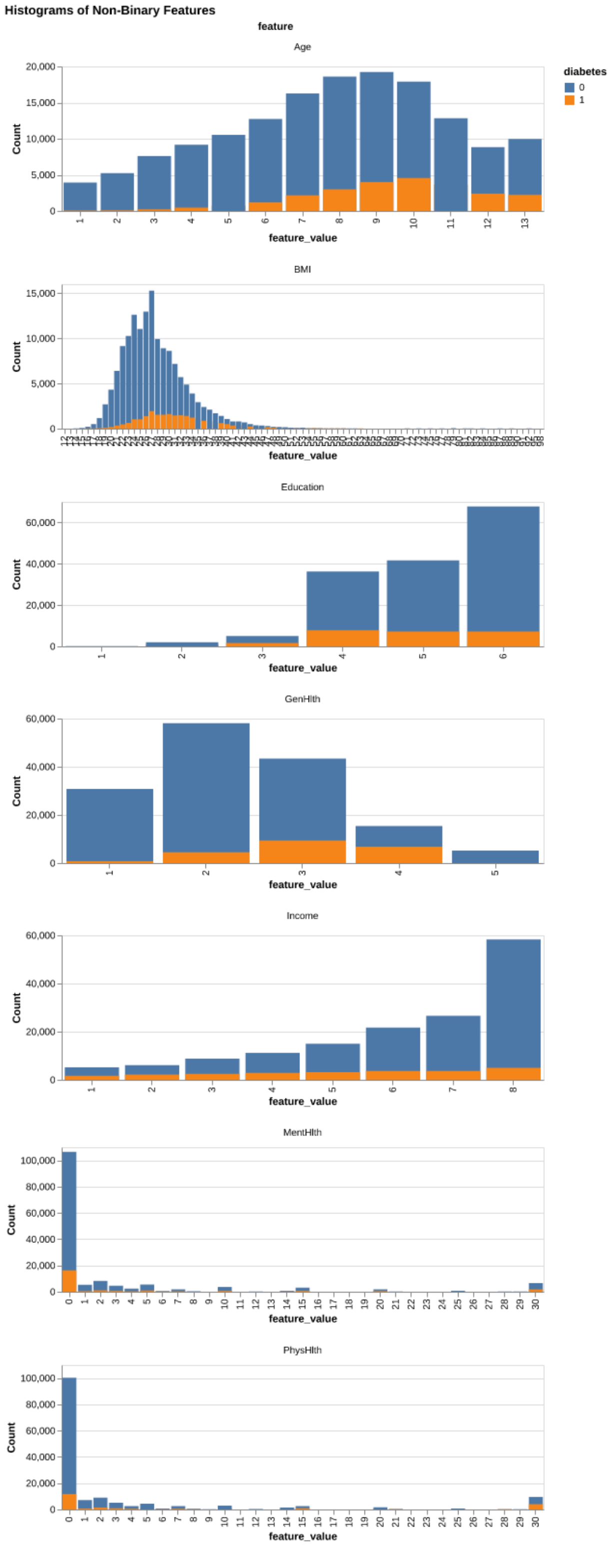

Figure 4: Histograms of the non-binary numeric features.

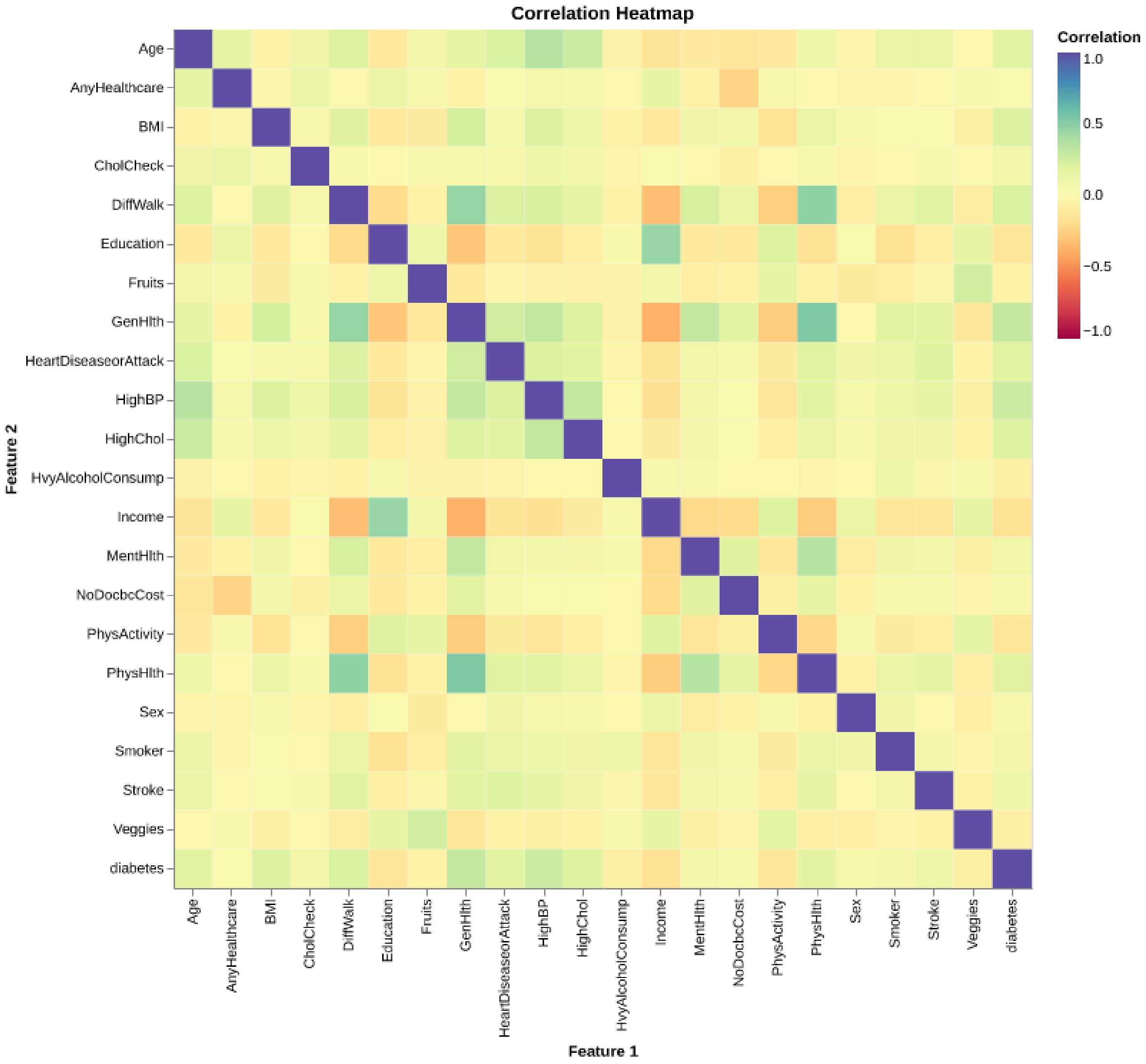

Figure 5: Feature-feature correlation plot of all features.

Modeling

Classification Analysis

Model Selection and Hyperparameter Tuning

We performed grid search cross-validation to optimize hyperparameters for both models:

Decision Tree Classifier: - Best max_depth: 10 - Best min_samples_leaf: 225

Naive Bayes Classifier:
- Best alpha parameter: 0.001

The grid search process identified optimal hyperparameters that balance model complexity with predictive performance. The Decision Tree was configured with a maximum depth of 10 and minimum samples per leaf of 225, while the Naive Bayes model used an alpha smoothing parameter of 0.001.

Result Visualizations

Model Performance Comparison

The following table summarizes the performance metrics for our two models evaluated on the test set:

Table 6: Comparison of model performance metrics on the test set

Unnamed: 0	Model	Test Accuracy	Test f2-score	Test recall	Test precision
0	Decision Tree	0.706	0.587	0.784	0.293
1	Naive Bayes	0.814	0.46	0.489	0.373

As shown in Table 6, the Decision Tree model demonstrates superior performance across all evaluation metrics compared to the Naive Bayes model. The Decision Tree achieves higher accuracy, f2-score, recall, and precision scores, making it the optimal choice for our diabetes prediction task.

Performance Visualization

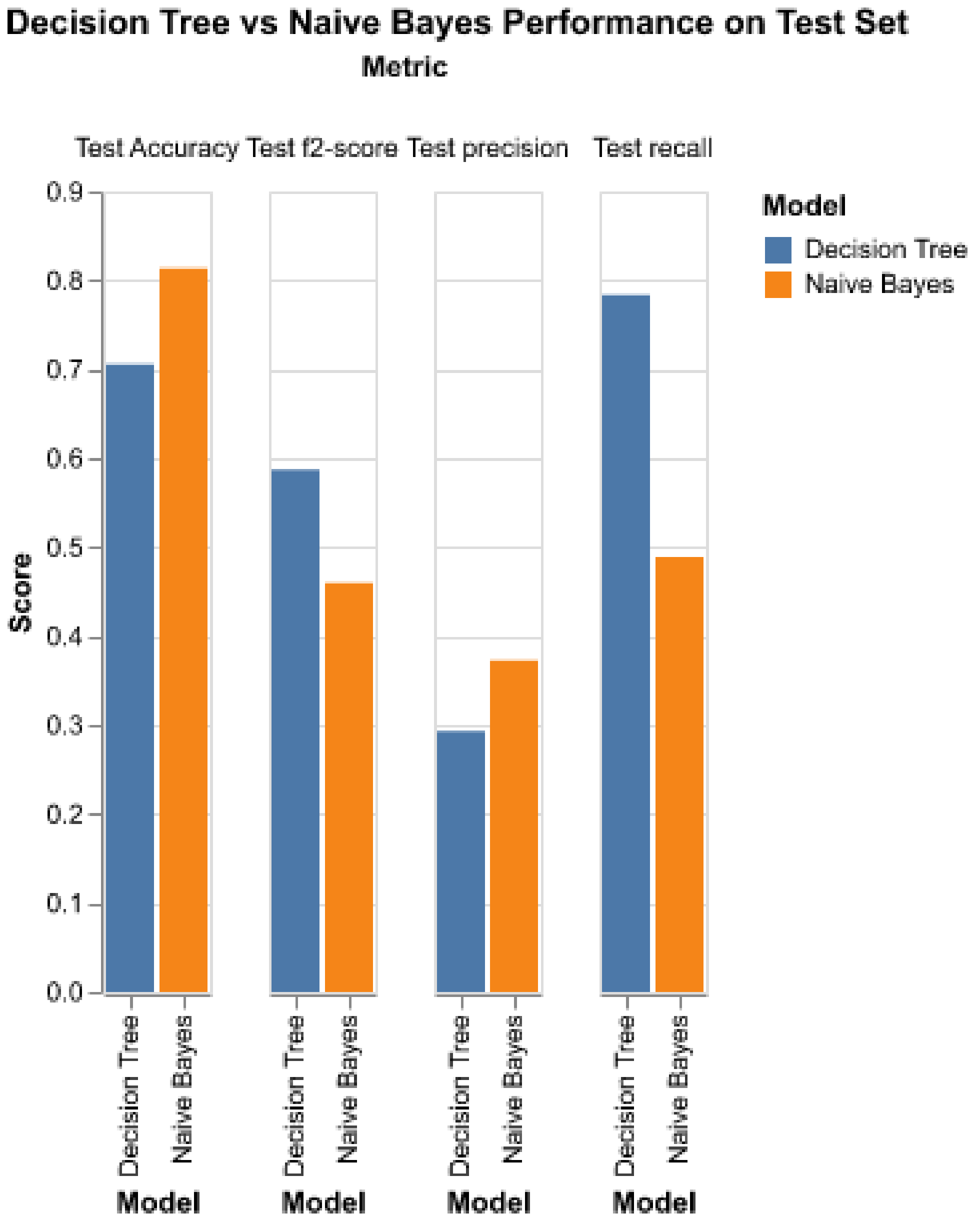

Figure 6: Comparison of model performance metrics on the test set

The performance comparison in Figure 6 demonstrates that the Decision Tree model outperforms the Naive Bayes model across all evaluation metrics.

Confusion Matrix Analysis

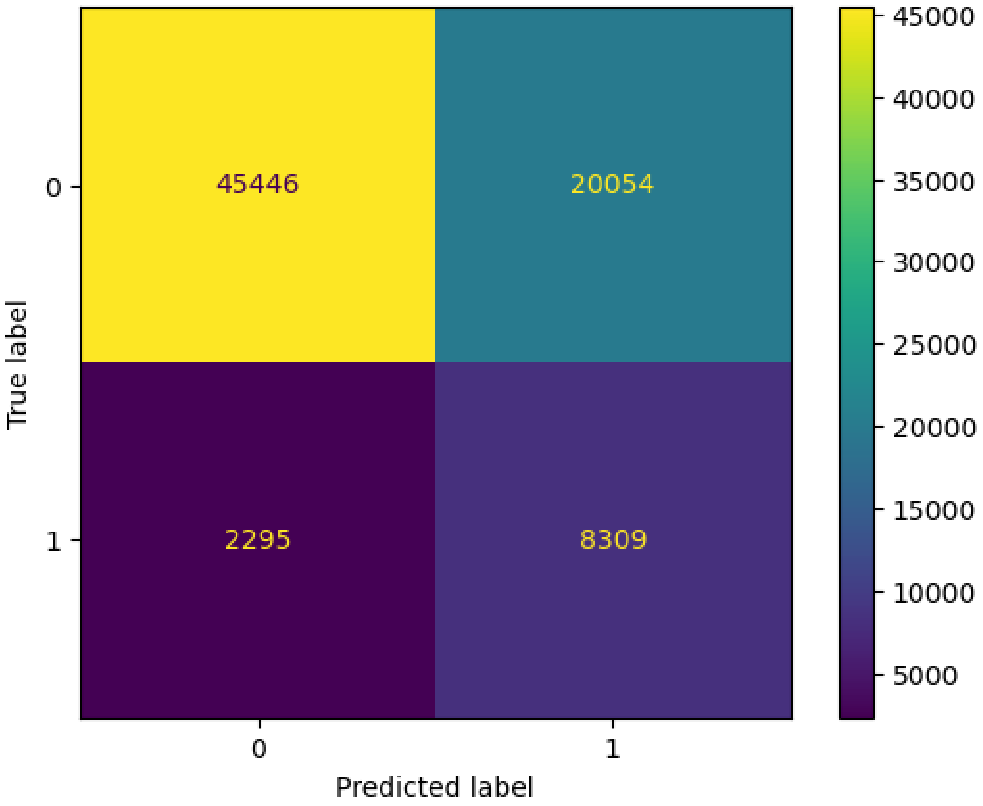

Figure 7: Confusion matrix for the best performing model (Decision Tree)

The confusion matrix in Figure 7 shows the detailed classification performance of our best model. The model correctly identified a significant portion of the test cases, demonstrating strong predictive capability.

Discussion

The current performance of the Decision Tree model is likely already good enough to offer some benefit in the real world given the large number of people with undiagnosed diabetes. However, the recall score of 78.4% could likely be improved, and the precision score of 29.3% definitely leaves something to be desired.

We were surprised by the high rate of false positives shown in Figure 7, which might be an indication of how many non-diabetic people are at risk. Further improvements to predicting diabetes could likely be found by trying a wider variety of model types and using a wider hyperparameter search, and possibly through more feature engineering. A future study could be done to find a smaller set of the most easy to obtain features, as such a model would be more usable by the average person.

Some work is needed to determine this smaller number of easy to obtain features that doesn't significantly reduce model performance. Another question is whether a regression model could be made that predicts a person's risk as a percent chance of developing diabetes, though longitudinal data might be required for this type of prediction.

References

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, Christophe Dervieux, and Gordon Woodhull. 2025. "Quarto." https://doi.org/10.5281/zenodo.5960048.

Astanin, Sergeio. 2025. "Tabulate: Pretty-Print Tabular Data in Python." https://github.com/astanin/python-tabulate.

Bantilan, Niels. 2020. "Pandera: Statistical Data Validation of Pandas Dataframes." In *Proceedings of the 19th Python in Science Conference*, edited by Meghann Agarwal, Chris Calloway, Dillon Niederhut, and David Shupe, 116–24. https://doi.org/ 10.25080/Majora-342d178e-010 .

Chorev, Shir, Philip Tannor, Dan Ben Israel, Noam Bressler, Itay Gabbay, Nir Hutnik, Jonatan Liberman, Matan Perlmutter, Yurii Romanyshyn, and Lior Rokach. 2022. "Deepchecks: A Library for Testing and Validating Machine Learning Models and Data." https://arxiv.org/abs/2203.08491.

Dane, Sohier, and Alex Teboul. 2021. "Diabetes Health Indicators Dataset." https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data; Center of Disease Control.

Kelly, Markelle, Rachel Longjohn, and Kolby Nottingham. 2021. "The UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. https://doi.org/ 10.25080/Majora-92bf1922-00a .

NumPy Documentation. 2008-2022. https://numpy.org/doc/1.26/.

Pallets. 2020. *Click*. https://click.palletsprojects.com/.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Python 3.11.6 Documentation. 2021-2025. https://docs.python.org/release/3.11.6/reference/index.html.

Rios, Nilka Burrows, Isreal Hora, Linda S. Geiss, Edward W. Gregg, and Ann Albright. 2017. "Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico." *Incidence of End-Stage Renal*

Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico, no. 66(43): 1165–70.

Satyanarayan, Arvind, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. "Vega-Lite: A Grammar of Interactive Graphics." *IEEE Transactions on Visualization and Computer Graphics* 23 (1): 341–50.

"Snapshot of Diabetes in Canada, 2023." 2023. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/snapshot-diabetes-canada-2023.html; Public Health Agency of Canada.

Stafford, Lauryn K, Anna Gage, Yvonne Yiru Xu, Madeleine Conrad, Ismael Barreras Beltran, and Edward J Boyko. 2025. "Global, Regional, and National Cascades of Diabetes Care, 2000–23: A Systematic Review and Modelling Analysis Using Findings from the Global Burden of Disease Study." *The Lancet Diabetes & Endocrinology*, no. 13(11): 924–34.

team, The pandas development. 2020. "Pandas-Dev/Pandas: Pandas." Zenodo. https://doi.org/10.5281/zenodo.3509134.

VanderPlas, Jacob, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (32): 1057. https://doi.org/10.21105/joss.01057.