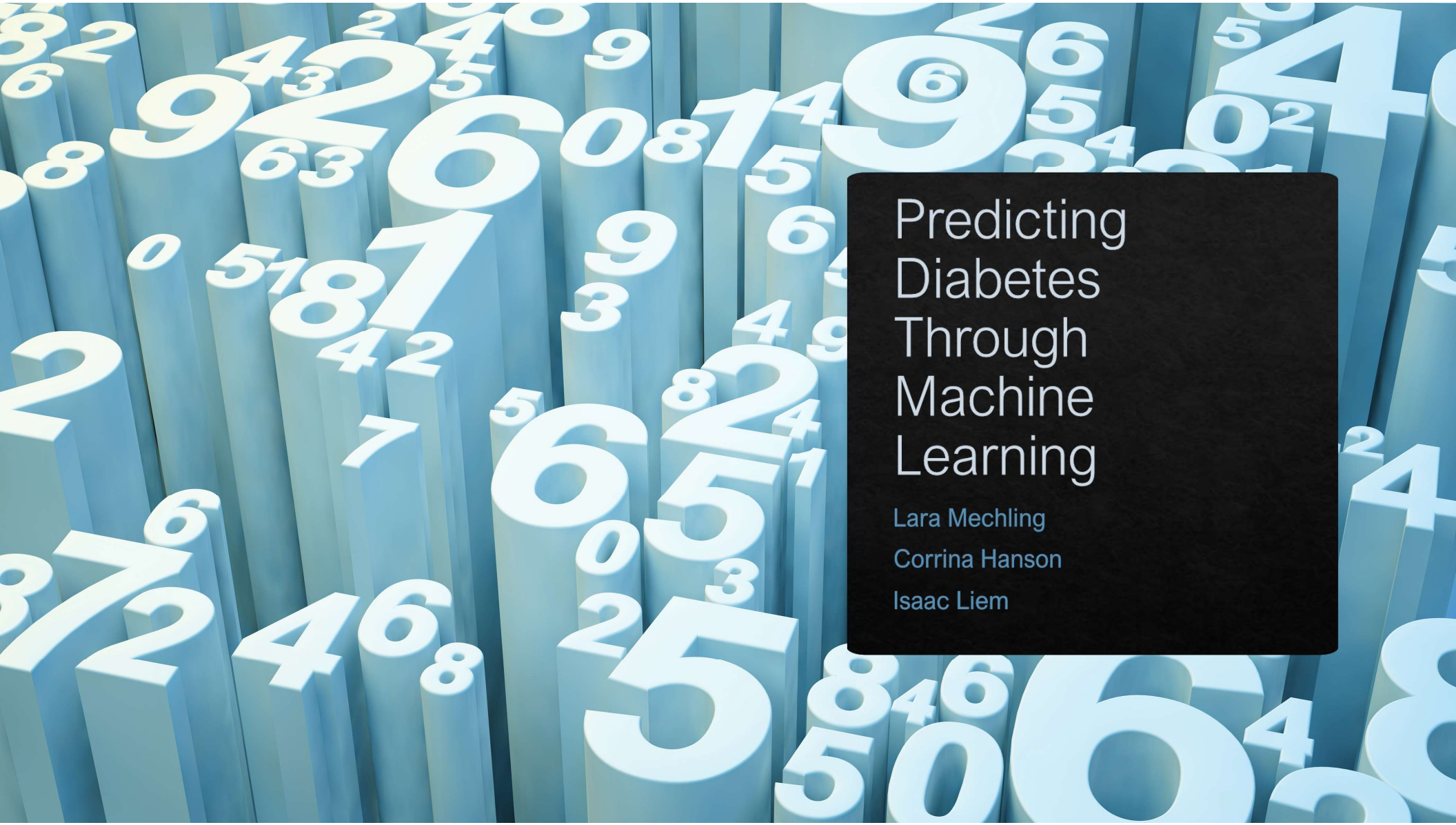


Predicting Diabetes Through Machine Learning

Lara Mechling

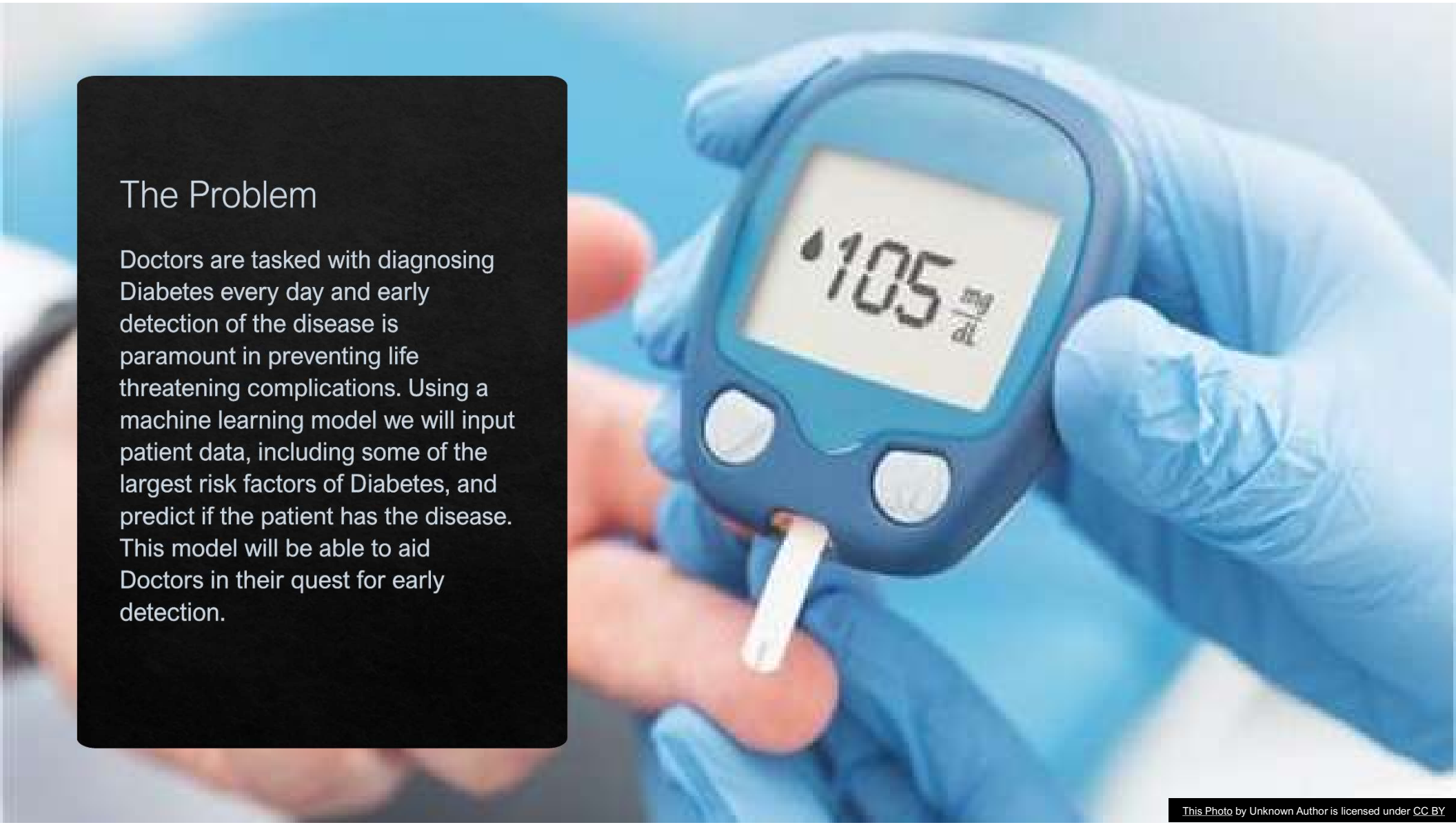
Corrina Hanson

Isaac Liem

“Early detection is key in diabetes because early treatment can prevent serious complications. When a problem with blood sugar is found, doctors and patients can take steps to prevent permanent damage to the heart, kidneys, eyes, nerves, blood vessels, and other vital organs.”

(Falcone, 2020)

- The disease effects over 37 million Americans (CDC, n.d.)
- The disease can cause significant health concerns (CDC, n.d.)



The Problem

Doctors are tasked with diagnosing Diabetes every day and early detection of the disease is paramount in preventing life threatening complications. Using a machine learning model we will input patient data, including some of the largest risk factors of Diabetes, and predict if the patient has the disease. This model will be able to aid Doctors in their quest for early detection.

Database Attributes

Skin Thickness

According to Collier et al. “skin thickness was increased and significantly related to duration of diabetes” (1989)

Age

According to Helmer “age is a big risk factor and an estimated 14% of Americans ages 45 to 64are diagnosed with diabetes which is almost five times the rate for those 18 to 44” (2022)

BMI

“Individuals affected by excess weight, particularly obesity and morbid obesity, are more likely to develop diabetes as a related condition of their excess weight” (Understanding excess weight and its role in type 2 diabetes, n.d.)

Blood Pressure

“High blood pressure is twice as likely to strike a person with diabetes than a person without diabetes" (Johns Hopkins, n.d.)

Other Attributes

Number of Pregnancies

Glucose

Insulin

Diabetes Pedigree Function

The Modeling Process

Data Wrangling

Cleansing

Exploratory Data Analysis

Feature Engineering

Data Modeling

Random Forest Classifier

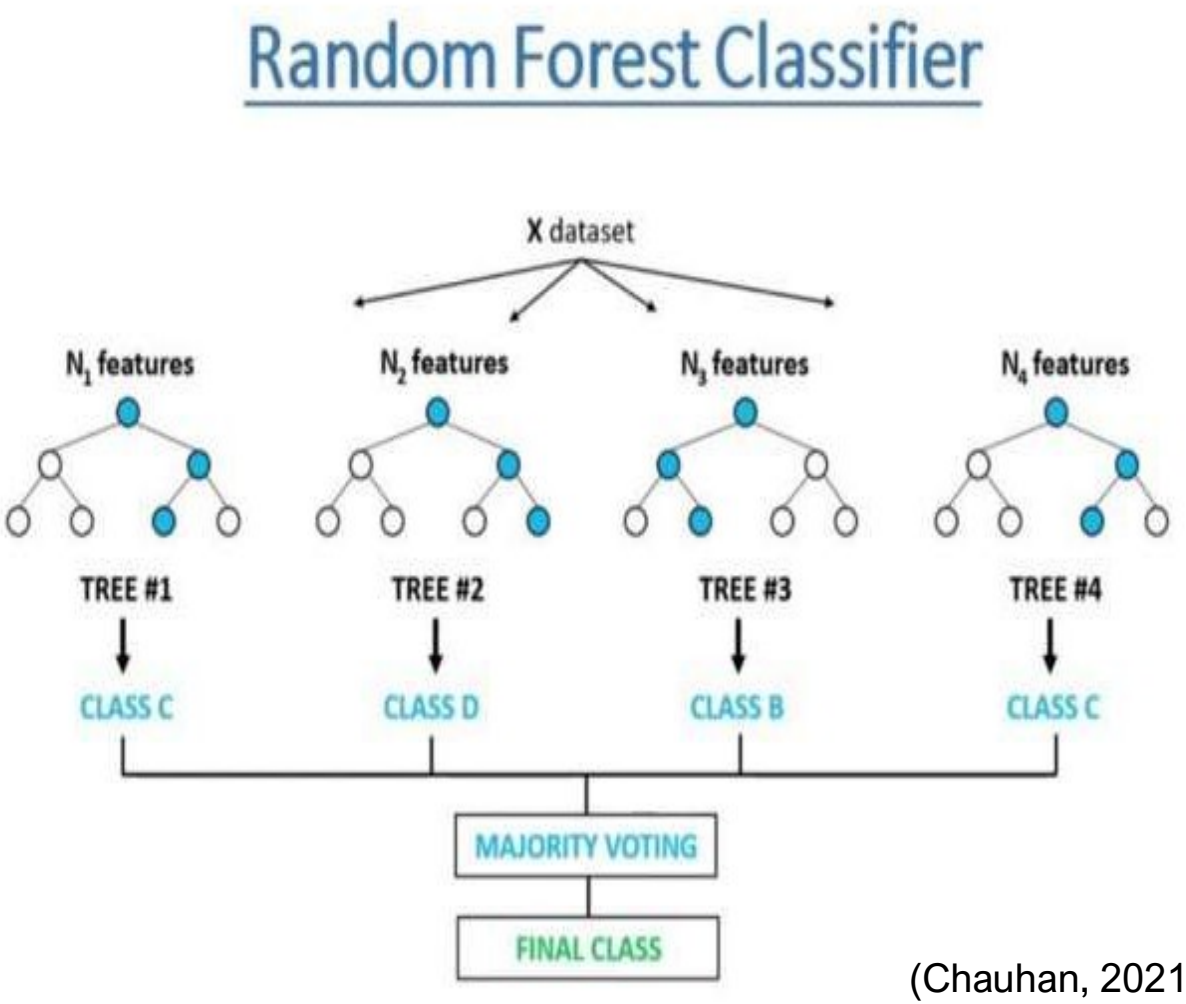
sklearn

Model Validation

Measure Used

Confusion Matrix

72% F1

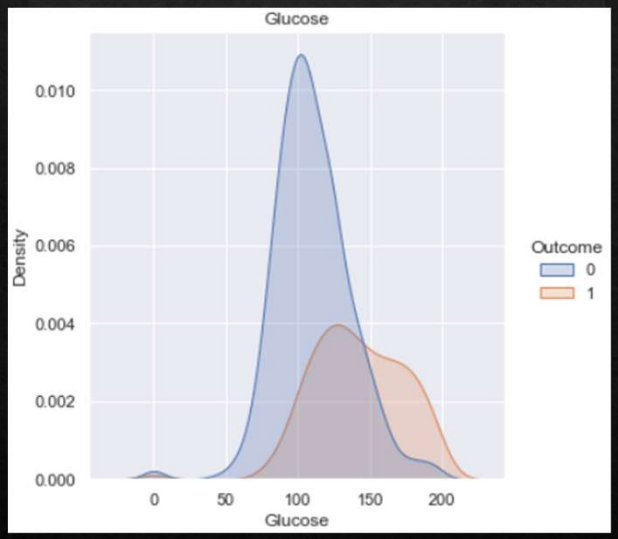


(Chauhan, 2021)

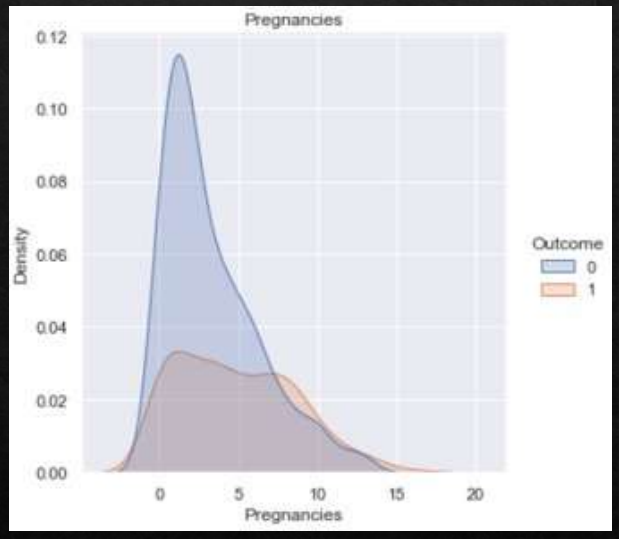
Random Forest Classifier

Random forest classifier (RF) is an ensemble method that uses many decision trees in order to make the final prediction. "RF is a multifunctional machine learning method. It can perform the tasks of prediction and regression. In addition, RF is based on bagging, and it plays an important role in ensemble machine learning" (Zou, et al., 2018).

Variable Probability Density



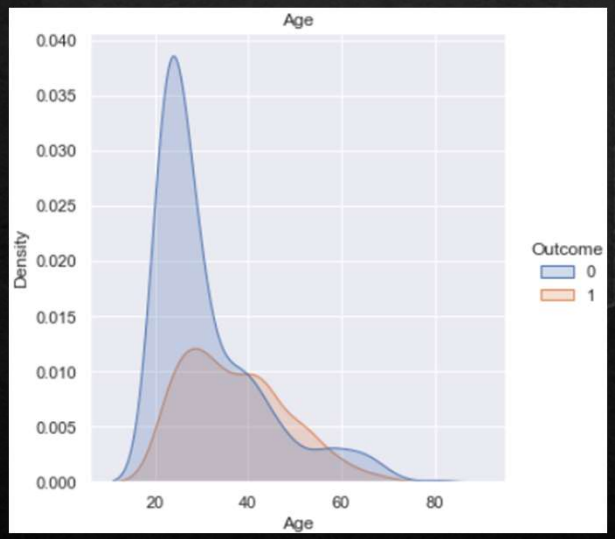
Glucose

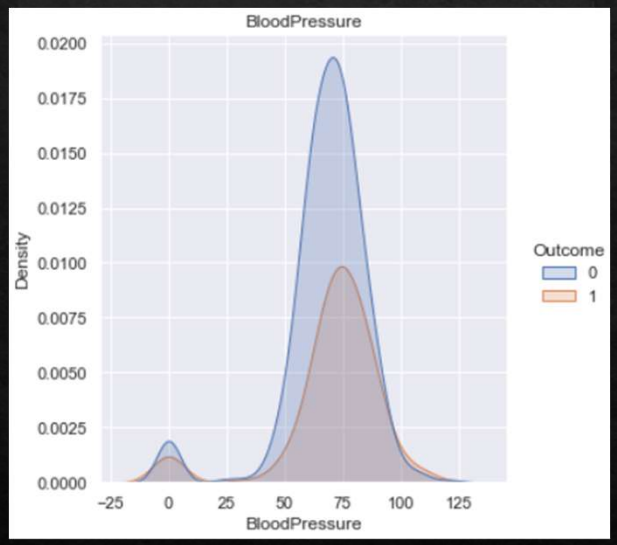


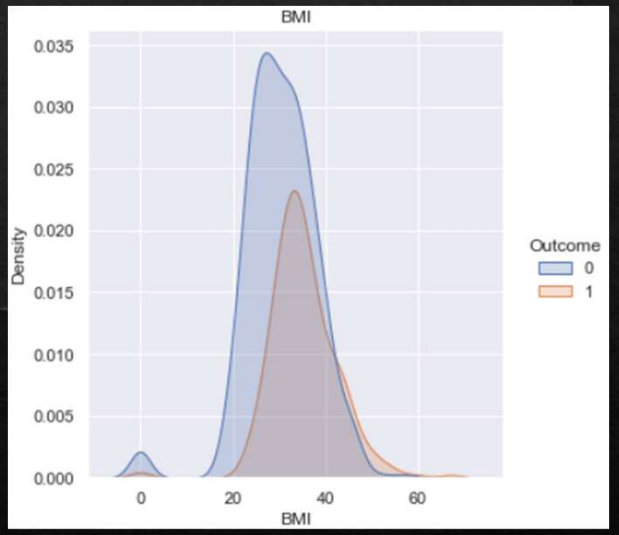
Pregnancies

Several of the attributes have a higher probability for those with Diabetes (1), than those without (0). The orange KDE graphs represent the probability densities of those with the disease. Namely it is more likely for those with diabetes to have had more children, lower insulin, and higher glucose, pedigree function, BMI, age, and blood pressure. The skin thickness is relatively similar for both groups.

Variable Probability Density





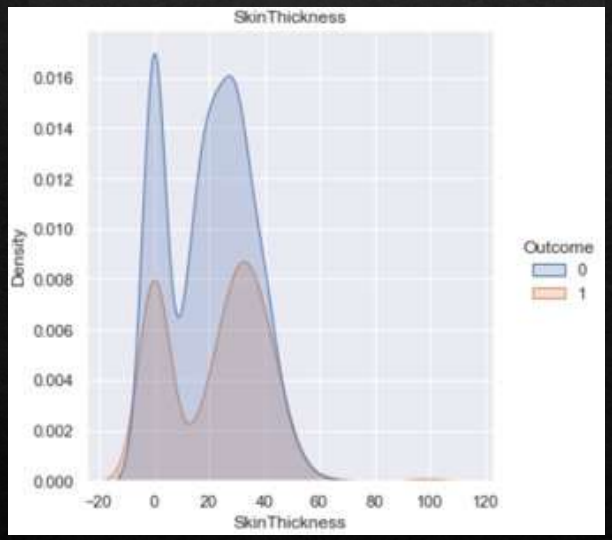


Age

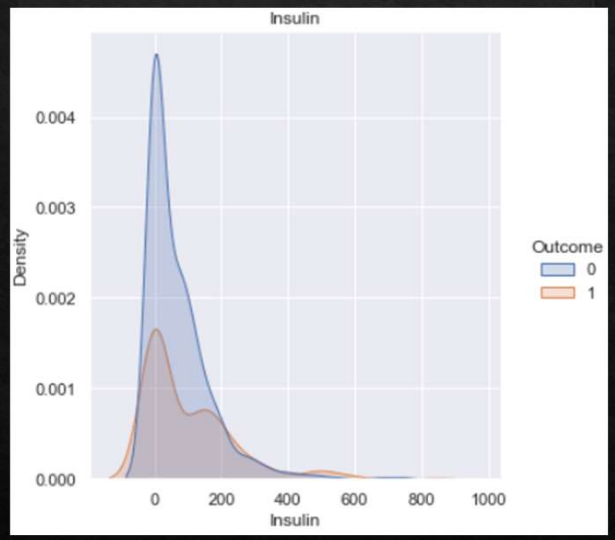
Blood Pressure

BMI

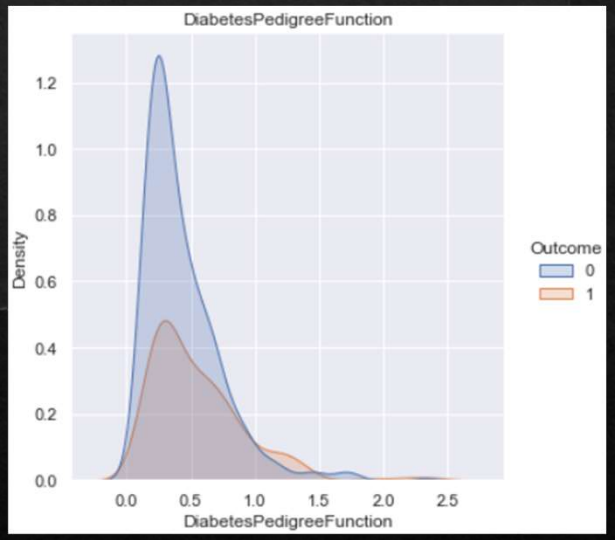
Variable Probability Density



Skin Thickness



Insulin



Diabetes Pedigree Function

Model Metrics

Accuracy Score

77% accuracy in predicting whether patients had diabetes.

Confusion Matrix

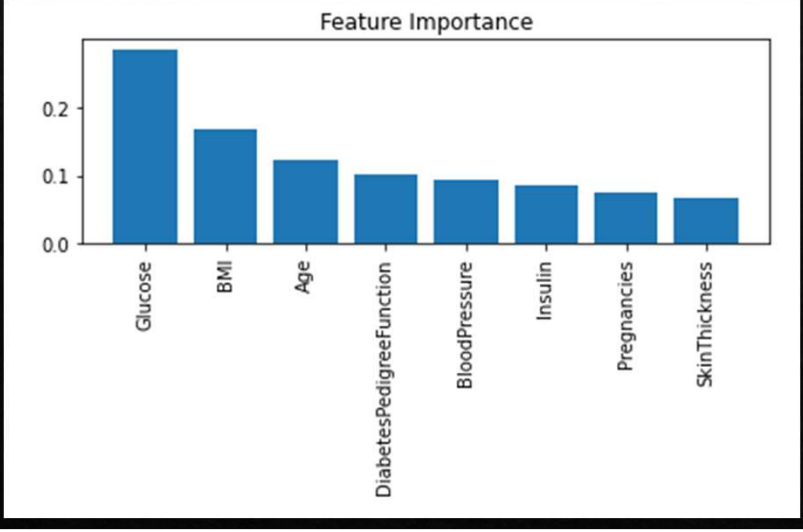
Correctly predicted 130 false cases

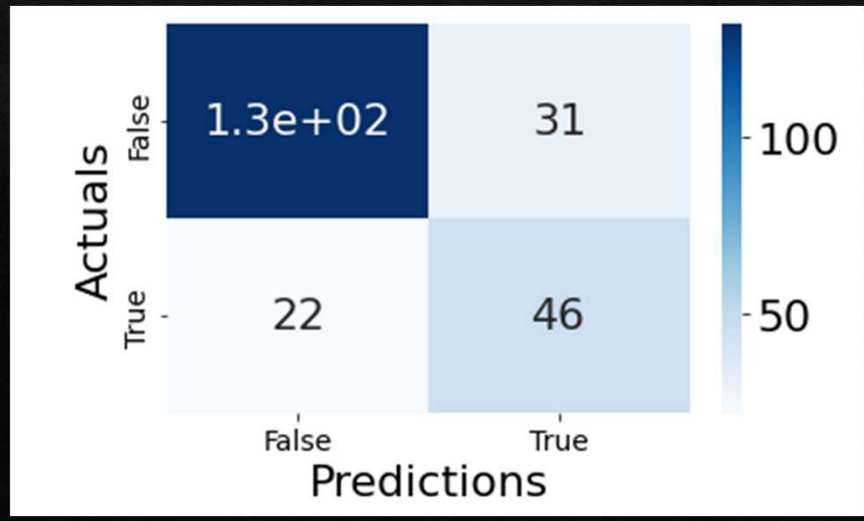
Correctly predicted 46 true cases

Incorrectly predicted 22 false cases

Incorrectly predicted 31 true cases

Feature Importance





Further Learning

Predictive modeling using machine learning techniques has come a long way. Our model had a 77.4% accuracy score, while not a bad accuracy score for a first attempt at a model to predict diabetes, can likely be refined.

Areas that need more research include selecting the correct attributes which best aid in doctor's abilities to correctly identify whether a patient is diabetic or not, as well as further looking at machine learning models and fine tuning them to get the best result.

Larger and more robust datasets could further enable to machine learning model to be more precise and take in more data to improve its accuracy.

The accuracy score of 77% "can indicate machine learning can be used for predicting diabetes, but finding suitable attributes, classifier and data mining methods are very important" (Zou, et al., 2018).

References

CDC. (n.d.). *The Facts, Stats, and Impacts of Diabetes*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/diabetes/library/spotlights/diabetes-facts-stats.html#:~:text=37.3%20million%20Americans%E2%80%94about%201,t%20know%20they%20have%20it.

Chauhan, A. (2021, Feb 23). *Random Forest Classifier and it's Hyperparameters*. Retrieved from Towards Data Science: https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6

Collier, A., Patrick, A. W., Bell, D., Matthews, D. M., Macintyre, C. C., Eing, D. J., & Clarke, B. F. (1989). Relationship of skin thickness to duration of diabetes, glycemic control, and diabetic complications in male IDDM patients. *National Library of Medicine*, 309 - 312.

Falcone, S. (2020, November 30). *Why Early Detection is Key in Diabetes*. Retrieved from My Virtual Physician: https://myvirtualphysician.com/2020/11/30/why-early-detection-is-key-in-diabetes/#:~:text=Early%20detection%20is%20key%20in%20diabetes%20because%20early%20treatment%20can,vessels%2C%20and%20other%20vital%20organs

Helmer, J. (2022, April 9). *How Age Relates to Type 2 Diabetes*. Retrieved from Web MD: https://www.webmd.com/diabetes/diabetes-link-age#091e9c5e81edf172-2-6

Johns Hopkins. (n.d.). *Diabetes and High Blood Pressure*. Retrieved from Johns Hopkins Medicine: https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure

Understanding excess weight and its role in type 2 diabetes. (n.d.). Retrieved from Honor Health: https://www.honorhealth.com/medical-services/bariatric-weight-loss-surgery/patient-education-and-support/comorbidities-type-2-diabetes#:~:text=Being%20overweight%20(BMI%20of%2025,to%20your%20own%20insulin%20hormone

Zou, Q., Qu, K., Lou, Y., Yin, D., Ju, Y., & Tang, H. (2018, November 6). Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontier Genetics*. Retrieved from https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full