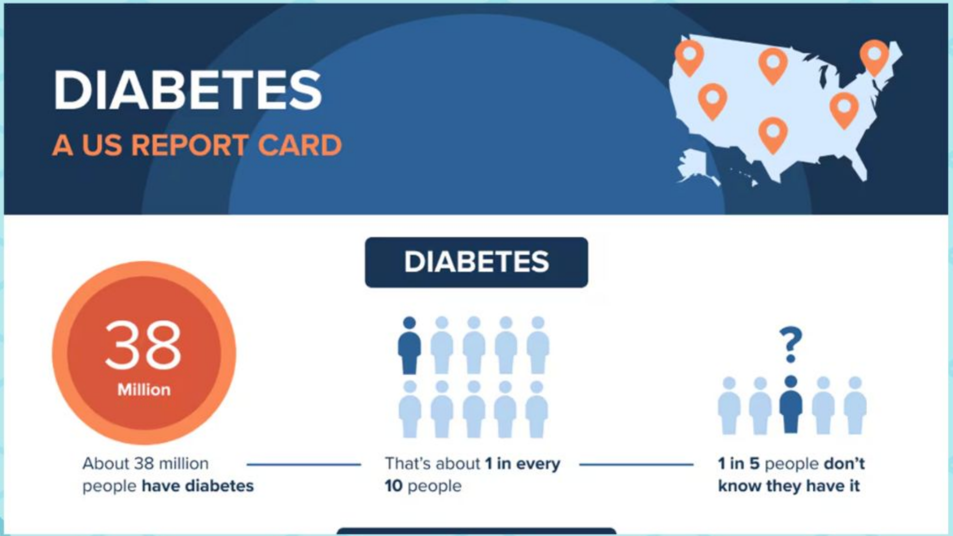
What factors might influence the risk of getting diabetes?

Lanxi Liu, Laura Wang, Minjoo Kim, Zoey Zeng

Introduction

What is the problem?



Why is it interesting/important

A lot of hospitalizations and ER visits due to diabetes are placing a significant burden on healthcare resources

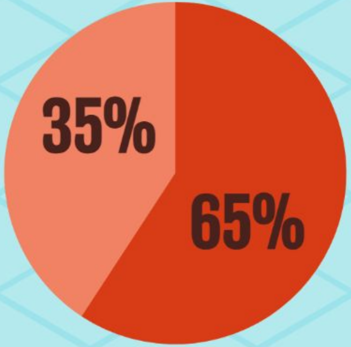
Family genetics and lifestyle drive diabetes risk, so we want to see how ML can help in early detection and risk assessment

Dataset

Pima Indians Diabetes Database

768 rows x 9 columns

- **Pregnancies**: # of pregnancies
- **Glucose**: Plasma glucose concentration
- **Blood Pressure**: Diastolic blood pressure
- **Skin Thickness**: Triceps skin fold thickness
- **Insulin**: 2-Hour serum insulin
- **BMI**: Body mass index
- **Diabetes pedigree function**: Genetic influence
- **Age**
- **Outcome**: Diabetes diagnosis (0 = No, 1 = Yes)



Non-diabetic cases

500 individuals

Diabetic cases

268 individuals

Experiment Setup

Methodology

- Train-test split: 80% training 20% testing

Data Cleaning

- There were no NA values, but a lot of 0 values
- We converted 0 values into the mean of each feature

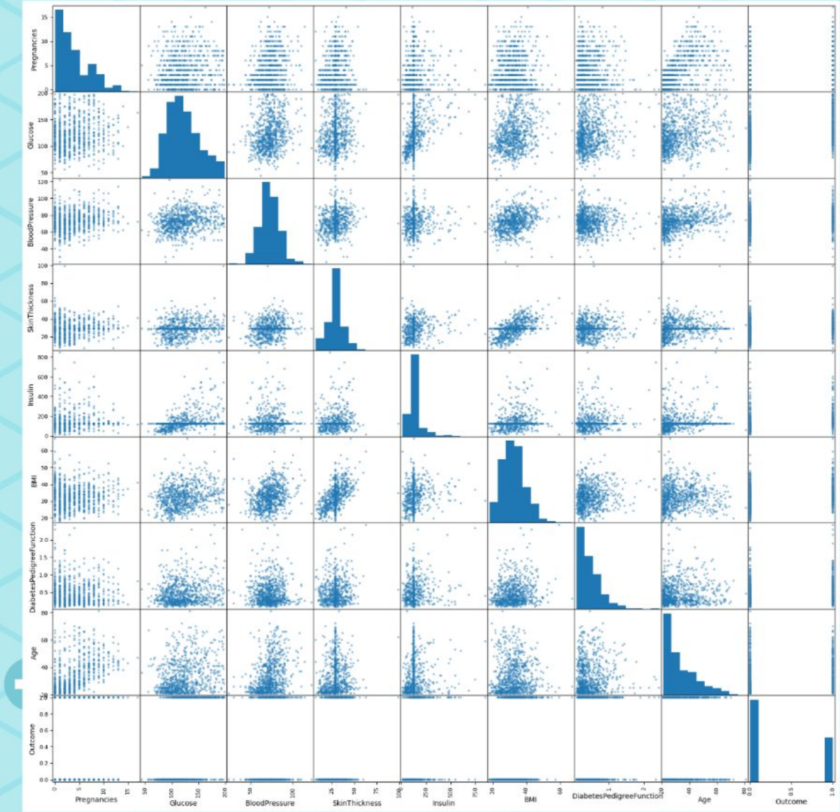
Models Evaluated

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- Neural Network

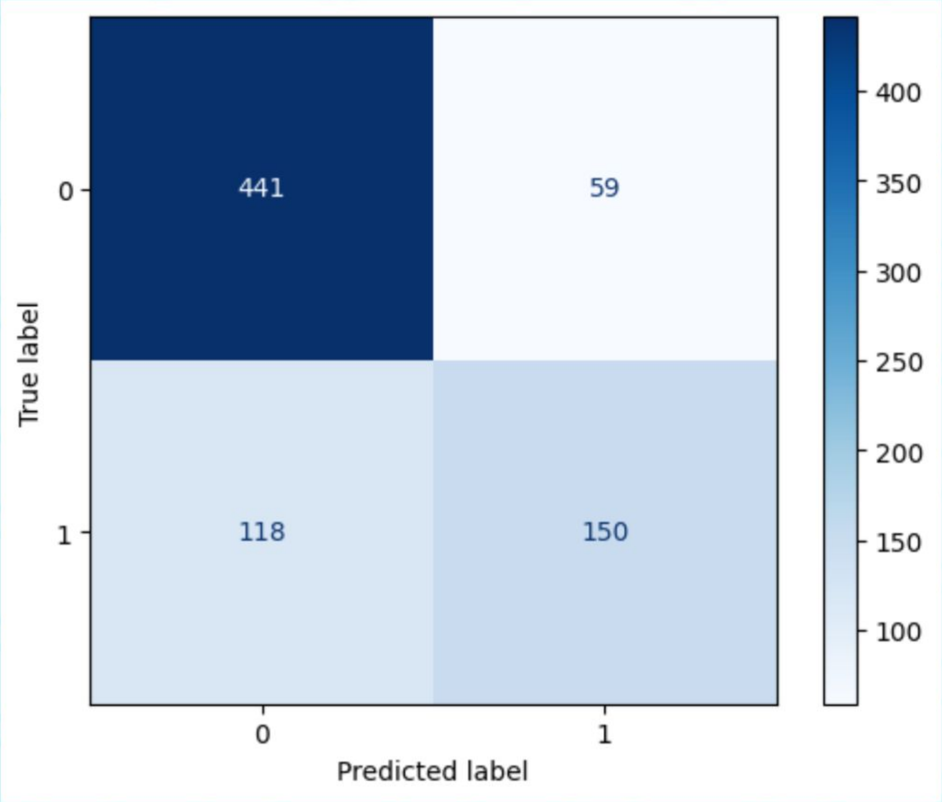
Cross Validation

- K-fold validation

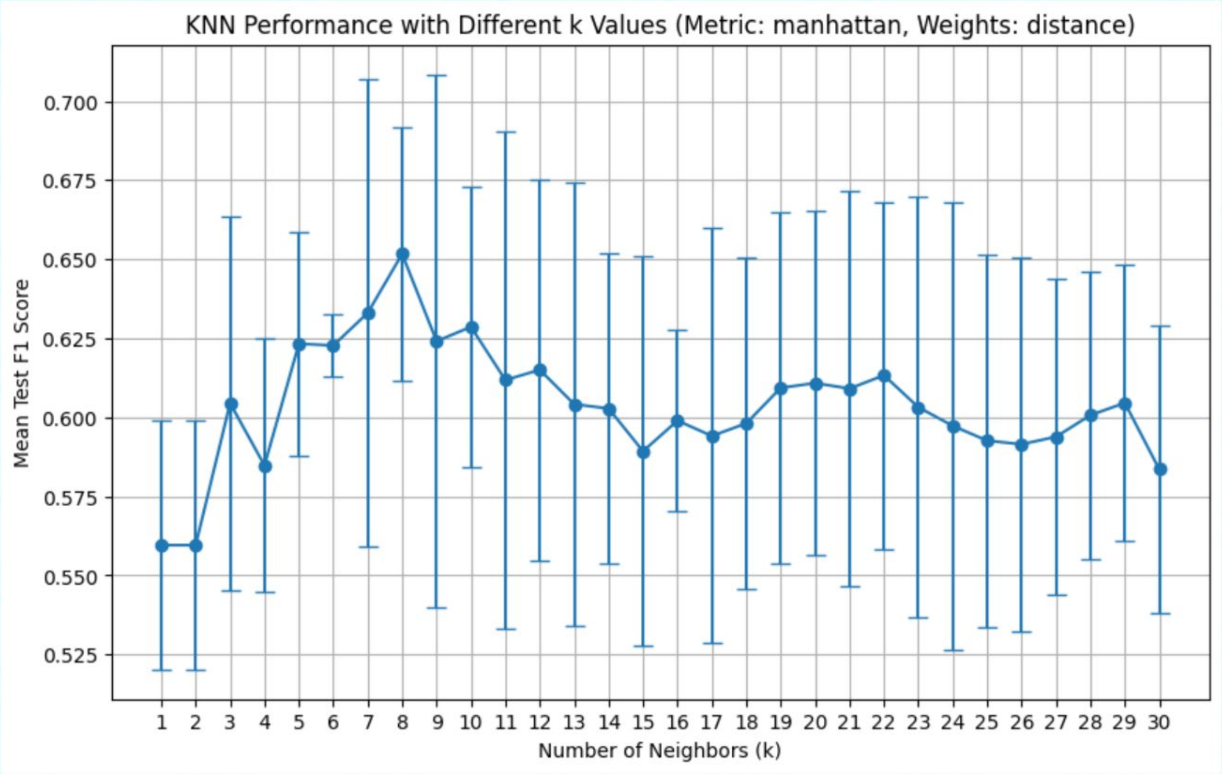
Pairplot and Heat Map



Cross-Validation & Confusion Matrix



KNN

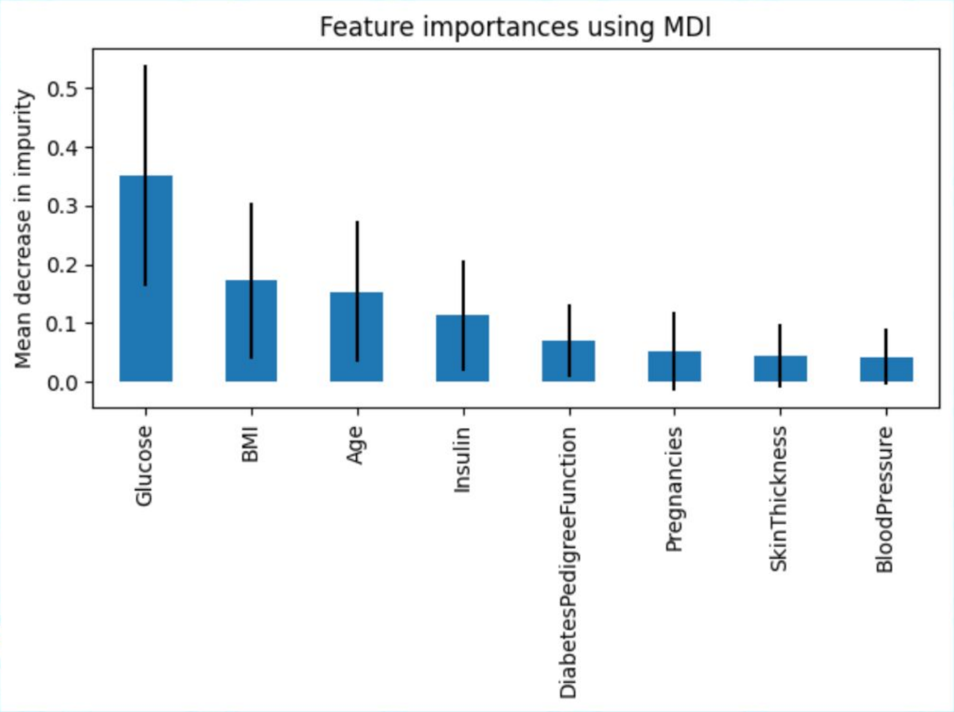

KNN Performance with Different k Values (Metric: manhattan, Weights: distance)
Mean Test F1 Score
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Number of Neighbors (k)

Results Comparison

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Cross-Val Mean	Cross-Val Std
Logistic Regression	0.7532	0.6667	0.6182	0.6415	0.8231	0.7688	0.0304
KNN	0.7273	0.6032	0.6909	0.6441	0.7642	0.7444	0.0289
Decision Tree	0.6948	0.5667	0.6182	0.5913	0.6778	0.6807	0.0453
Random Forest	**0.7597**	0.6607	0.6727	**0.6667**	**0.8254**	**0.7737**	0.0311
Neural Network	0.7468	0.6379	0.6727	0.6549	0.8000	0.7720	0.0275

We are assuming Random Forest as the best model with the highest Accuracy, F1 Score, ROC AUC, and Cross-Val Mean.

Random Forests and Features Importance




```
Feature correlations with Outcome:
 Glucose                      0.489082
BMI                          0.319116
Age                          0.280654
SkinThickness                0.211854
Pregnancies                  0.207550
Insulin                      0.188590
BloodPressure                0.159846
DiabetesPedigreeFunction     0.154560
```

Discussion and limitation

- **Random Forest** achieved the highest ROC AUC (0.8254), making it the best-performing model for predicting diabetes risk.
- Key predictors include **Glucose**, **BMI**, and **Age**, with Glucose having the strongest impact on outcomes.
- Limitation includes not fully represent broader populations, and potential biases in feature selection, data quality, or sample imbalance could impact the generalizability of the results.
- Future studies could benefit from integrating larger, more diverse datasets and incorporating explainable AI techniques to enhance the interpretability of machine learning models.
- Longitudinal data could be used to analyze temporal trends and causal relationships, providing deeper insights into the progression of diabetes.