Diabetes Data-Diving

Anvitha Kosaraju, Sithara Samudrala, Nilai Vemula

MATH 2820L

Intro

- The data was collected from the Sylhet Diabetes Hospital in Bangladesh through direct questionnaires to 520 patients.
- All answers were approved by a doctor to ensure validity within the data.
- Provides information that can be used to determine possible correlations between having diabetes and other common risk factors.

Data from the UCI Machine Learning Repository

Features

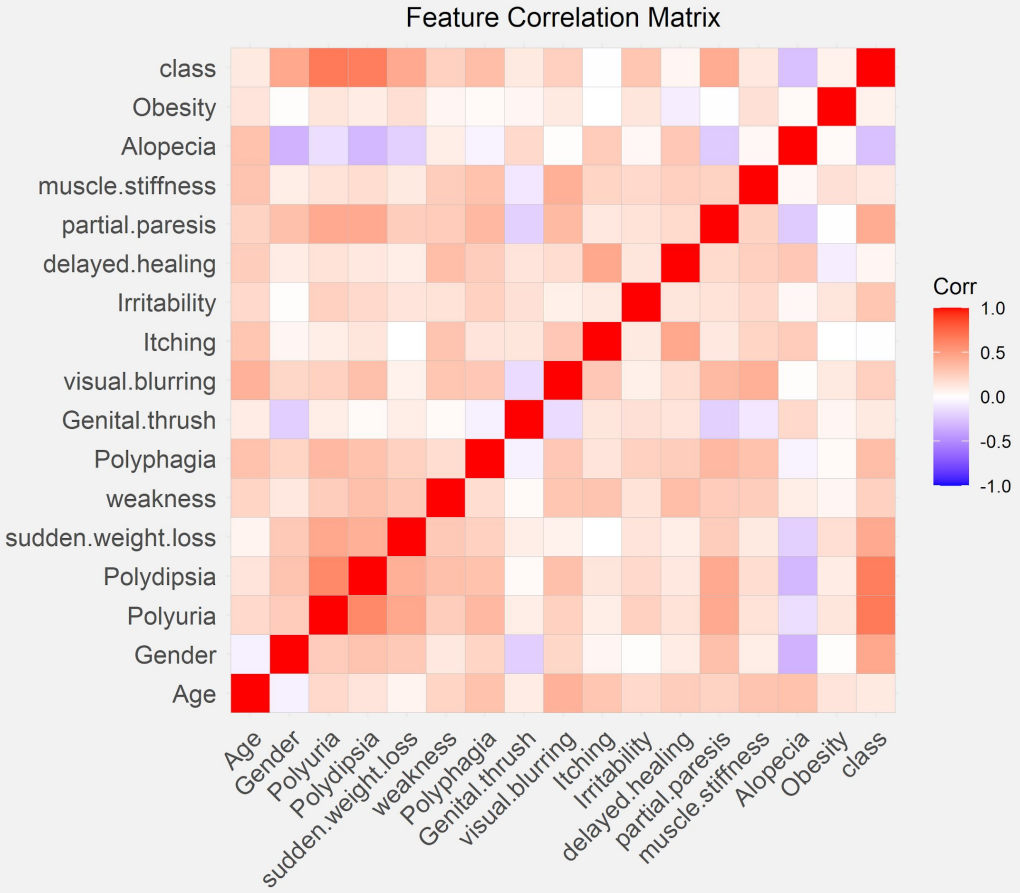
- **Variables included within the data:**
  - Age, Sex, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital Thrush, Visual Blurring, Itching, Irritability, Delayed Healing, Partial Paresis, Muscle Stiffness, Alopecia, and Obesity
- **Response Variable:** Class (Positive or Negative for Diabetes)

Our Goals

1. Analyze the data to make novel conclusions about diabetes and related medical conditions.
2. Build a predictive model to be able to accurately determine if a patient has diabetes given their medical data.
3. Determine which factors are the best predictors for diabetes.

Correlation Matrix

We computed Pearson correlation coefficients for each combination of features.



$\chi 2$ Feature Selection

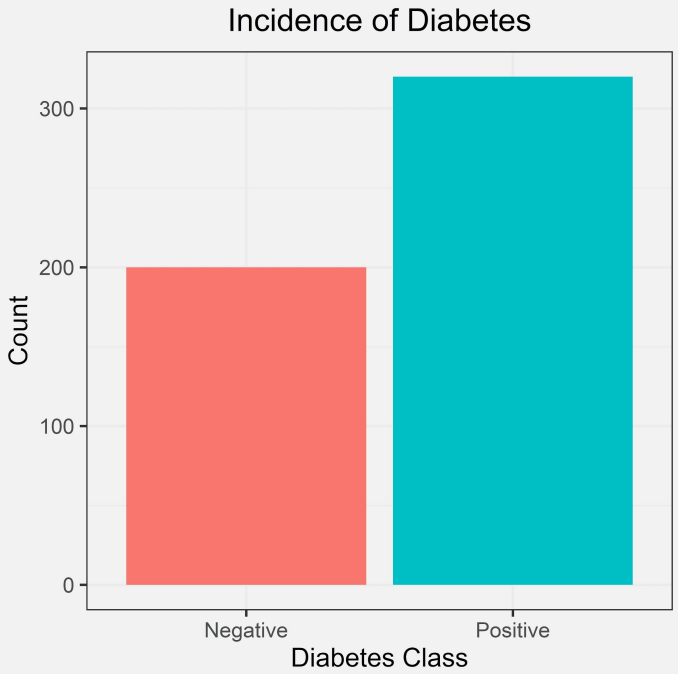
Use a $\chi 2$ test of independence to select features that less independent from the diabetes status.

Feature	$\chi 2$ Test Statistic
Polyuria	227.9
Polydipsia	216.2
Gender	103.1
Sudden Weight Loss	97.3
Partial Paresis	95.4
Polyphagia	59.6
Irritability	45.2
Alopecia	36.1
Visual Blurring	31.8
Weakness	29.8

Selected Features

- Age
- Gender
- Polydipsia
- Polyurea
- Sudden Weight Loss
- Partial Paresis

Diabetes Diagnosis



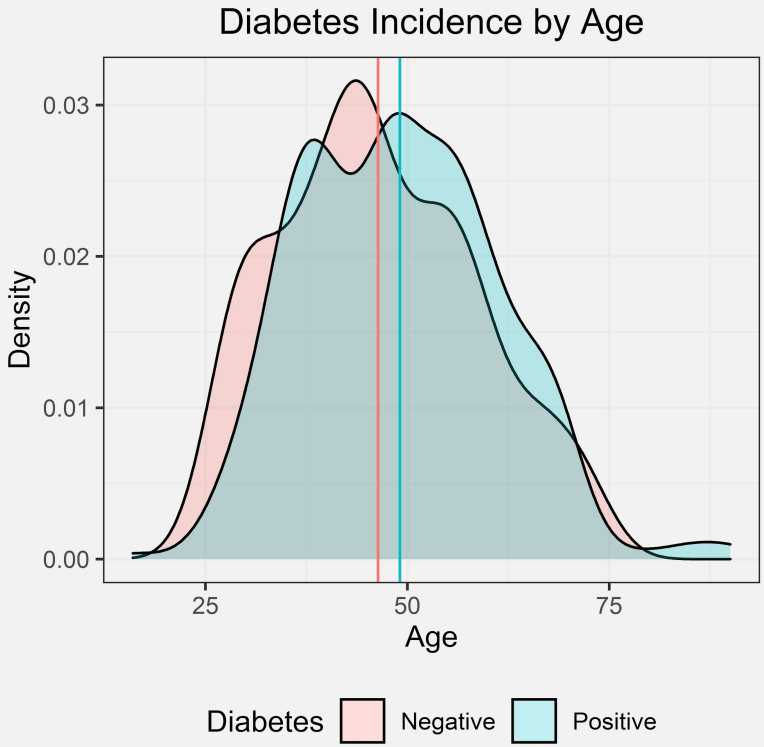
OUT OF 520 PATIENTS

Number of Patients Diagnosed with Diabetes : 320

Number of Patients NOT Diagnosed with Diabetes : 200

Age

Initial Hypothesis:
The greater an individual's age is, the higher the chance they are diagnosed with diabetes.



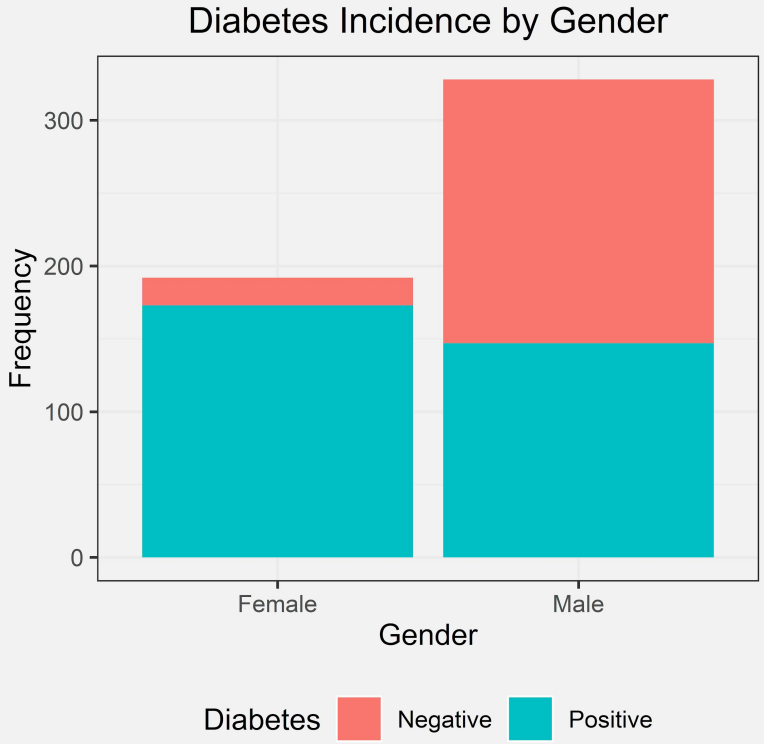
Conclusion:
There is only a slightly higher average age for patients with diabetes compared to patients without diabetes.

Mean:

- Positive: 49.07
- Negative: 46.36

Gender

Initial Hypothesis:
There will be no correlation between gender and the diagnosis of diabetes. Therefore, the diagnosis of diabetes will be fairly evenly distributed among the two genders presented in the data.



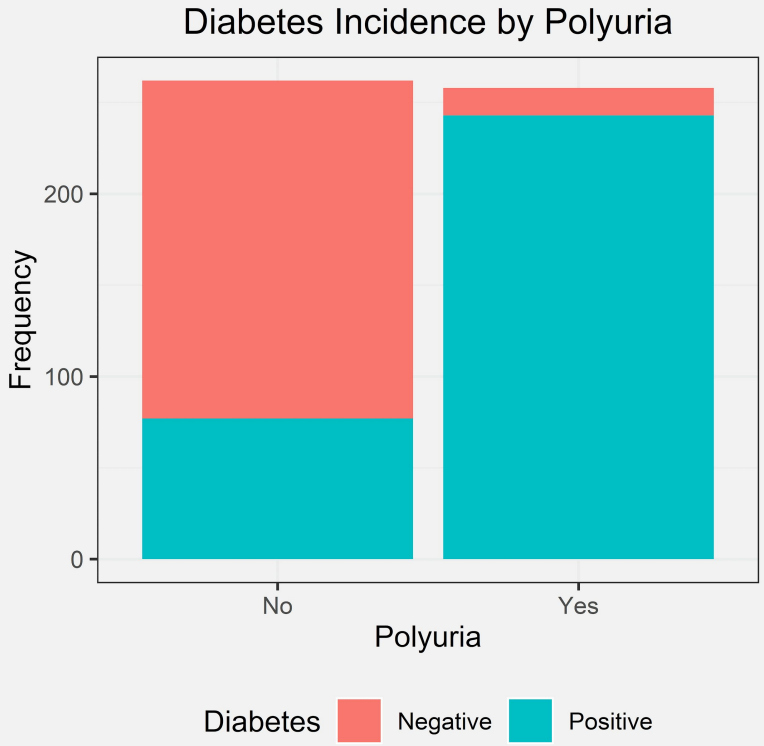
Male: 328
Female: 192

Conclusion:
In this dataset, the overwhelming majority of female patients were diagnosed with diabetes compared to male patients.

Polyuria

Polyuria: production of abnormally large volumes of dilute urine.

Initial Hypothesis: If the individual has polyuria, then they have a much higher chance of being diagnosed with diabetes.

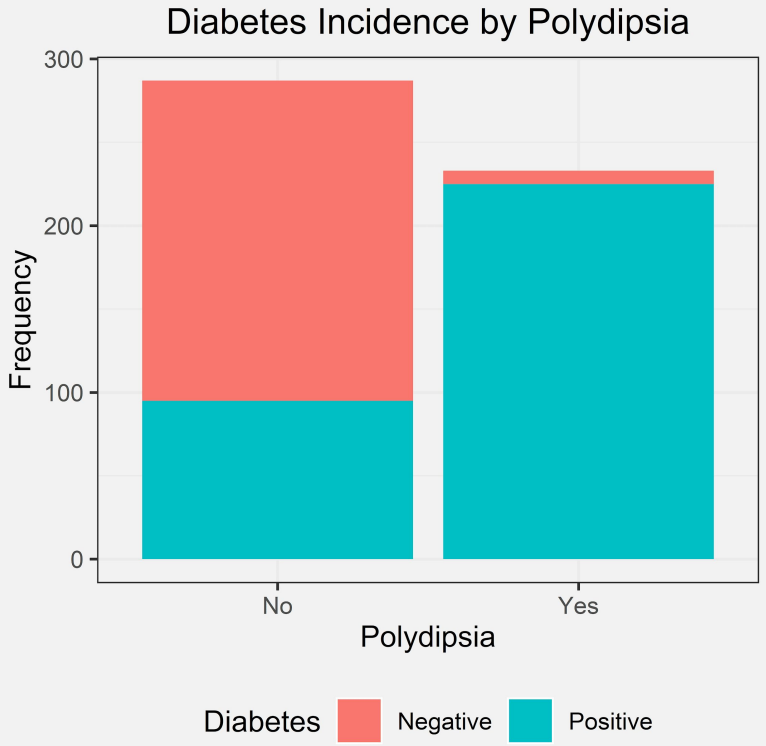


Polyuria: 258
No Polyuria: 262

Conclusion: Nearly all patients with polyuria have diabetes. Therefore, if a patient experienced polyuria, they have a higher chance of being diagnosed with diabetes.

Polydipsia

Polydipsia: Abnormally excessive thirst.
Initial Hypothesis:
If the individual has polydipsia, then they have a much higher chance of being diagnosed with diabetes.

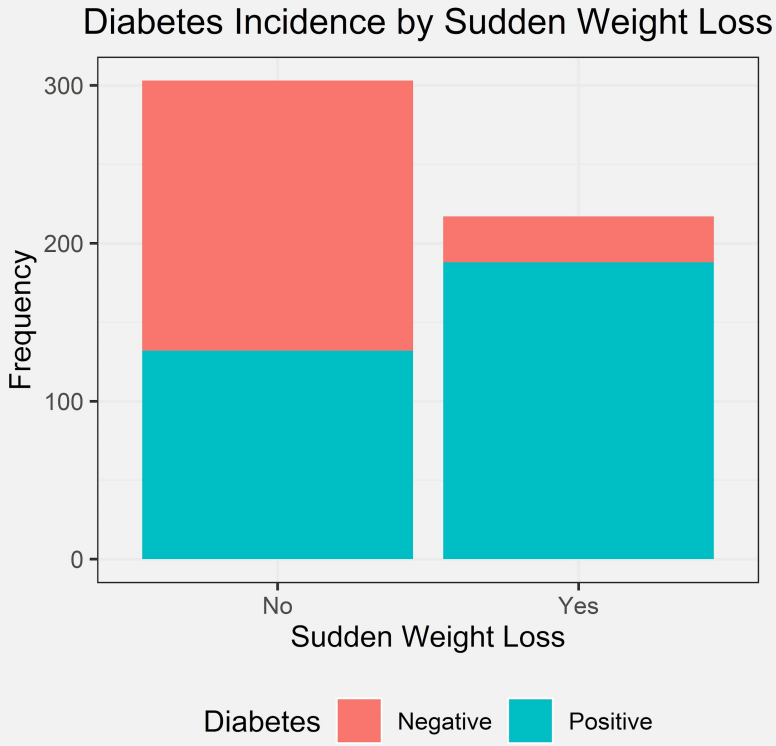


Polydipsia: 233
No Polydipsia: 287

Conclusion:
Nearly all patients with polydipsia have diabetes. Therefore, if a patient experienced polydipsia, they have a higher chance of being diagnosed with diabetes.

Sudden Weight Loss

Initial Hypothesis:
If the individual had recently experienced sudden weight loss, then they have a much higher chance of being diagnosed with diabetes.



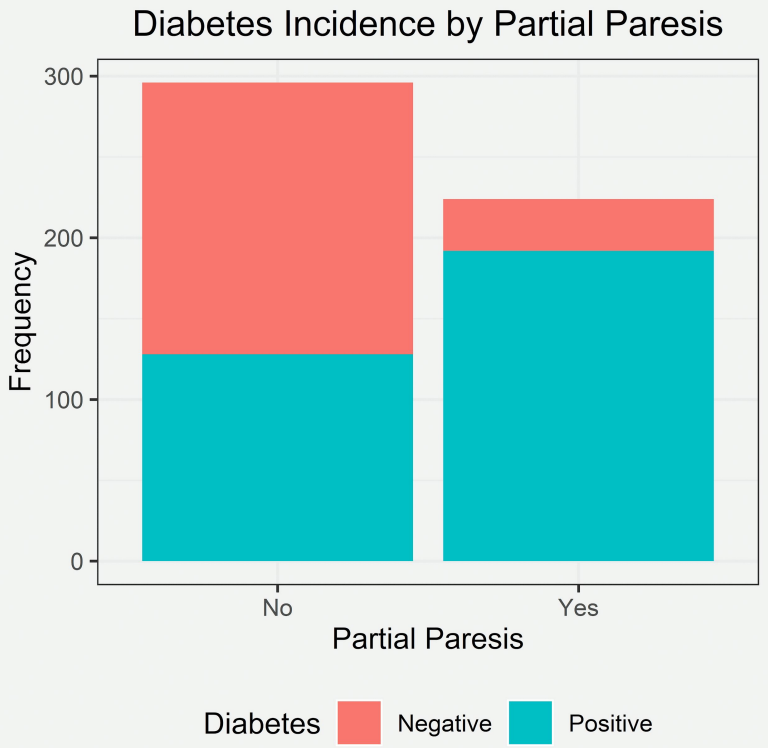
Conclusion:
If a patient has had sudden weight loss, it is very likely that they have diabetes, but the converse is not true.

Sudden Weight Loss: 217
No Sudden Weight Loss: 303

Partial Paresis

Paresis: condition typified by a weakness of voluntary movement

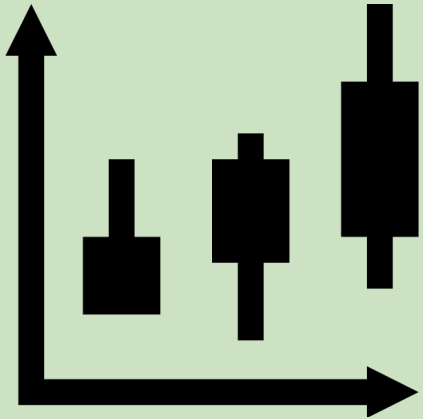
Initial Hypothesis:
If the individual has partial paresis, then they have a much higher chance of being diagnosed with diabetes.



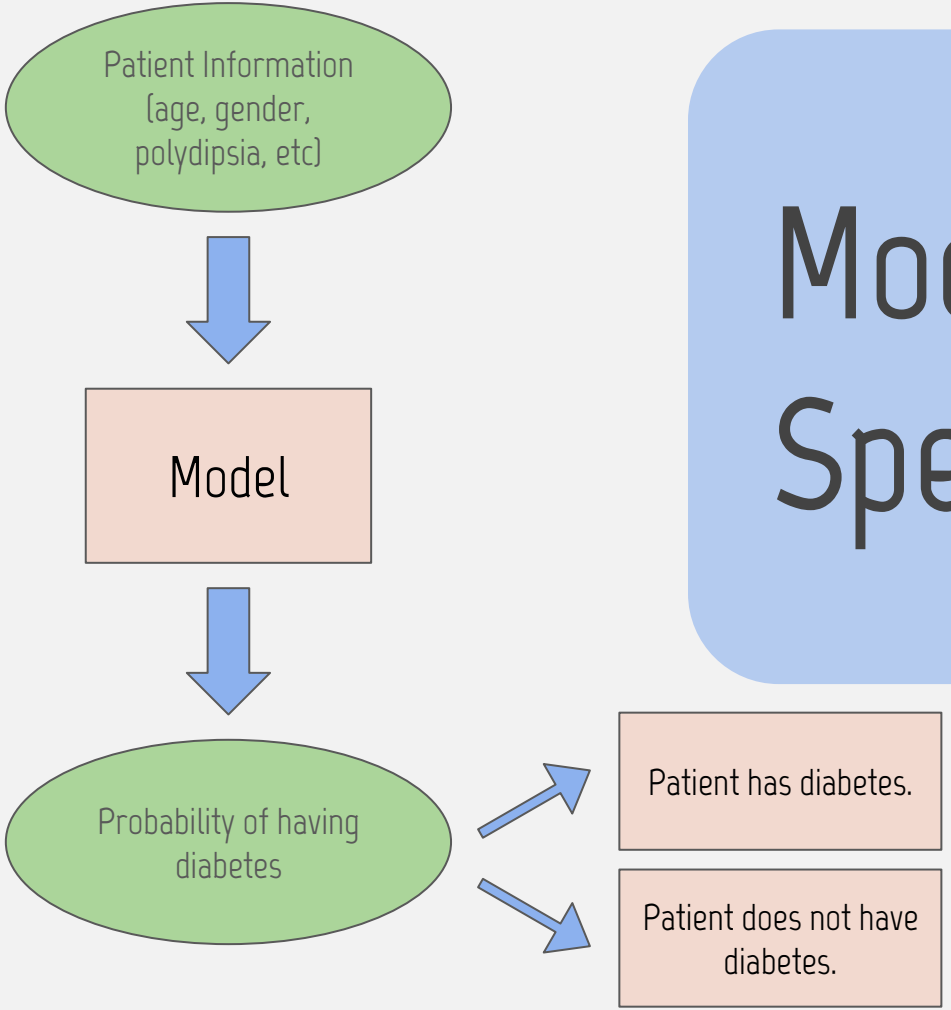
Partial Paresis: 224
No Partial Paresis: 296

Conclusion:
If a patient has partial paresis, it is very likely that they have diabetes, but the converse is not true.

Model Building



Goal: Build a model that can predict if a patient has diabetes.



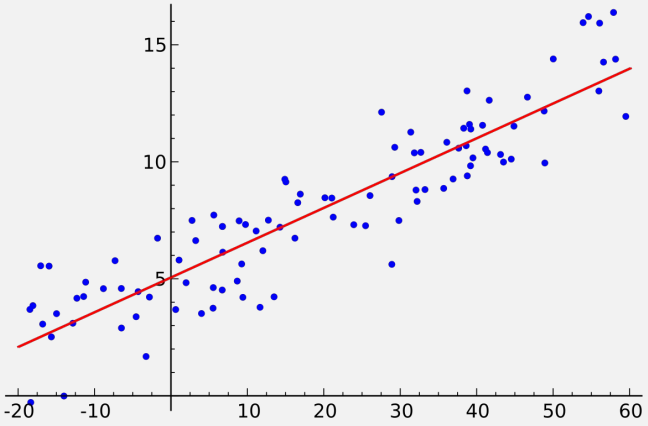
Model Specifications

Our model is a binary classifier.

Types of Models

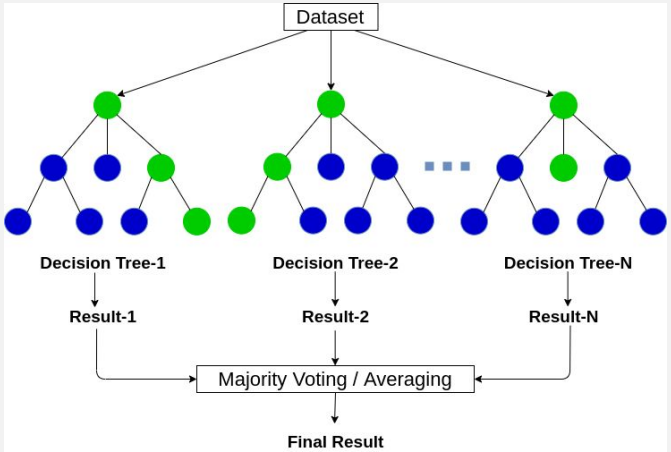
Linear Regression

Simple and easy to interpret but limited to linear relationships.



Random Forest

Complex ensemble of many decision trees. High accuracy.



Linear Regression

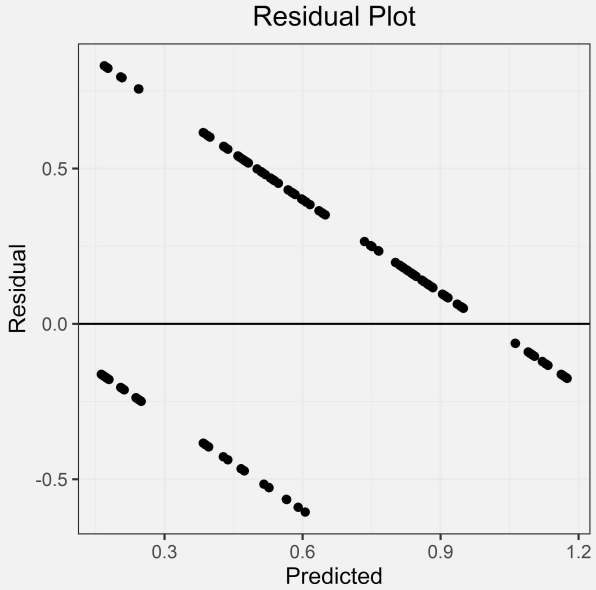
Performance Metrics

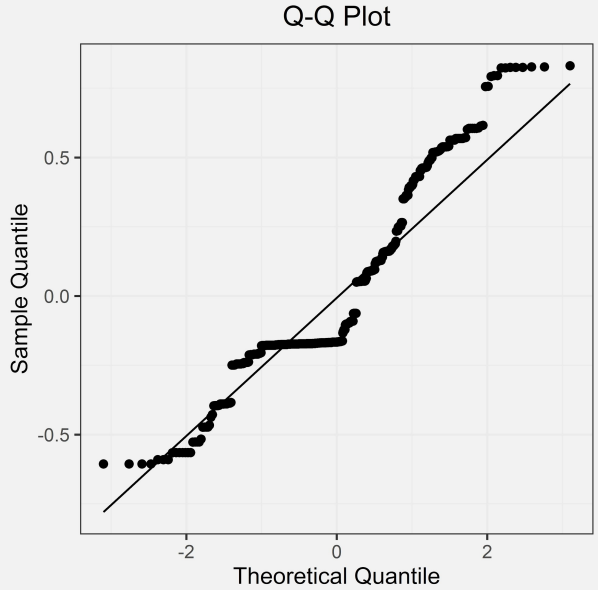
MSPE: 0.0956
Adjusted R^2: 59.6%

	Estimate	P-value
(Intercept)	0.19	**0.001**
Age	-3.6E-4	0.75
Gender	0.22	**9.8E-12**
Polyuria	0.36	**<2E-16**
Polydipsia	0.31	**3.47E-16**
Sudden Weight Loss	0.07	**0.03**
Partial Paresis	0.04	0.23

Linear Regression Assumptions

The features are not necessarily independent.
Our residuals are not randomly dispersed and the Q-Q plot does not closely fit the line.

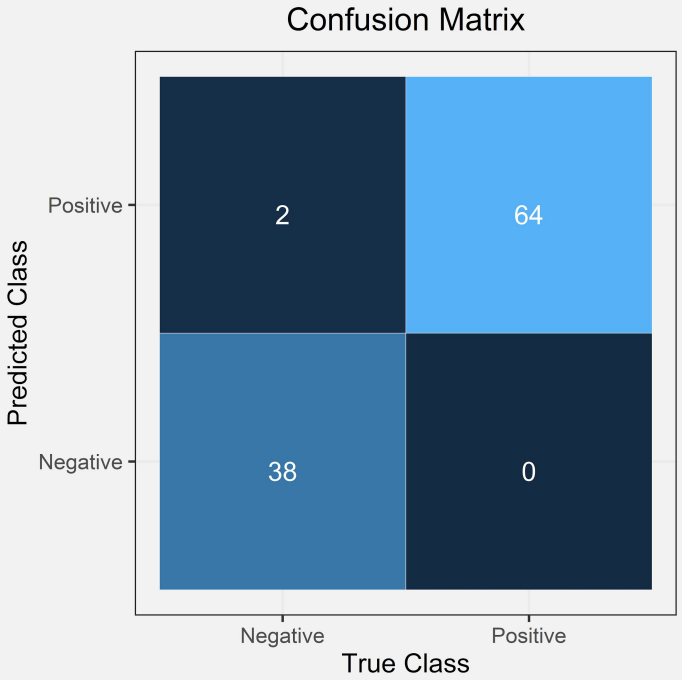




Random Forest Performance

Trained on 416/520 samples with 96.9% accuracy.

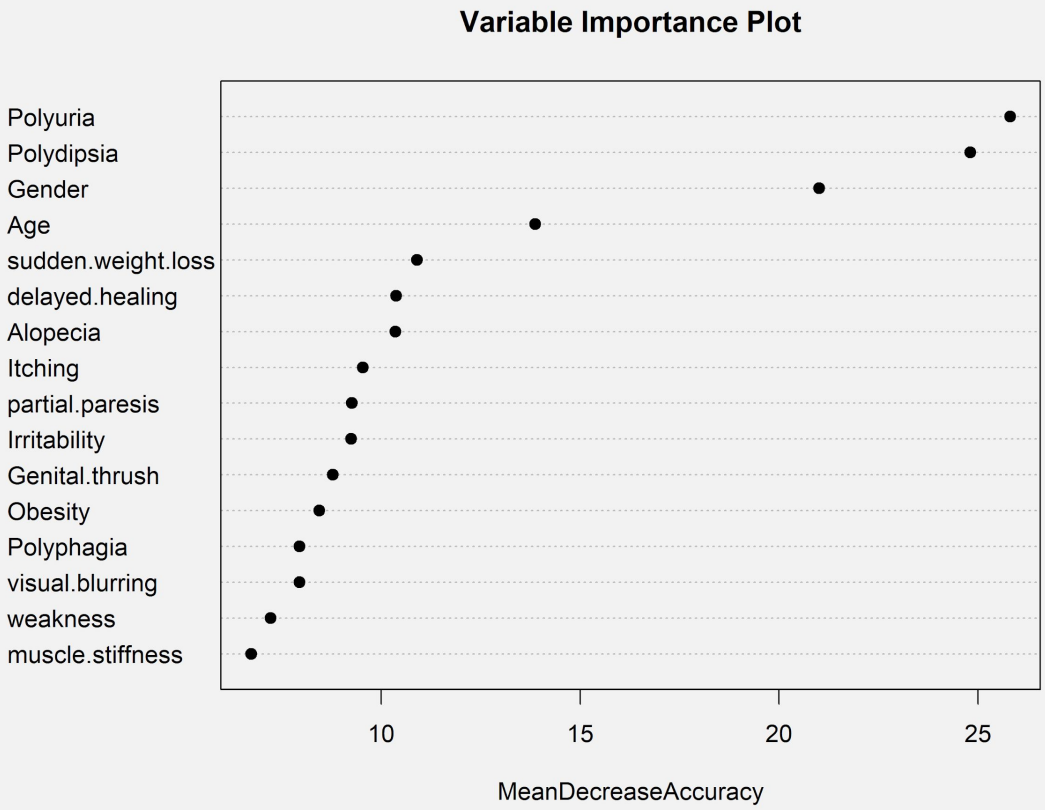
Tested on 104/520 samples with 98.1% accuracy.



- A **confusion matrix** is a special type of contingency table that lets us see the true/false positives/negatives.
- Our model is very accurate and the only errors are **false positives**.

Random Forest Insights

We can use the random forest model structure to rank how important each feature is to the overall result.



Conclusions

- The presence of conditions such as polydipsia, polyuria, sudden weight loss, and partial paresis makes it very likely that a patient has diabetes.
- A random forest model is very accurate when predicting for diabetes.
- The random forest model indicates that polyuria, polydipsia, age, and gender are the most useful features for predicting the diagnosis of diabetes.