Diabetes Risk: The Prediction of Diabetes Risk Through Supervised Learning

By: Katherine Nguyen

Data Science
UNIVERSITY OF COLORADO BOULDER

Agenda

- Outline/Agenda
- Introduction
- Approach
  - Data
  - Rundown
  - Data Cleaning
  - Models
  - Model Analysis
- Results
- Conclusion

Introduction

- **Problem**
  - Diabetes is prevalent and being able to predict whether one has diabetes is useful
  - How can we predict diabetes based on potential risk factors of diabetes?
- **Purpose**
  - To understand the diabetes risk factors that can help predict whether an individual will acquire diabetes based on those risk factors in the future
- **Why Is it Important?**
  - To raise awareness and potential address prevention strategies for earlier stages before diabetes

Approach: Data

- **Dataset:** "Diabetes Risk Prediction"
  - Diabetes Database: https://www.kaggle.com/datasets/rcratos/diabetes-risk-prediction
  - Categorical Data; Binary

Approach: Data

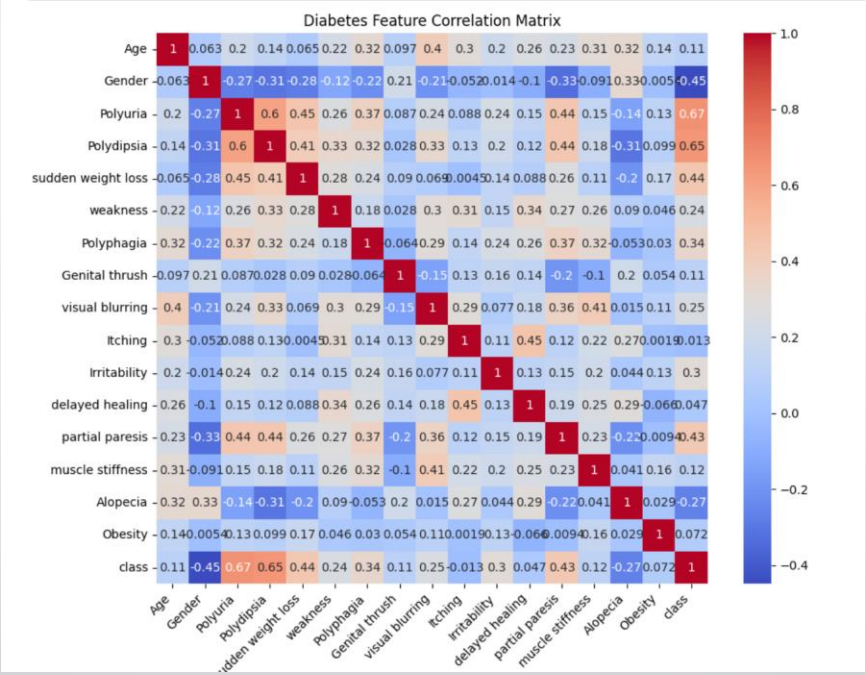
	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
1	58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	39	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	1
516	48	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1
517	58	0	1	1	1	1	1	0	1	0	0	0	1	1	0	1	1
518	32	0	0	0	0	1	0	0	1	1	0	1	0	0	1	0	0
519	42	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

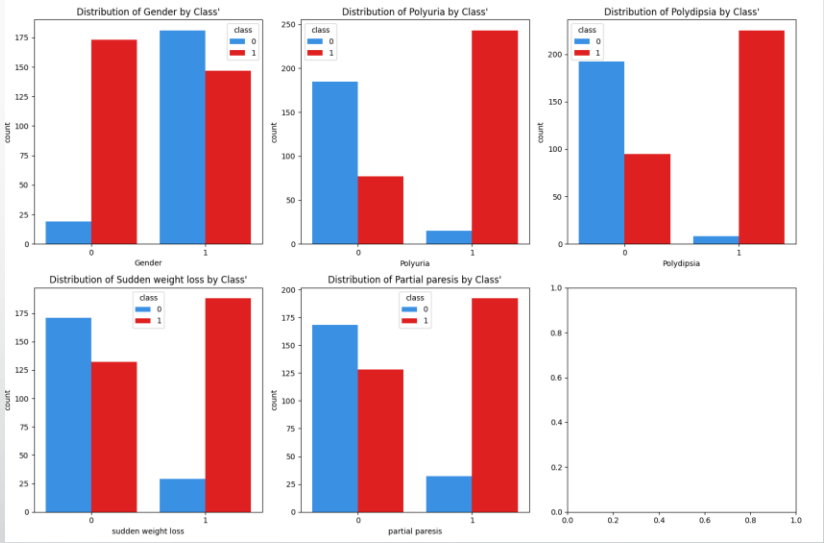
520 rows × 17 columns

Approach: Rundown

- **Data Cleaning :**
  - Simplified the Data
    - Convert labels into binary values (e.g. Yes = 1, No = 0)
    - Removed unnecessary features for prediction
- **Exploratory Data Analysis :**
  - Used Bar Plots to show classifications of diabetes in respect to other features

Approach: Data Cleaning





Approach: Data Cleaning

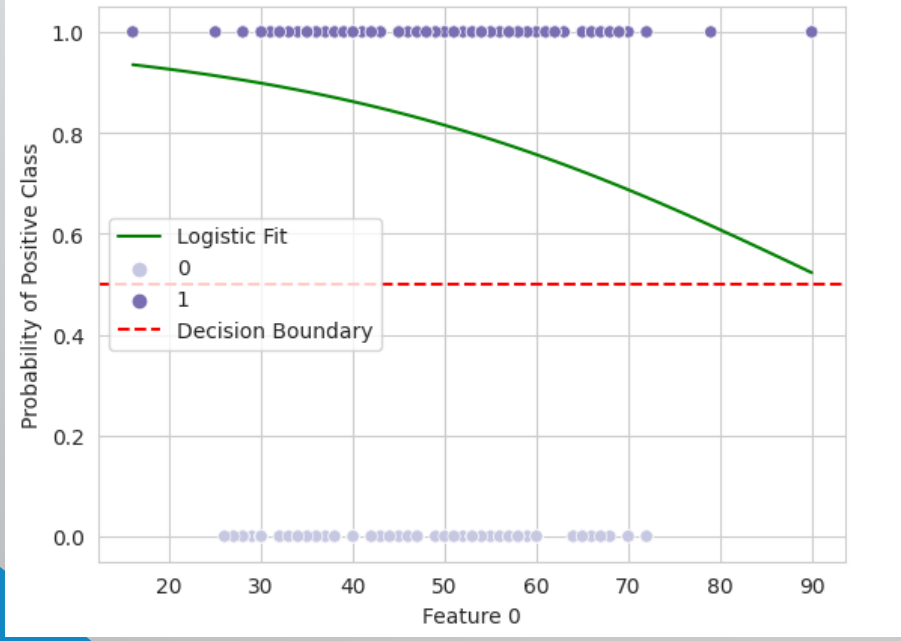
	Age	Gender	Polyuria	Polydipsia	sudden weight loss	partial paresis	class
0	40	1	0	1	0	0	1
1	58	1	0	0	0	1	1
2	41	1	1	0	0	0	1
3	45	1	0	0	1	0	1
4	60	1	1	1	1	1	1

Approach: Models

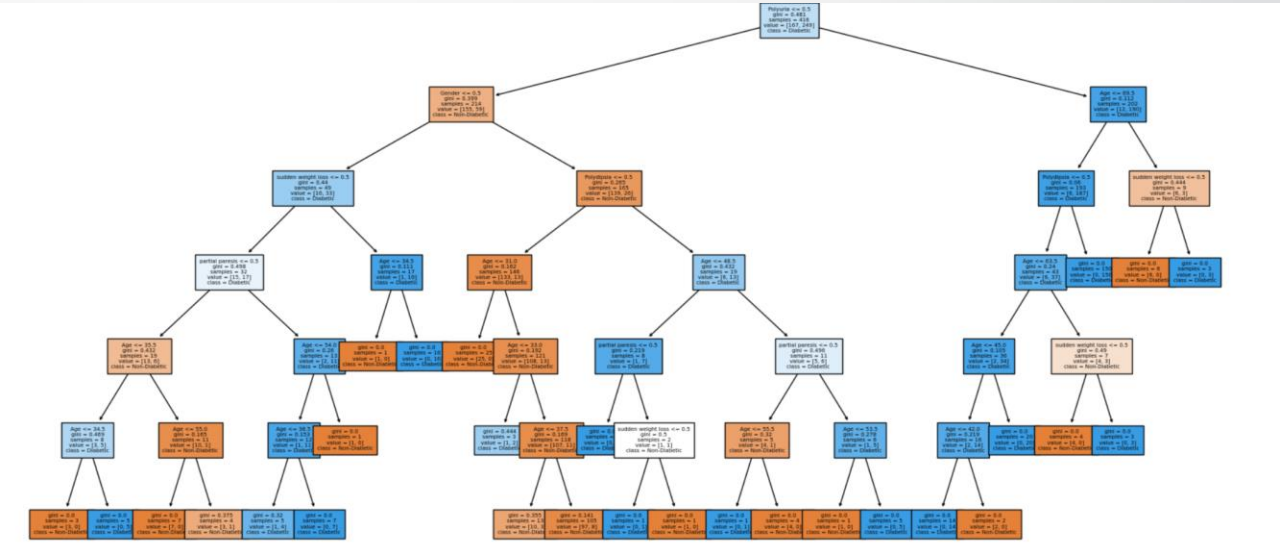
- **Supervised Learning:**
  - Logistic Regression
    - Evaluate and predict whether binary classifications can predict an outcome
  - Decision Tree Classifier
    - Evaluate traits of a classification based on other variables

Approach: Models

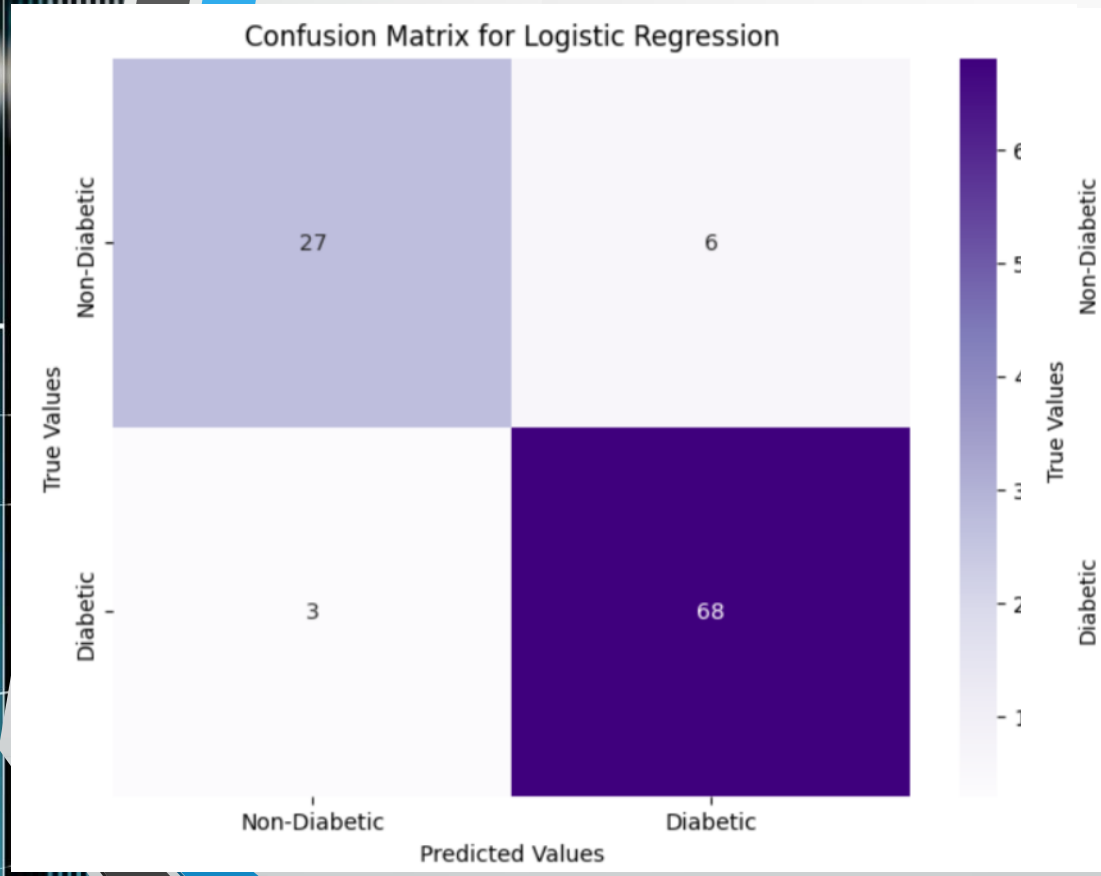
Logistic Regression

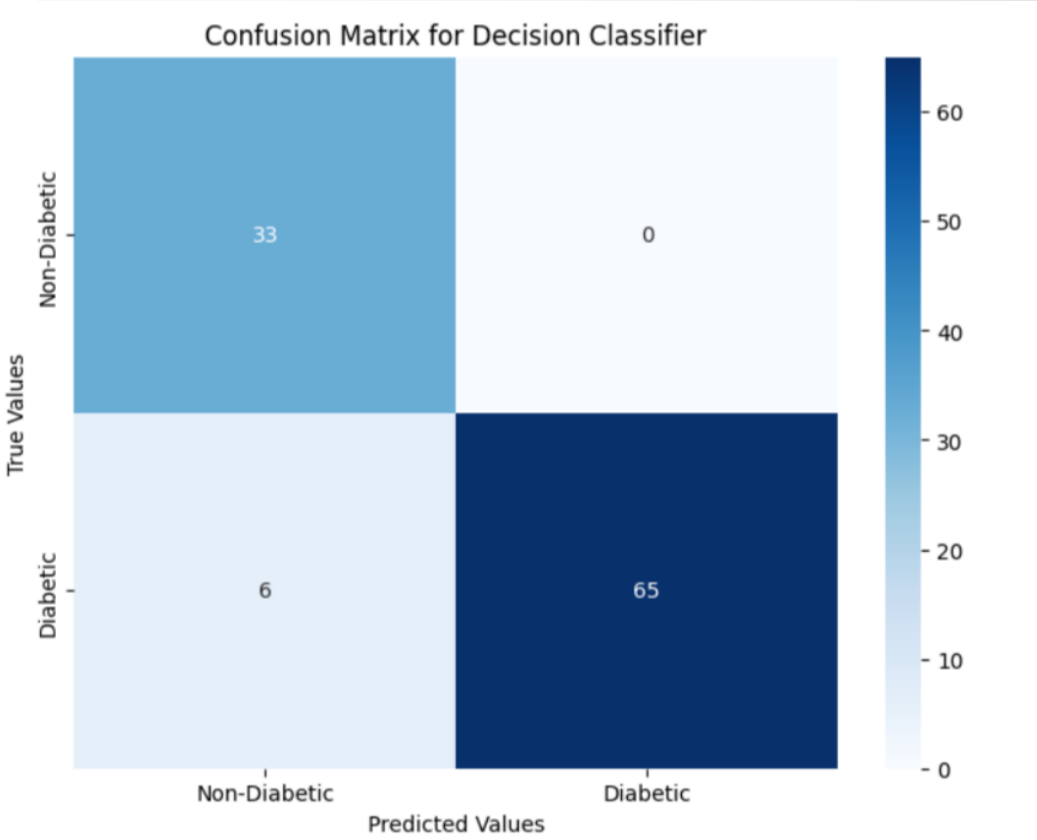


Decision Tree Classifier



Results: Model Analysis






```
Accuracy: 0.9134615384615384
Precision: 0.918918918918919
```

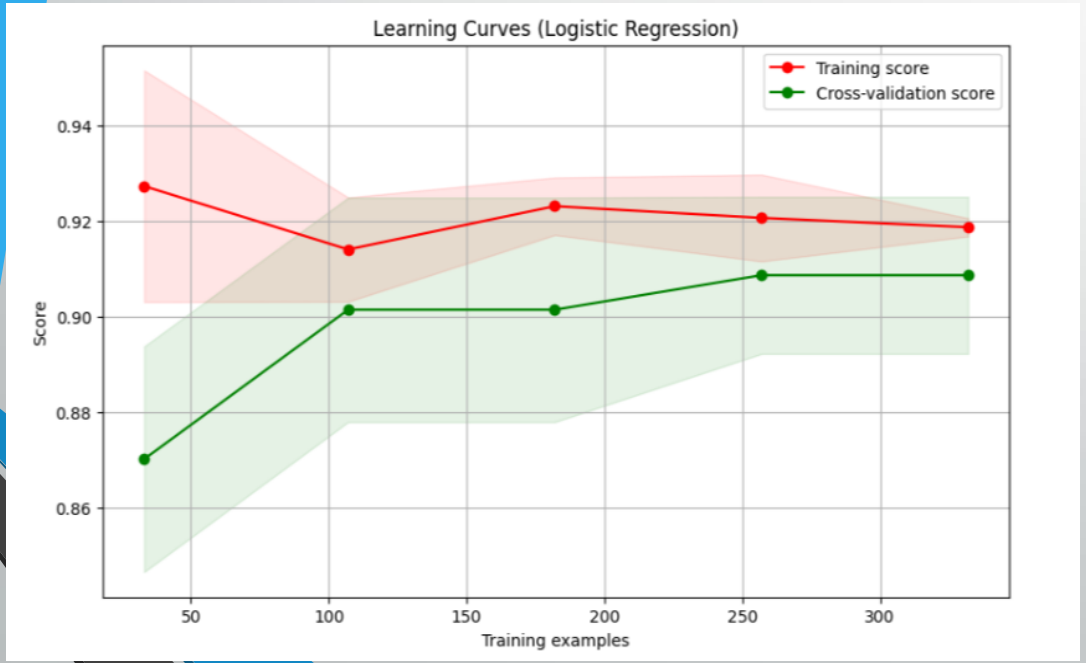
```
Accuracy: 0.9423076923076923
Precision: 1.0
```

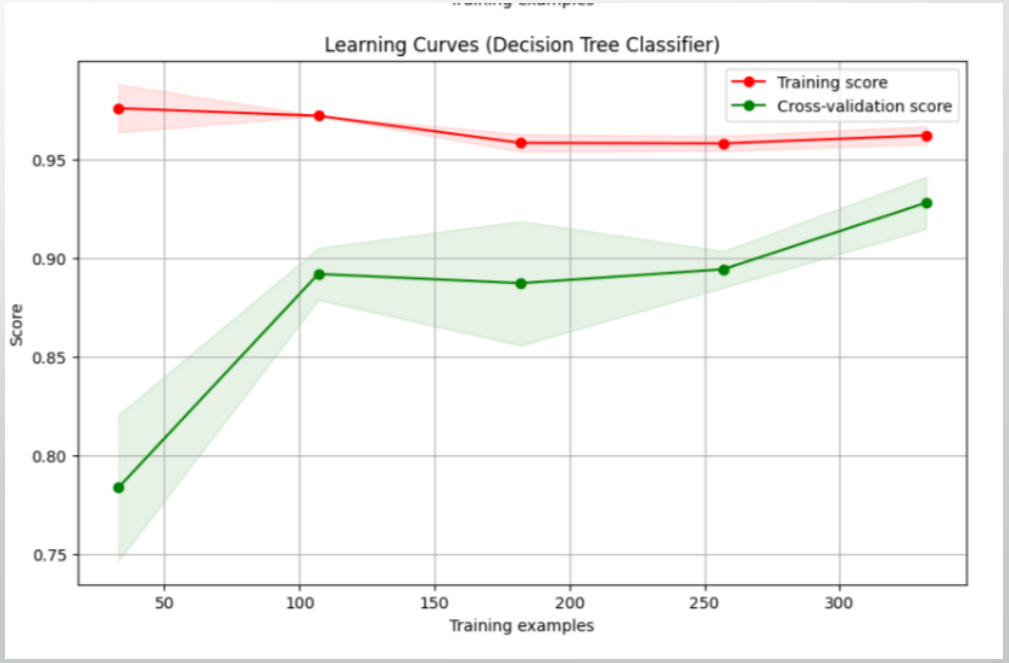
Results

- **Evaluation:**
  - Logistic Regression
    - Cannot deal with too many features
    - The more ideal model with the simplified data
    - High accuracy and precision; pretty reliable predicting model
  - Decision Tree classifier
    - High accuracy and precision too high; might be good or overfitted model
    - Matches the original assumptions from the barplot

Learning Curve

- Evaluate model performance:
  - **Logistic Regression:** Converges, meaning it is performing ideally for supervised learning
  - **Decision Tree Classifier:** Converges but starts diverging, meaning through more training examples, it starts to overfit





Conclusion

- **Logistic Regression** is the best model to predict diabetes, based on diabetes risk factors
- **Decision Tree Classifier** is also a good model but is slightly overfitted.
- Decision Tree Classifier needs changes in either selected features or parameters