Overview

Diabetes is a common condition. The CDC found that 30.3 million adults in the U.S. have diabetes. They also estimate that 80.4 million adults have prediabetes. Diabetes can be difficult to live with and be expensive to treat. Therefore, diagnosing diabetes early can be helpful to a person's way of life and wallet.

Data

A dataset of people with diabetic symptoms was found on Kaggle. It was put there by the UC Irvine (UCI) Center for Machine Learning and Intelligent Systems. The data contains information collected from questionnaires from newly diabetic or would be diabetic patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. There are 520 instances (patients) with 15 possible features (symptoms) and their class (whether they have or do not have diabetes). 320 instances are Positive for diabetes and 200 are Negative for diabetes.

Data Wrangling

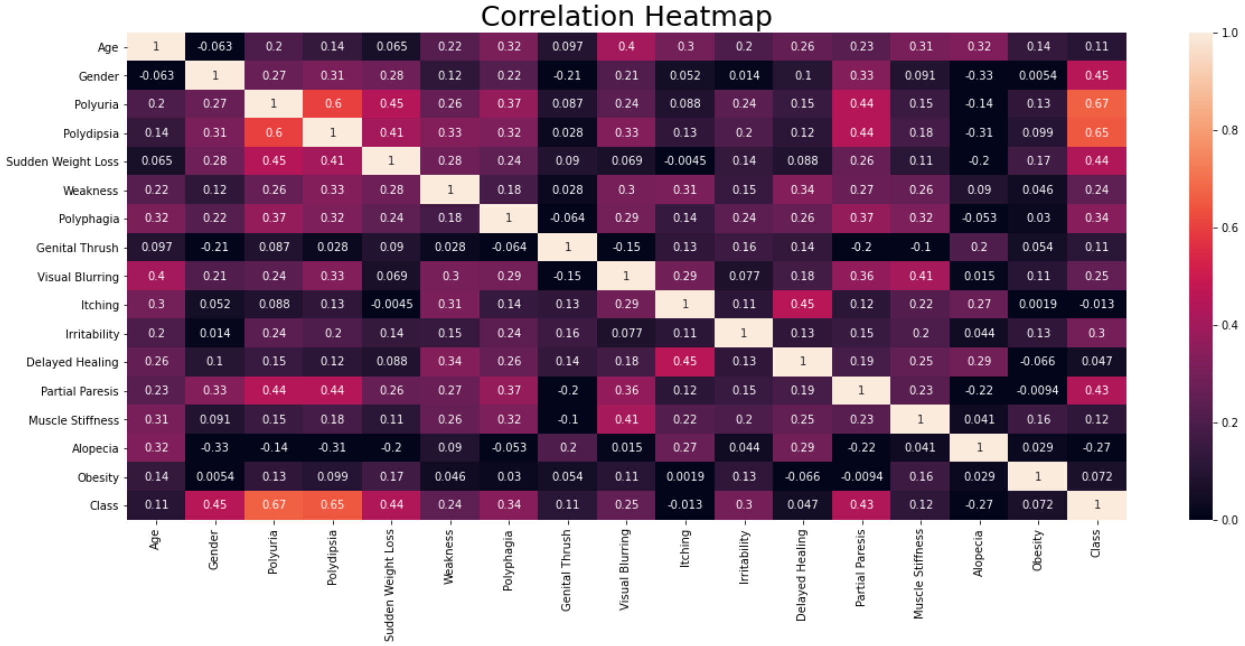
The dataset did not have any missing values.

Some of the analysis requires the data to be integers rather than text. Therefore, a dataframe was created changing the literals ‘Yes/No’, ‘Female/Male’, and ‘Positive/Negative’ to the integers 1/0, respectively.

Exploratory Data Analysis

Correlation Heatmap

A correlation heatmap was created to see what features (symptoms) relate to the class (whether they have or do not have diabetes). Only two features had more than a 50% correlation to the class (whether they have or do not have diabetes). Polyuria has a 67% correlation and Polydipsia has a 65% correlation. That is not too good by itself so further analysis is needed.



	Age	Gender	Polyuria	Polydipsia	Sudden Weight Loss	Weakness	Polyphagia	Genital Thrush	Visual Blurring	Itching	Irritability	Delayed Healing	Partial Paresis	Muscle Stiffness	Alopecia	Obesity	Class
Age	1	-0.063	0.2	0.14	0.065	0.22	0.32	0.097	0.4	0.3	0.2	0.26	0.23	0.31	0.32	0.14	0.11
Gender	-0.063	1	0.27	0.31	0.28	0.12	0.22	-0.21	0.21	0.052	0.014	0.1	0.33	0.091	-0.33	0.0054	0.45
Polyuria	0.2	0.27	1	0.6	0.45	0.26	0.37	0.087	0.24	0.088	0.24	0.15	0.44	0.15	-0.14	0.13	0.67
Polydipsia	0.14	0.31	0.6	1	0.41	0.33	0.32	0.028	0.33	0.13	0.2	0.12	0.44	0.18	-0.31	0.099	0.65
Sudden Weight Loss	0.065	0.28	0.45	0.41	1	0.28	0.24	0.09	0.069	-0.0045	0.14	0.088	0.26	0.11	-0.2	0.17	0.44
Weakness	0.22	0.12	0.26	0.33	0.28	1	0.18	0.028	0.3	0.31	0.15	0.34	0.27	0.26	0.09	0.046	0.24
Polyphagia	0.32	0.22	0.37	0.32	0.24	0.18	1	-0.064	0.29	0.14	0.24	0.26	0.37	0.32	-0.053	0.03	0.34
Genital Thrush	0.097	-0.21	0.087	0.028	0.09	0.028	-0.064	1	-0.15	0.13	0.16	0.14	-0.2	-0.1	0.2	0.054	0.11
Visual Blurring	0.4	0.21	0.24	0.33	0.069	0.3	0.29	-0.15	1	0.29	0.077	0.18	0.36	0.41	0.015	0.11	0.25
Itching	0.3	0.052	0.088	0.13	-0.0045	0.31	0.14	0.13	0.29	1	0.11	0.45	0.12	0.22	0.27	0.0019	-0.013
Irritability	0.2	0.014	0.24	0.2	0.14	0.15	0.24	0.16	0.077	0.11	1	0.13	0.15	0.2	0.044	0.13	0.3
Delayed Healing	0.26	0.1	0.15	0.12	0.088	0.34	0.26	0.14	0.18	0.45	0.13	1	0.19	0.25	0.29	-0.066	0.047
Partial Paresis	0.23	0.33	0.44	0.44	0.26	0.27	0.37	-0.2	0.36	0.12	0.15	0.19	1	0.23	-0.22	-0.0094	0.43
Muscle Stiffness	0.31	0.091	0.15	0.18	0.11	0.26	0.32	-0.1	0.41	0.22	0.2	0.25	0.23	1	0.041	0.16	0.12
Alopecia	0.32	-0.33	-0.14	-0.31	-0.2	0.09	-0.053	0.2	0.015	0.27	0.044	0.29	-0.22	0.041	1	0.029	-0.27
Obesity	0.14	0.0054	0.13	0.099	0.17	0.046	0.03	0.054	0.11	0.0019	0.13	-0.066	-0.0094	0.16	0.029	1	0.072
Class	0.11	0.45	0.67	0.65	0.44	0.24	0.34	0.11	0.25	-0.013	0.3	0.047	0.43	0.12	-0.27	0.072	1

Feature Analysis

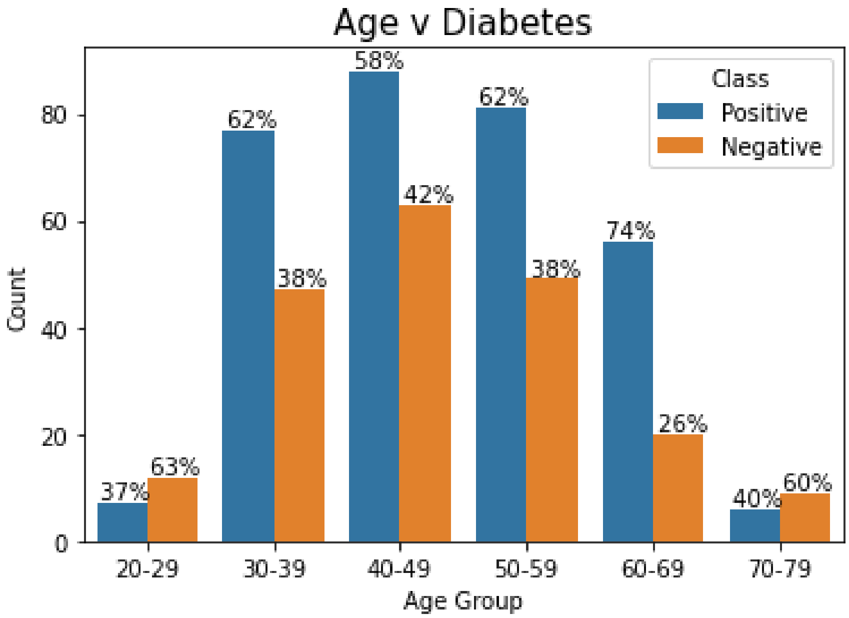
Each of the features was analyzed with respect to whether a patient had diabetes or not.

Positive v Negative

There are 320 instances of patients with diabetes (Positive) and 200 instances of patients without diabetes (Negative).

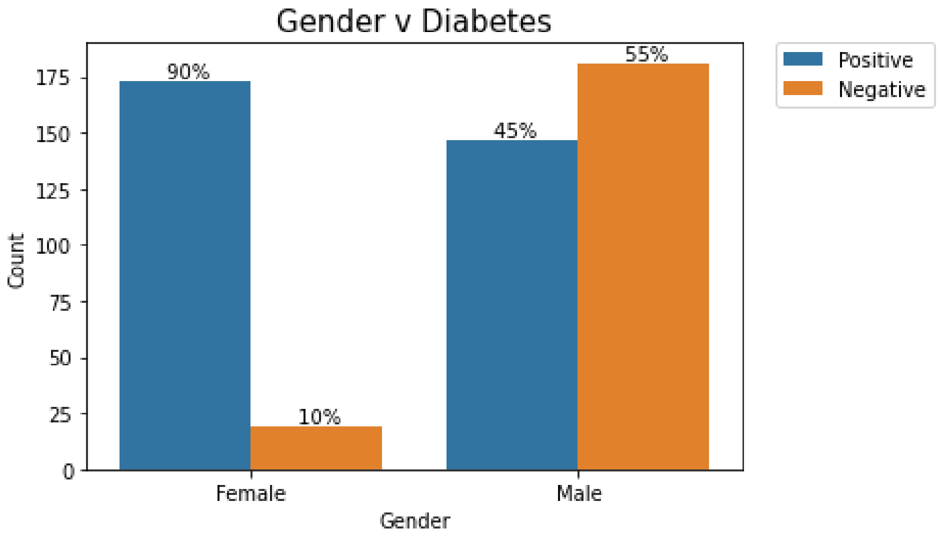
Age

Patients in their 60's had the highest chance of diabetes at 74%. Patients in their 30's and 50's had a 62% chance of having diabetes and patients in their 40's had a 58% chance of having diabetes. Patients in their 20's and 70's had less than a 50% chance of having diabetes.



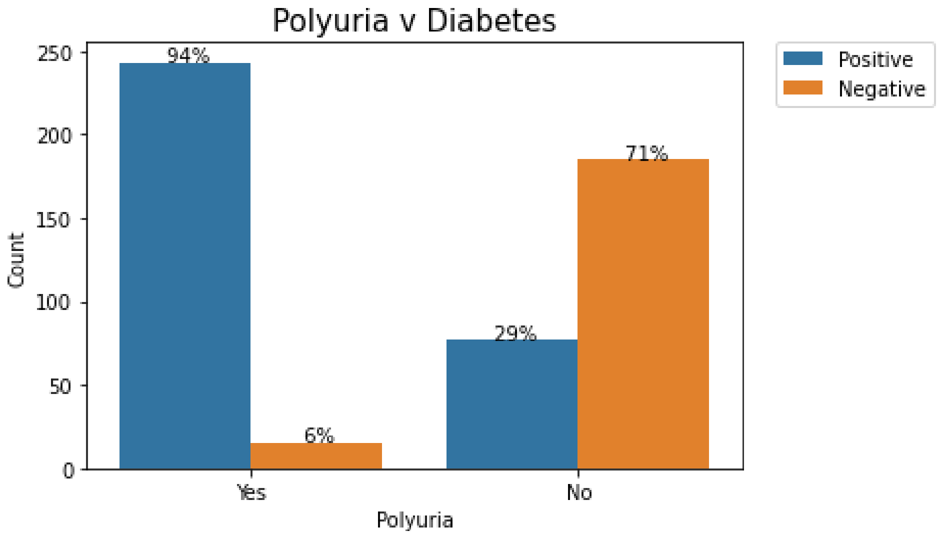
Gender

Females had a 90% chance of having diabetes and males had a 45% chance of having diabetes.



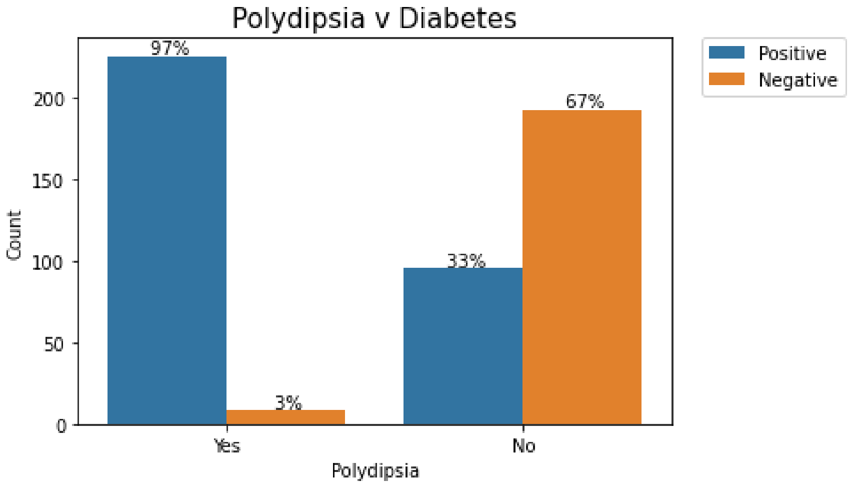
Polyuria

Patients with Polyuria had a 94% chance of having diabetes.



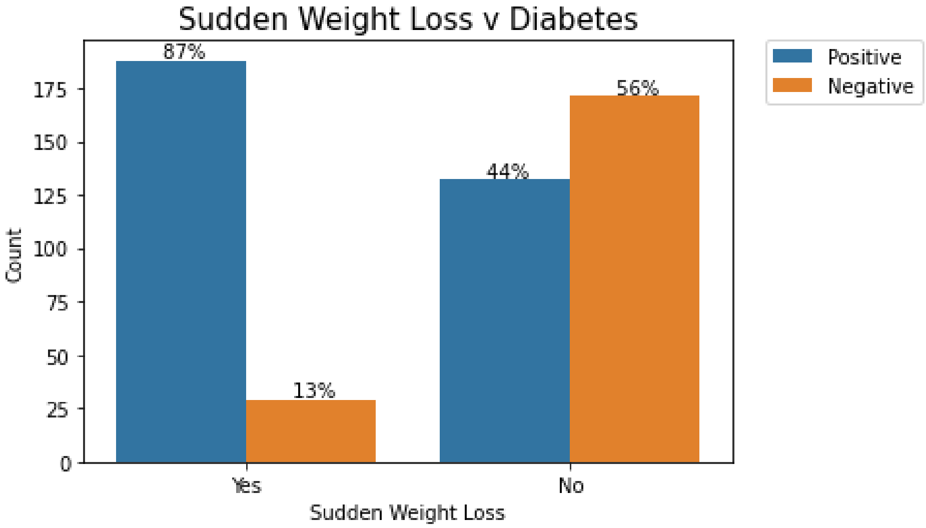
Polydipsia

Patients with Polydipsia had a 97% chance of having diabetes.



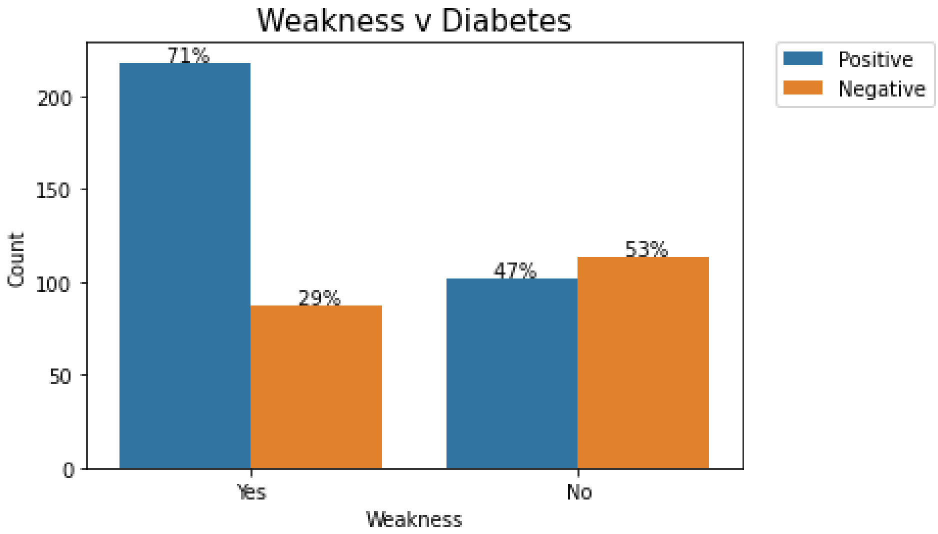
Sudden Weight Loss

Patients with Sudden Weight Loss had a 87% chance of having diabetes.



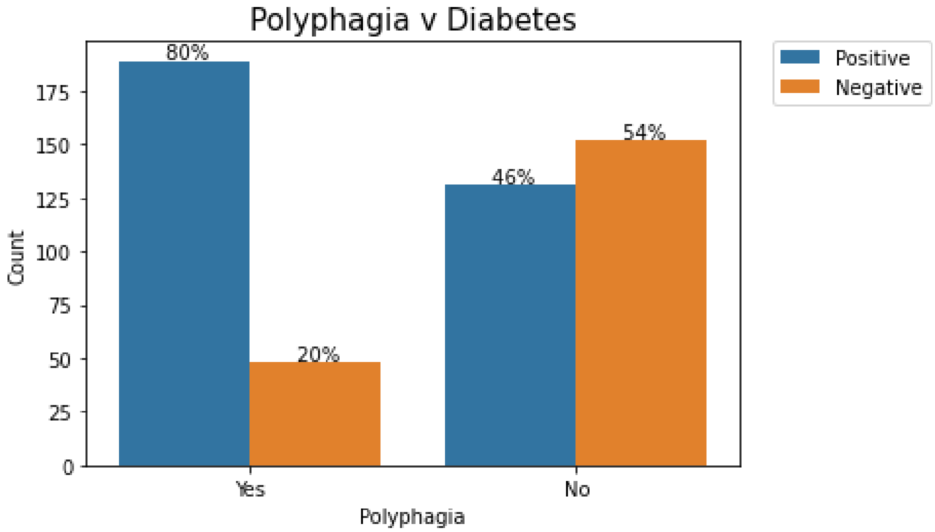
Weakness

Patients with Weakness had a 71% chance of having diabetes.



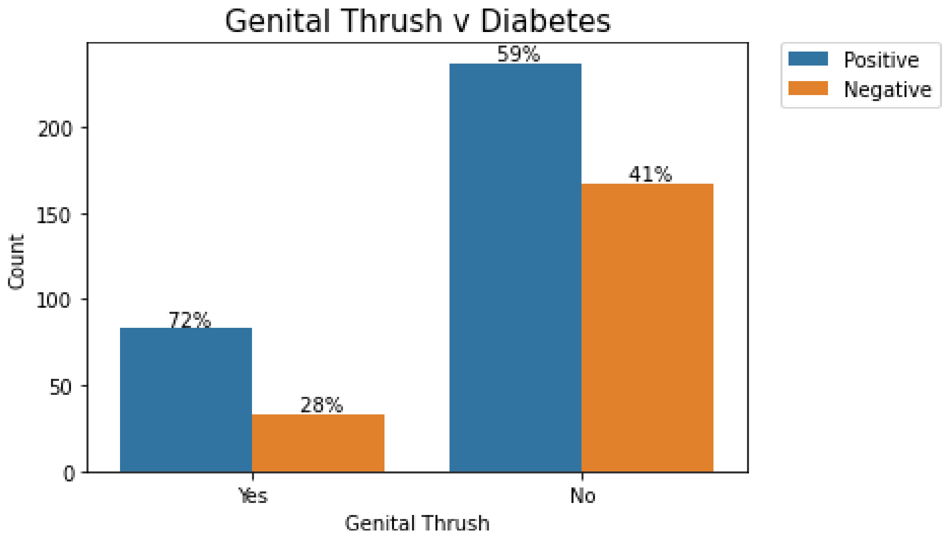
Polyphagia

Patients with Polyphagia had a 80% chance of having diabetes.



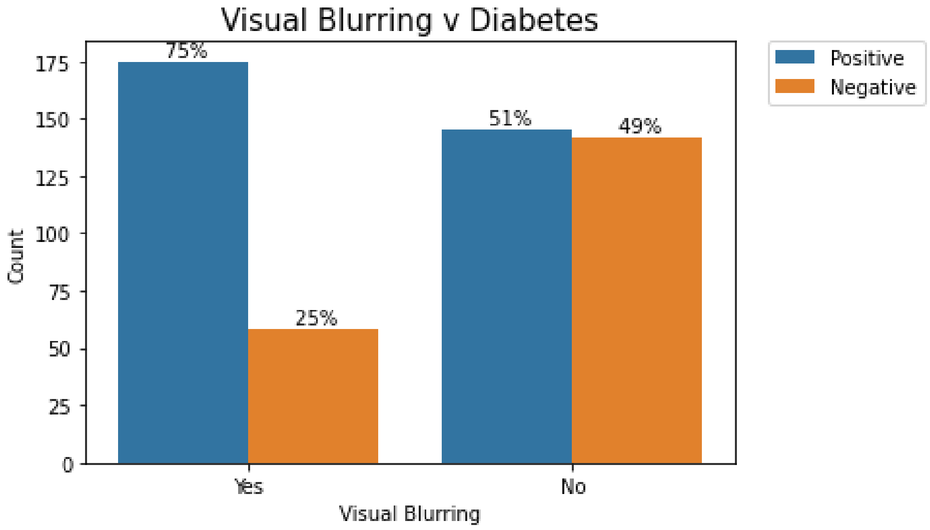
Genital Thrush

Patients with Genital Thrush had a 72% chance of having diabetes.



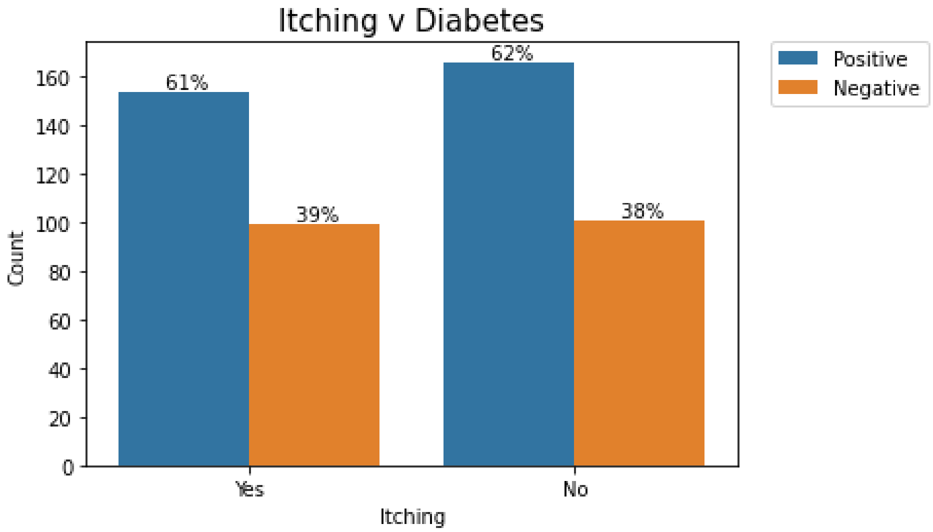
Visual Blurring

Patients with Visual Blurring had a 75% chance of having diabetes.



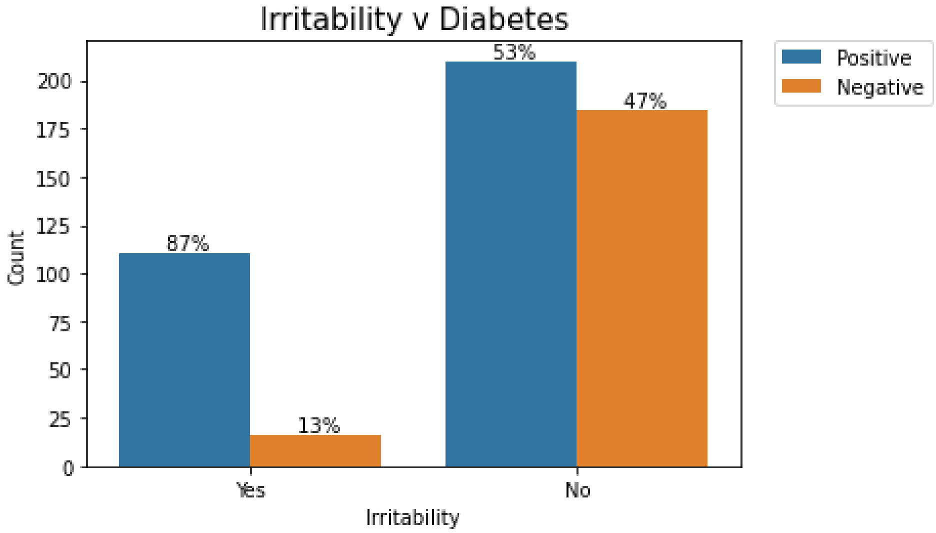
Itching

Patients with Itching had a 61% chance of having diabetes.



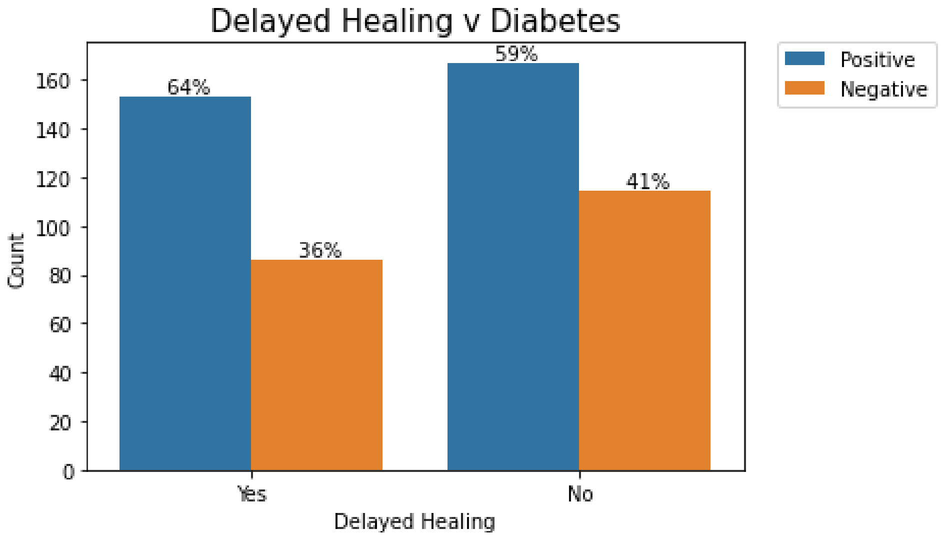
Irritability

Patients with Irritability had a 87% chance of having diabetes.



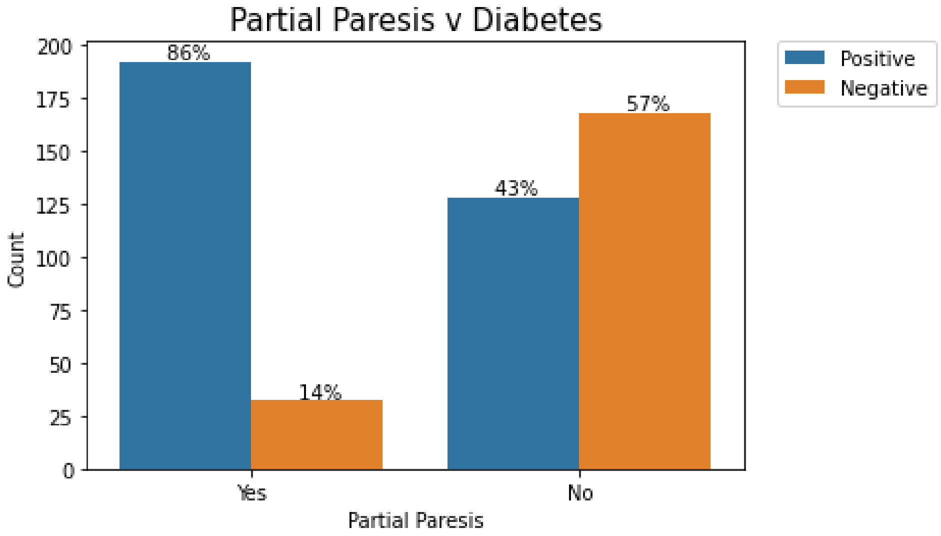
Delayed Healing

Patients with Delayed Healing had a 64% chance of having diabetes.



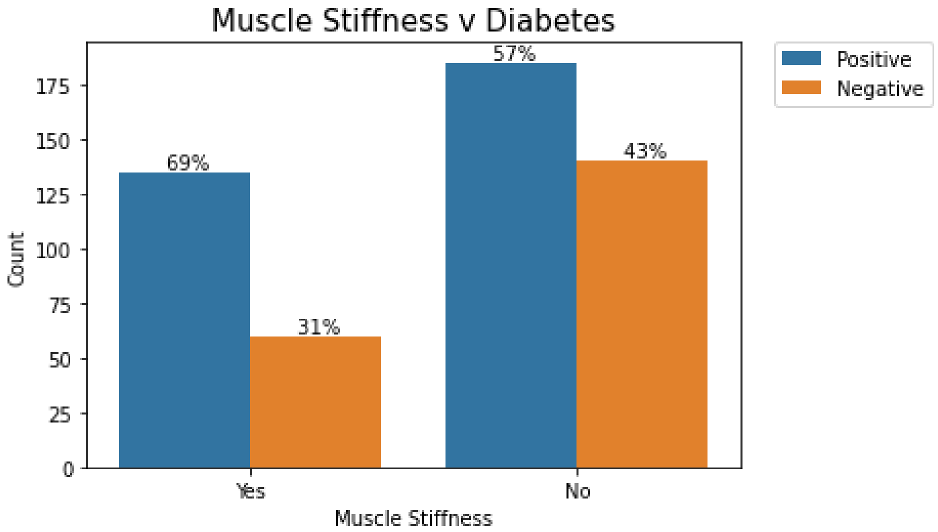
Partial Paresis

Patients with Partial Paresis had a 86% chance of having diabetes.



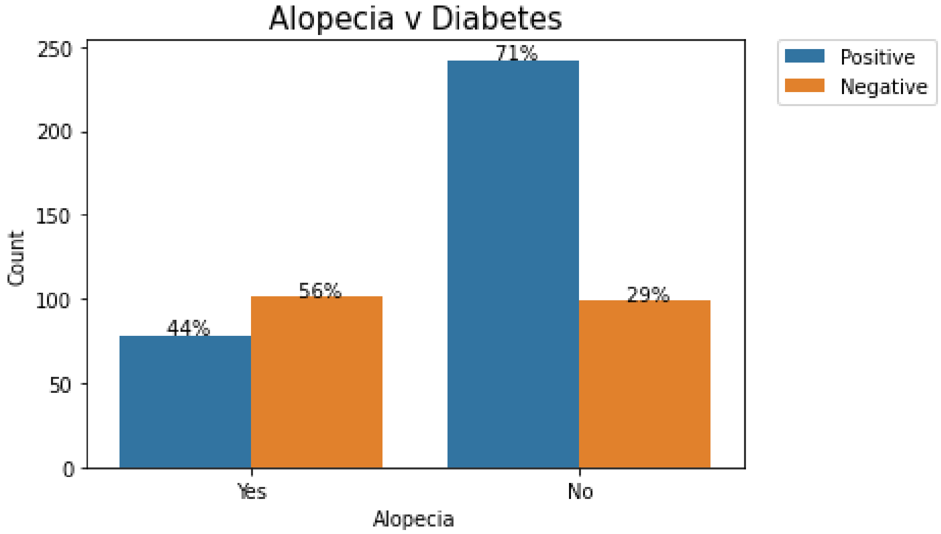
Muscle Stiffness

Patients with Muscle Stiffness had a 69% chance of having diabetes.



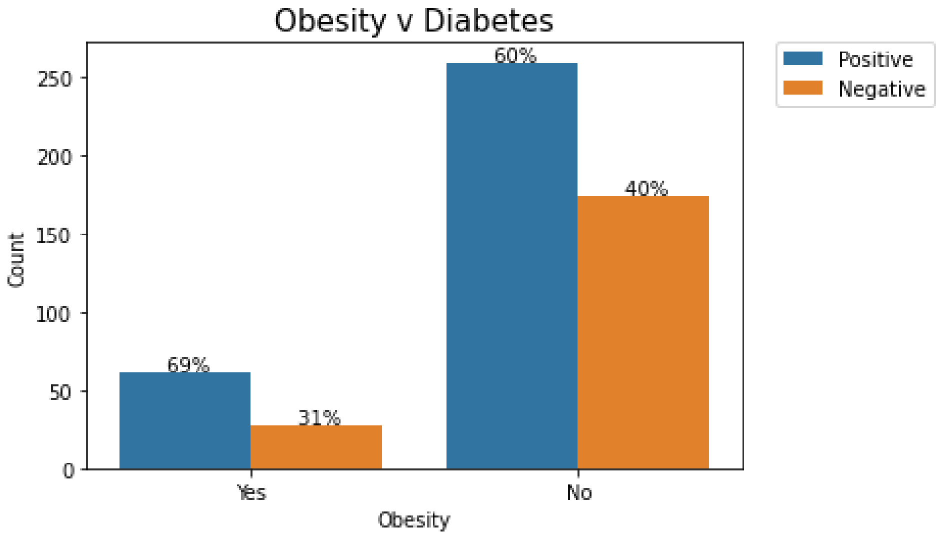
Alopecia

Patients with Alopecia had a 44% chance of having diabetes.



Obesity

Patients with Obesity had a 69% chance of having diabetes.

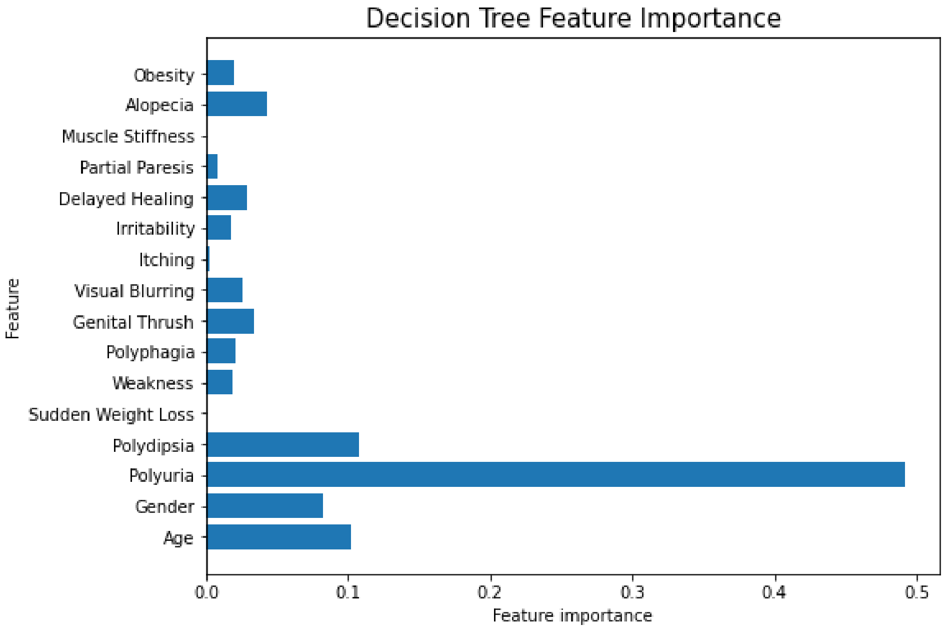


Model Analysis

Logistic Regression with Scaling
Accuracy on training set: 0.940
Accuracy on test set: 0.929

Note: The data needed to be scaled in order to do Logistic Regression

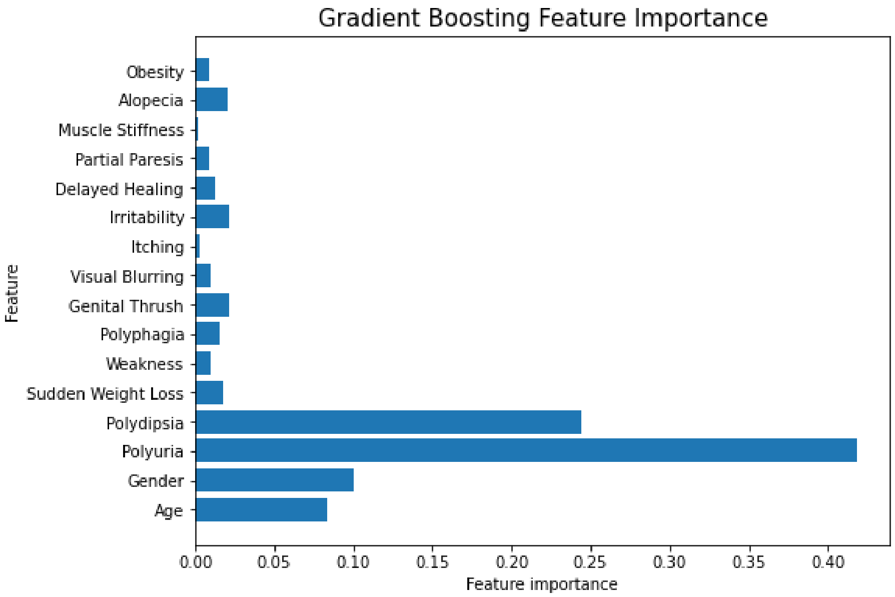
Decision Tree
Accuracy on training set: 0.995
Accuracy on test set: 0.936



Gradient Boosting

Accuracy on training set: 0.997

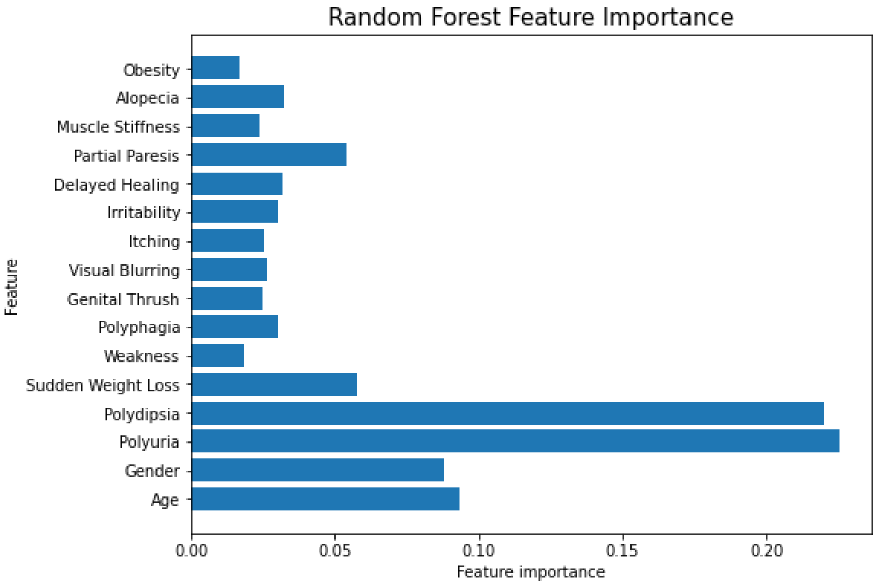
Accuracy on test set: 0.949



Random Forest

Accuracy on training set: 1.000

Accuracy on test set: 0.981



Summary of Results **(greater than 40%)**

Correlation Heatmap	Polyuria	67%
Correlation Heatmap	Polydipsia	65%
Correlation Heatmap	Gender	45%
Correlation Heatmap	Sudden Weight Loss	44%
Correlation Heatmap	Partial Paresis	43%
Feature Analysis	Age – 60’s	74%
Feature Analysis	Age – 30’s, 50’s	62%
Feature Analysis	Age – 40’s	58%
Feature Analysis	Age – 70’s	40%
Feature Analysis	Gender – Female	90%
Feature Analysis	Gender – Male	45%
Feature Analysis	Polyuria	94%
Feature Analysis	Polydipsia	97%
Feature Analysis	Sudden Weight Loss	87%
Feature Analysis	Weakness	71%
Feature Analysis	Polyphagia	80%
Feature Analysis	Genital Thrush	72%
Feature Analysis	Visual Blurring	75%
Feature Analysis	Itching	61%
Feature Analysis	Irritability	87%
Feature Analysis	Delayed Healing	86%
Feature Analysis	Partial Paresis	86%
Feature Analysis	Muscle Stiffness	69%
Feature Analysis	Alopecia	44%
Feature Analysis	Obesity	69%
Model	Logistic Regression with Scaling	93%
Model	Decision Tree	94%
Model	Gradient Boosting	95%
Model	Random Forest	98%

Conclusion

The Correlation Heatmap gave us a good initial idea about what features may be useful. 5 features gave 40% - 67% correlation to determine whether a person has diabetes. 67% correlation is not very high.

Feature Analysis gave us 20 features with 40% - 97% probability of determining whether a person has diabetes. Some features had a pretty good probability of determining whether a person has diabetes but not all patients had those symptoms so this is good if the patient has the high probability symptom but not as good if the patient does not have a high probability symptom.

Modelling gave us 93% - 98% probability of determining whether a person has diabetes. This is very good. Random Forest gives us the maximum probability of determining whether a person has diabetes at 98%.

Early detection of diabetes can be very beneficial to a person's way of life and to their pocketbook. The use of machine learning can help predict diabetes with 98% accuracy.

For Future Consideration

The data had only 520 instances. Collecting more data may help to make the model prediction more accurate.