Predicting Diabetes with K-Nearest Neighbor and Logistic Regression

Abel Varghese

ITCS 3156

May 11, 2025

Dr. Xue

Introduction

Diabetes mellitus is a chronic disease that affects millions of individuals worldwide, leading to serious health complications if not managed effectively. Machine learning offers powerful tools for early detection and risk stratification of diabetes, enabling clinicians to identify high-risk patients and intervene proactively. The objective of this project is to build and evaluate a comprehensive machine learning pipeline using a publicly available diabetes prediction dataset. This pipeline - including data exploration, preprocessing, model selection, and result analysis- is explored in this report and further built upon with provided potential improvements. This workflow adheres to best practices in reproducible research and aims to balance predictive performance with interpretability. I had chosen this subject as diabetes has a history in my family as well as in my native country, so furthering my understanding of the disease would be personal and academic motivation. Two machine learning models were built and analyzed in the project with varying degrees of success: K-Nearest Neighbor and Logistic Regression.

This project can be accessed from the following GitHub repository:
https://github.com/AbelV0538/diabetes-prediction-with-ml/

Data

1. Data Introduction

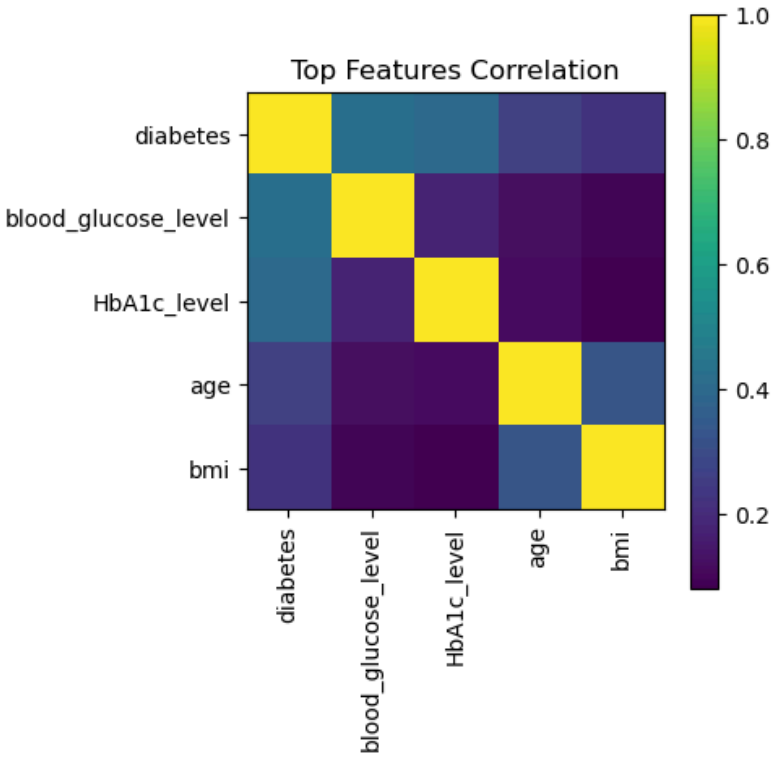
The dataset used in this study was obtained from Kaggle (Diabetes Prediction Dataset) and consists of 100,000 patient records. Each record includes demographic information (gender, age), comorbidity indicators (hypertension, heart disease), lifestyle factors (smoking history), physiological measurements (BMI, HbA1c_level, blood_glucose_level), and the binary target variable diabetes indicating disease presence (1) or absence (0).

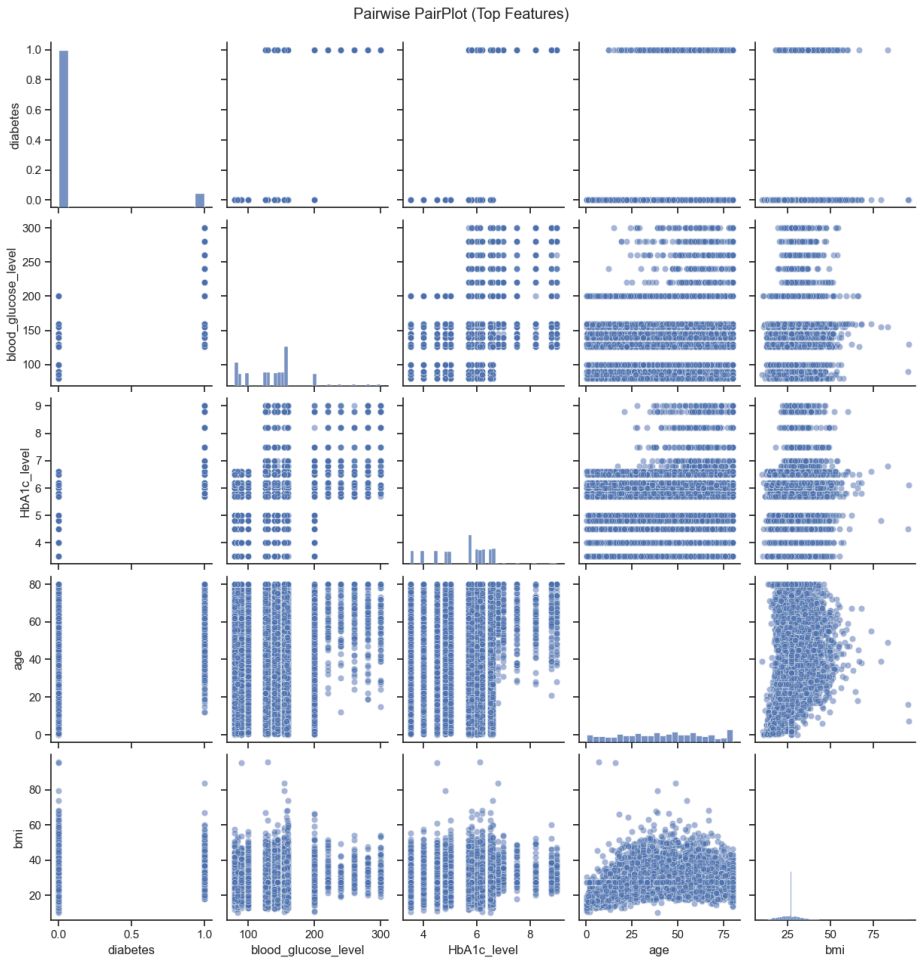
2. Exploratory Analysis

The data was first loaded into a pandas DataFrame, and a schema was created to confirm data integrity. Basic statistics revealed:

- **Gender distribution:** approximately balanced male/female.
- **Age range:** 1 month to 80 years, mean age ~41 years.
- **Physiological measures:**
  - bmi mean: ~27, considered overweight.
  - HbA1c_level mean: ~5.5, considered prediabetic
  - Blood_glucose_level mean: 138 mg/dl, 100-125 is considered prediabetic,
  - Hypertension mean: ~0.07 [0 is not having hypertension and 1 is having hypertension]
- **Behavior measure:** smoking_history mean:~ 2.7 [0 →No Info, 1→Current, 2→ One time ever, 3 →Former , 4→Never, 5→ Not Current]

To visualize the correlation between the features, a correlation heat map and a pairwise plot were used. These graphics only accounted for the top five features in correlation, according to Panda's correlation function. Due to the sheer size of the dataset, 10,000 random samples of the 100,000 samples were used to create these graphics to avoid lengthy computational times.





Preprocessing

1. Handling Missing and Categorical Data

Upon inspection, all necessary values were present in the dataset, but two features were not numerical values. These two were the gender and smoking_history features, which were described with string literals and had to be mapped from a string to an integer value. For gender, Female → 0, Male → 1, and Other → 2. The mapping for smoking_history has been stated in the exploratory analysis above.

2. Train/Test Split

For both models, a train/test split of 80 and 20 percent, respectively, was utilized, as it provided adequate training and testing data. However, in completing this research, K-NN was performed with only 10,000 samples of the total 100,000 samples in the dataset. As stated earlier, this was done to efficiently conduct the research.

Methods

1. k-Nearest Neighbors (k-NN)

- **Rationale:** Simple, non-parametric, and intuitive. Good baseline for classification when decision boundaries can be approximated by local neighborhoods.

- **Implementation:** We implemented k-NN from scratch using Euclidean distance and majority voting. The hyperparameter k was explored from one to 20, as a means of fine-tuning for best accuracy.

2. Logistic Regression (LR)

- **Rationale:** Interpretable coefficients direct probability estimates, and well-studied convex optimization with a guaranteed global minimum.

- **Implementation:** A scratch implementation using batch gradient descent, minimizing binary cross-entropy loss. Experimented with learning rates in {0.01, 0.0.5, 0.1} and iterations of {500, 1000, 1500} epochs for fine tuning.

Results

1. Hyperparameter Tuning

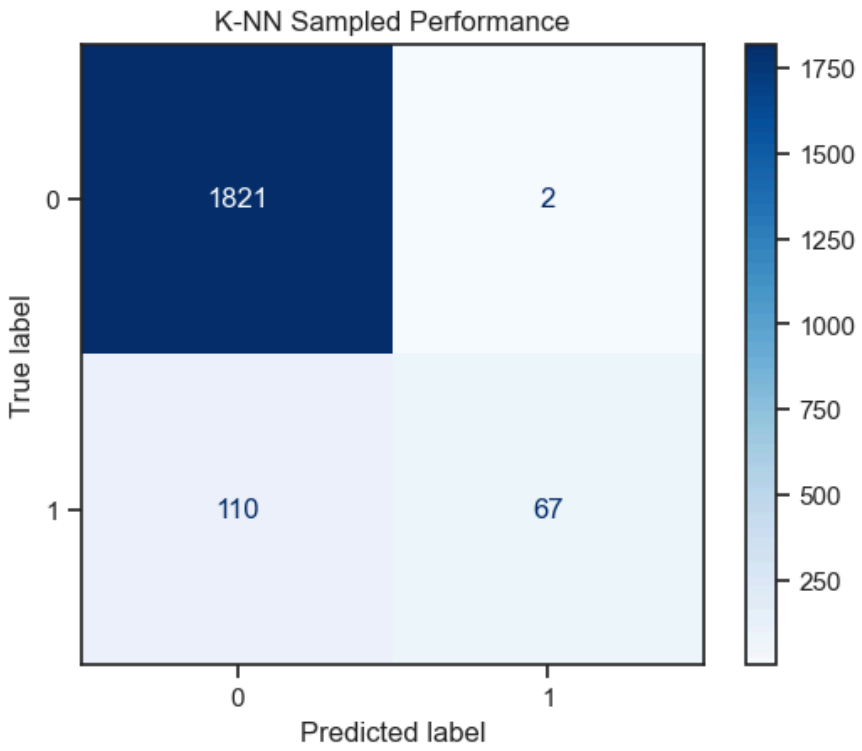
- **k-NN:** Iterating through 1-20 as possible values of k,k of 11 provided the best accuracy. This k value was also tested against the entire dataset and had a similarly strong performance. (This took 10 hours to compute on the laptop the program was running on)

- **LR:** Testing all possible combinations of learning rates and epochs listed above, the best turned out to be 500 epochs and a learning rate of 0.01. This was completed with two for loops, which redefined the best accuracies and hyperparameters when the previous values were outperformed by the current values.

2. Test Performance

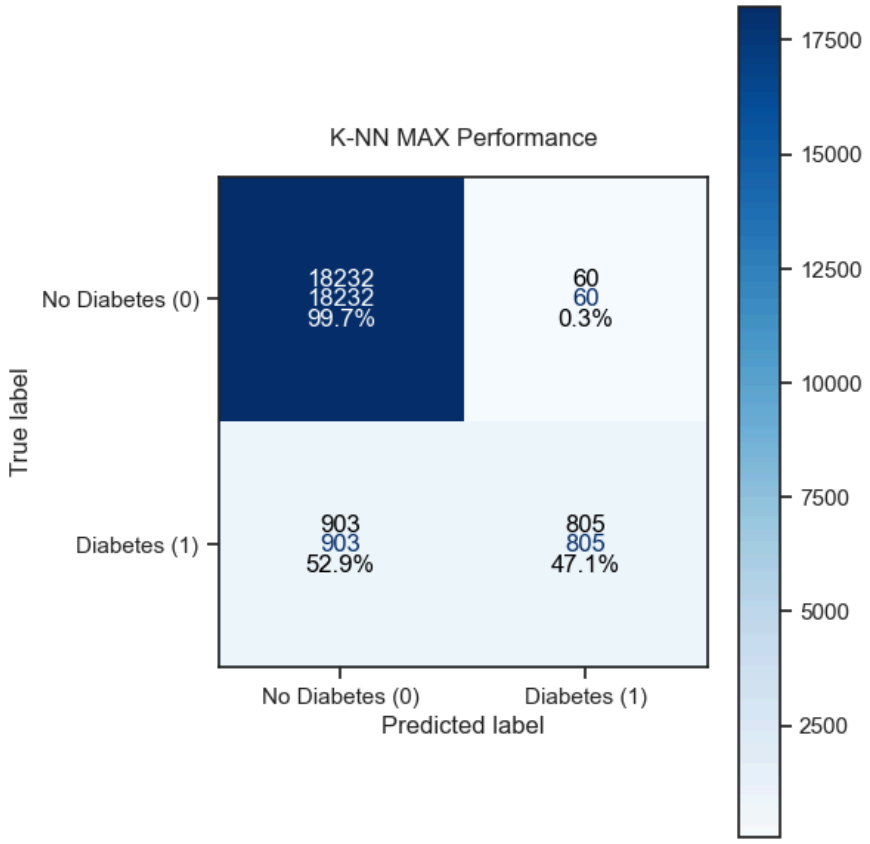
KNN (10,000 samples), Overall Accuracy: 0.944

Class	Precision	Recall	F_1-Score	Support
0	0.94	1.00	0.97	1823
1	0.97	0.38	0.54	177
macro avg	0.96	0.69	0.76	2000
weighted avg	0.95	0.94	0.93	2000



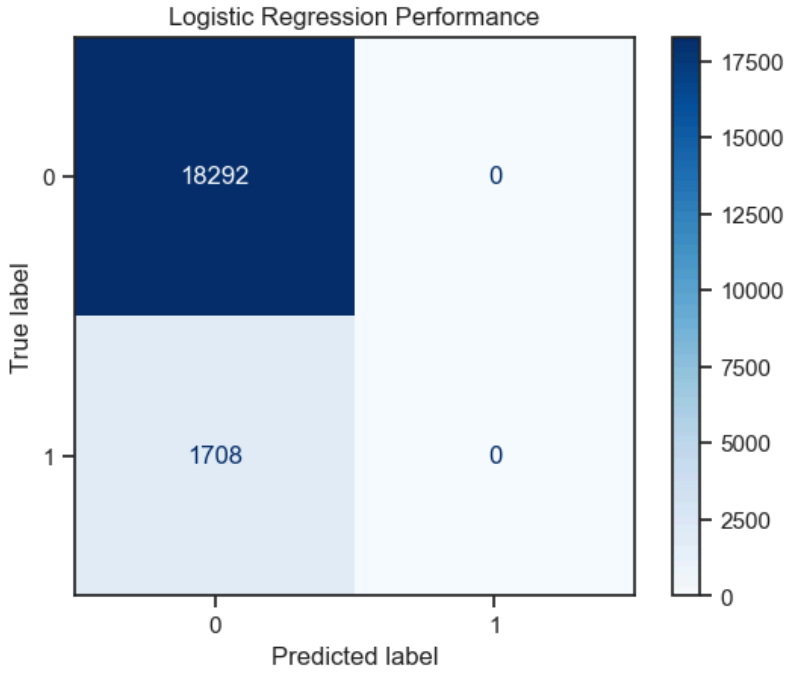
KNN(Full data set): Overall Accuracy 0.952

Class	Precision	Recall	F_1-Score	Support
0	0.95	1.00	0.97	18 292
1	0.93	0.47	0.63	1 708
Macro avg	0.94	0.73	0.80	20000
Weighted avg	0.95	0.95	0.94	20000

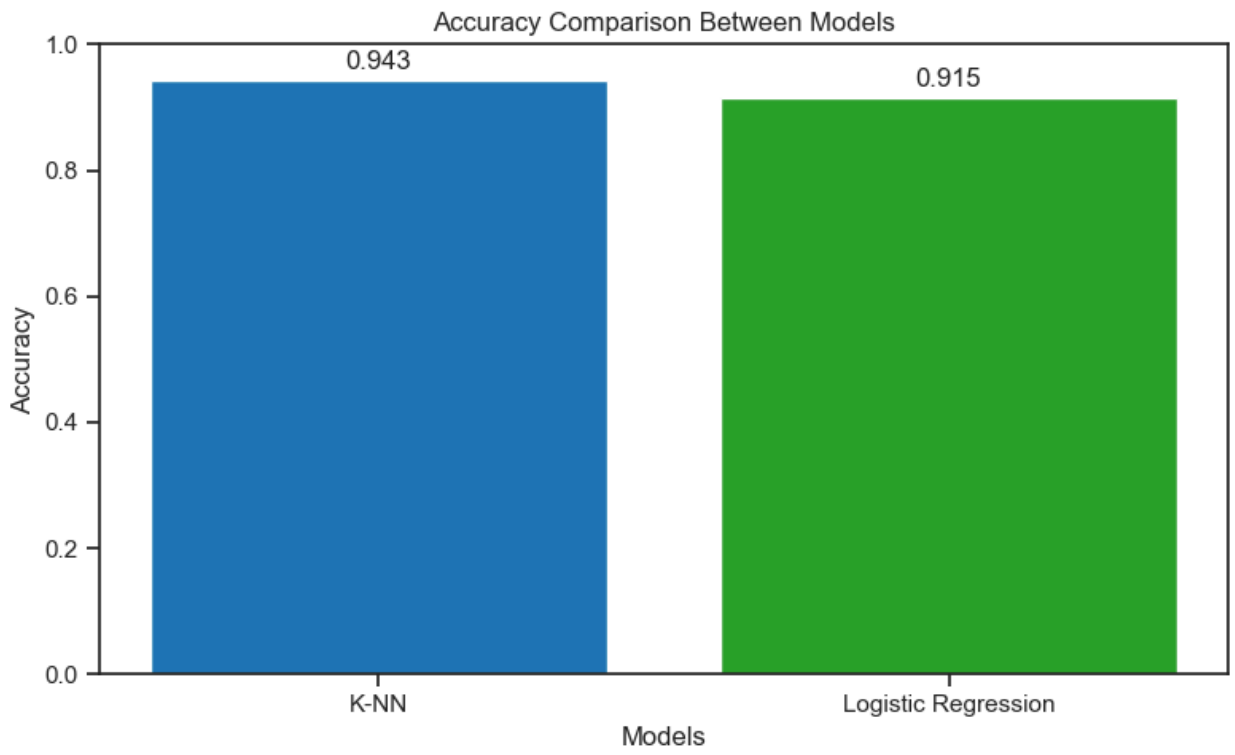


Log-Reg(0.01 learning rate, 500 epochs): Overall Accuracy 0.9146

Class	Precision	Recall	F_1-Score	Support
0	0.91	1.00	0.96	18 292
1	0.00	0.00	0.00	1 708
Accuracy	—	—	0.91	20 000
Macro avg	0.46	0.50	0.48	20 000
Weighted avg	0.84	0.91	0.87	20 000



Observations: By percentages alone, KNN outperforms Log-reg by about 3 percent. However, Log-reg produced the least number of errors compared to KNN and had a much faster computational time (KNN took 10 hours to compute for the whole dataset).



Discussion and Potential Improvements

1. **Non-linear Relationships:** Features like BMI and smoking history may interact non-linearly. Future work could include polynomial features, interaction terms, or kernelized methods (e.g., SVM with RBF kernel).

2. **Validation:** In-depth cross-validation would have saved vast amounts of time and produced similar results for both models.

3. **Imbalanced Learning:** Techniques such as SMOTE oversampling or class weighting may boost recall for diabetic cases without sacrificing precision.

4. **Standardization:** Normalization of features would have led to shorter computational times, allowing for more efficient research.

Conclusions

This project implemented a full machine learning pipeline for diabetes prediction, from exploratory analysis to model evaluation. Logistic Regression achieved the best balance of interpretability and numerical performance, whilst KNN had a higher percentage performance. It is imperative to mention the possibility of creator error, as if these test models were designed and tested professionally, the results would be much coherent. This project was quite the learning experience, as it demonstrated the concept of compromise very well, with marginal differences creating exponential loss in time or accuracy.

References

1. Kaggle Diabetes Prediction Dataset. By Mohammed Mustafa

   https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

   Statistics for health metrics in Exploratory Analysis:

2. Centers for Disease Control and Prevention. (n.d.). *Adult BMI categories*. Centers for Disease Control and Prevention.

   https://www.cdc.gov/bmi/adult-calculator/bmi-categories

3. U.S. National Library of Medicine. (n.d.). *Hemoglobin A1C (hba1c) test: Medlineplus medical test*. MedlinePlus.

   https://medlineplus.gov/lab-tests/hemoglobin-a1c-hba1c-test/

4. professional, C. C. medical. (2025, February 26). *What does my blood glucose test result mean?*. Cleveland Clinic.

   https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test

5. Mayo Foundation for Medical Education and Research. (2024, March 27). *Diabetes*. Mayo Clinic.

   https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

Acknowledgement

- Data exploration and modeling code utilized Python libraries: pandas, NumPy, Matplotlib, Seaborn, and scikit-learn.

- Various model improvements, bug fixes, and structural ideas were provided by Chat GPT-4o. Analysis and interpretations were done by the author and the author alone.