DEVELOPING CLASSIFICATION MODELS TO PREDICT DIABETES

LILLIAN MUELLER | ENPM808L | 4 DEC 2023

AGENDA

- INTRODUCTION
- METHODOLOGY
- RESULTS
- DISCUSSION

INTRODUCTION

Diabetes and Prediabetes

- Metabolic condition that affects millions of people globally
- 8th leading cause of death in United States
- Types of Diabetes
  - Type 1
  - Type 2
  - gestational diabetes
- Prediabetes
  - Higher than normal blood glucose levels
  - Early identification can allow individual to implement prevention strategies

Models to Facilitate Diabetes Prevention

- Identify individuals who are at risk or already affected by diabetes
- Use classification models to predict the likelihood of individuals have diabetes or prediabetes
  - Decision tree
  - Logistic regression
  - K-nearest neighbor
  - Naïve Bayes classifier
  - Linear discriminant analysis

METHODOLOGY

Cleaning the Dataset

- Dataset sourced from Kaggle
  - Health indicators
  - Demographics
  - Lifestyle attributes
- Cleaned each column to contain only numeric values
  - Scaled by ranges
  - Boolean features
- Target Classification Prepressing
  - 0 = no diabetes
  - 1 = prediabetes
  - 2 = diabetes

Model Variation Selection

- Decision Tree
  - Entropy/Gini Criterion
  - Maximum depth
- Logistic Regression
  - SAG/SAGA solver
  - L2 penalty or None
- K-Nearest Neighbor
  - K values
  - Uniform/Distance weight
- Naïve Bayes Classifier
  - Gaussian/Bernoulli
- Linear Discriminant Analysis

Evaluation and Comparison

- Accuracy
  - Overall correctness
- Precision
  - Correctness when predicting a specific classification
- Confusion Matrices
  - Visualize FP, TP, FN, FP
- ROC Curves
  - Rank models on performance
- Cross Validation
  - General accuracy and precision

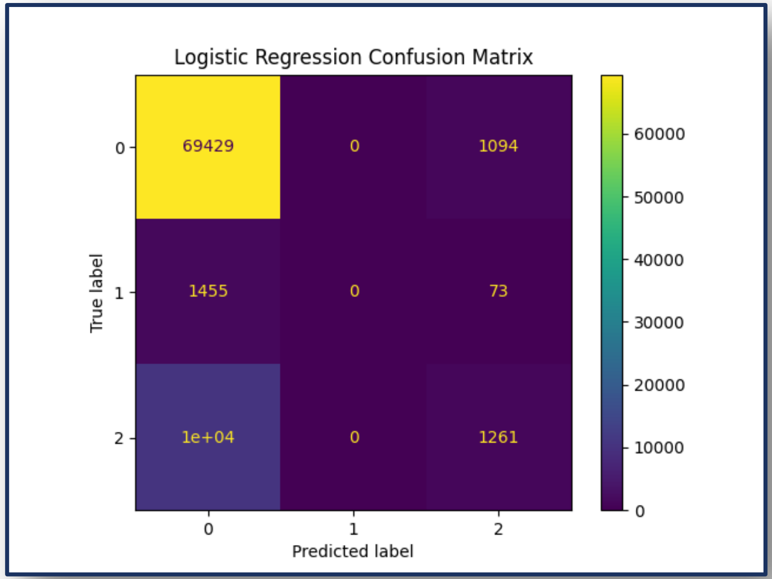
RESULTS

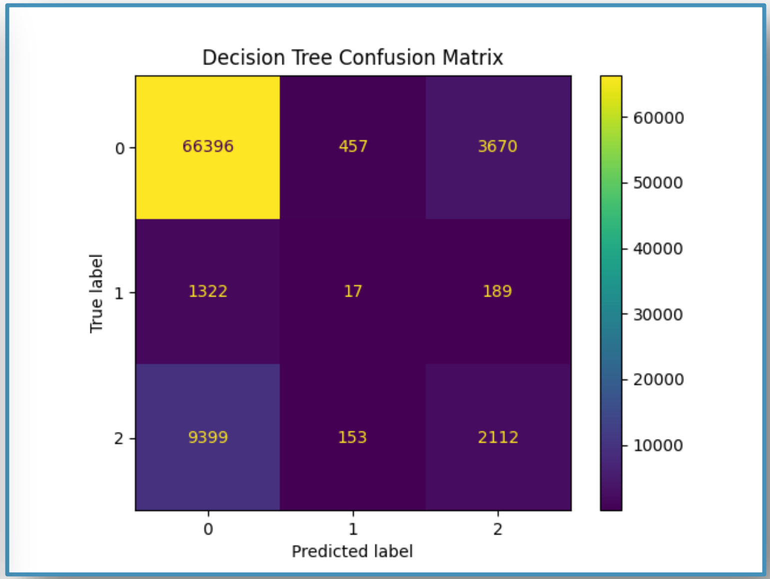
- Decision Tree
  - Gini Impurity Index
  - No maximum constraint
- Logistic Regression
  - Stochastic Average Gradient Descent (SAGA)
  - No penalty
- KNN
  - K = 25
  - Weighted distance using Euclidean
- Naïve Bayes
  - Gaussian
  - Bernoulli
- Linear Discriminant Analysis
  - Singular Value Decomposition (SVD)

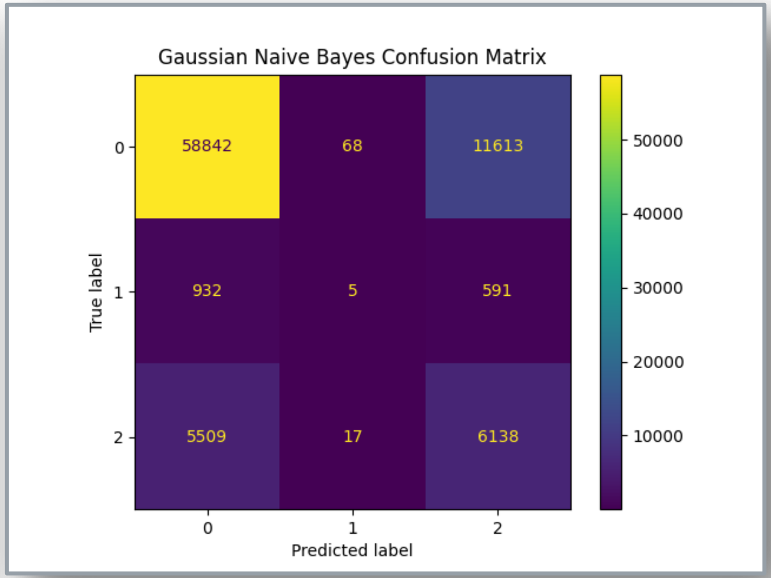
Model	Accuracy	Precision 0	Precision 1	Precision 2
Decision Tree	0.785	0.833	0.014	0.361
Logistic Reg.	0.815	0.826	0.	0.521
KNN	0.814	0.826	0.	0.502
Gaussian NB	0.739	0.875	0.018	0.341
Bernoulli NB	0.797	0.833	0.	0.392
LDA	0.814	0.830	0.	0.497

TABLE I
MODEL ACCURACY AND PRECISION

- Greatest Accuracy: Logistic Regression
- Greatest Precision: Gaussian Naïve Bayes







Accuracy	.815		
Precision	.826	.0	.521

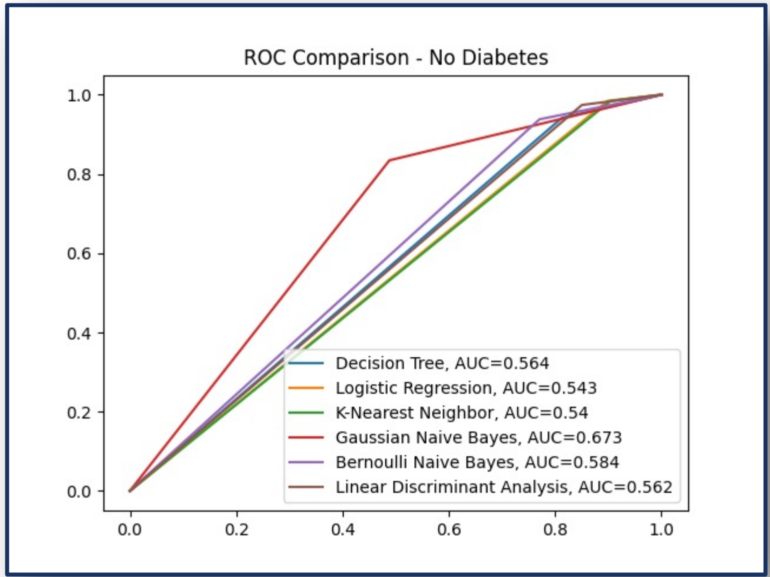
Accuracy	.785		
Precision	.833	.014	.361

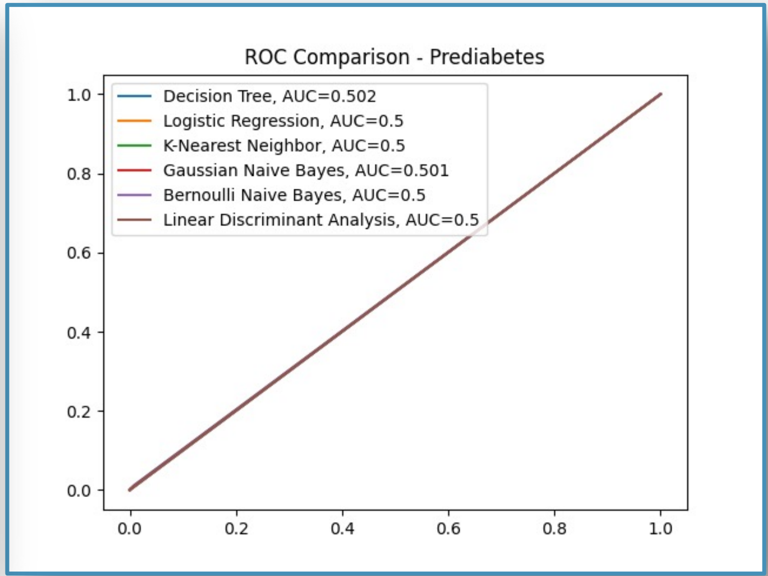
Accuracy	.739		
Precision	.875	.018	.341

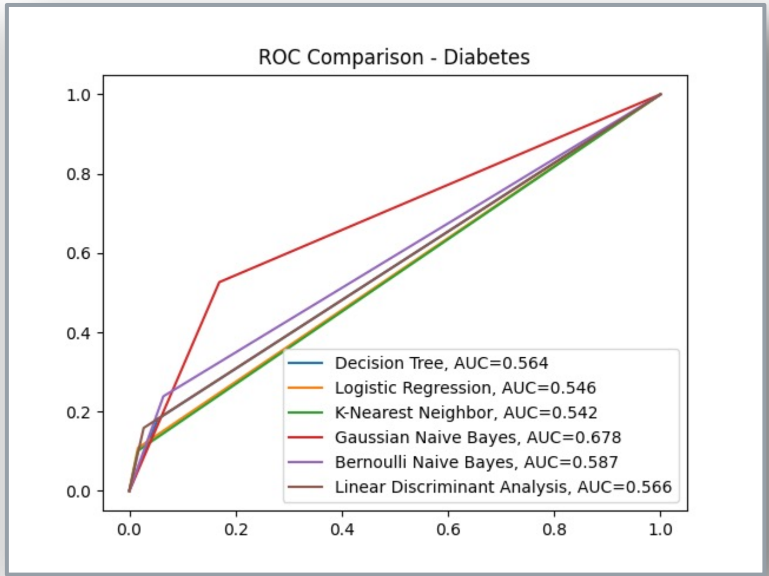
RESULTS

Confusion Matrices

Receiver Operating Characteristics Curves







RESULTS

Rank	No Diabetes	Prediabetes	Diabetes
1	Gaussian NB	Decision Tree	Gaussian NB
2	Bernoulli NB	Gaussian NB	Bernoulli NB
3	Decision Tree	-	Decision Tree
4	LDA	-	LDA
5	Logistic Reg.	-	KNN
6	KNN	-	Logistic Reg.

TABLE II
MODEL RANKINGS BASED ON AUC

DISCUSSION

- Gaussian naïve Bayes classifier proved to be the best classification model for predicting diabetes
- Accuracy versus Precision
  - When dealing with datasets with high imbalance, precision is a better performance indicator
  - High precision indicates fewer false positive predictions
  - High cost for misdiagnosis
- Addressing the dataset's imbalance to achieve better performing classification models
  - Resampling the data
  - Boosting or tree-based models
  - Collect more data

Model	Mean Accuracy	Standard Dev.
Decision Tree	0.820	0.003
Logistic Reg.	0.843	0.005
KNN	0.843	0.001
Gaussian NB	0.775	0.009
Bernoulli NB	0.824	0.002
LDA	0.841	0.005

TABLE III
10-FOLD CROSS VALIDATION ACCURACY

Model	Mean Precision	Standard Dev.
Decision Tree	0.010	0.005
Logistic Reg.	0.000	0.000
KNN	0.000	0.000
Gaussian NB	0.298	0.399
Bernoulli NB	0.000	0.000
LDA	0.000	0.000

TABLE IV
10-FOLD CROSS VALIDATION PRECISION OF PREDICTING PREDIABETES

CITATIONS

- [1] "About Prediabetes and Type 2 Diabetes — National Diabetes Prevention Program — CDC." Accessed: Nov. 23, 2023. [Online]. Available: https://www.cdc.gov/diabetes/prevention/about-prediabetes.html
- [2] J. P. Crandall et al., "The prevention of type 2 diabetes," Nat. Clin. Pract. Endocrinol. Metab., vol. 4, no. 7, pp. 382-393, Jul. 2008, doi: 10.1038/ncpendmet0843.
- [3] "Diabetes Health Indicators Dataset." Accessed: Nov. 23, 2023. [On-line]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
- [4] L. Mueller and R. Hong, "Investigating Decision Trees".
- [5] ArnavR, "Scikit-learn solvers explained," Medium. Accessed: Nov.25, 2023. [Online]. Available: https://medium.com/@arnavr/scikit-learn-solvers-explained-780a17bc322d
- [6] L. Mueller and R. Hong, "Iris Classification Using Logistic Regression".
- [7] L. Mueller and R. Hong, "Evaluating the Performance of K-NearestNeighbors Classification".7
- [8] S. Ray, "Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier," Analytics Vidhya. Accessed: Nov. 25, 2023. [Online]. Available:https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
- [9] L. Mueller, "Comparing Classifications Models Against the Naive Bayes Classifier and Linear Discriminant Analysis Model".
- [10] L. Mueller, "Evaluating Classifications Models using Confusion Matrices".
- [11] L. Mueller, "Ranking Classification Models using Receiver Operating Characteristics".
- [12] L. Mueller and R. Hong, "Using K-Fold Cross Validation on Decision Tree and Logistic Regression Models to Classify Iris Species".
- [13] "Accuracy vs. precision vs. recall in machine learning: what's the difference?" Accessed: Nov. 26, 2023. [Online]. Available:https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall
- [14] R. Feki, "Imbalanced data: best practices," Medium. Accessed: Nov. 26, 2023. [Online]. Available: https://rihab-feki.medium.com/imbalanced-data-best-practices-f3b6d0999f38