Utilizing Data in Early Detection of Diabetes

A presentation to Physicians of the Bangladesh National Public Health Committee

UChicago DABP1400
Group - Amit Bansal, Nicholas Brune, Jacob Hsu, and Brian Kim

Project Goal

1. Build a model that accurately predicts an individual's risk factor for diabetes to aid in early detection of diabetes in the public.

2. Implement the model in public health screenings to identify at risk individuals and detect diabetes in patients at earliest possible stage.

3. Individuals made aware with early detection lead healthier lives with the disease and avoid blindness, amputations, heart disease, stroke, and kidney failure.

Utilizing Data in Early Detection of Diabetes

A presentation to Physicians of the Bangladesh National Public Health Committee

UChicago DABP1400
Group - Amit Bansal, Nicholas Brune, Jacob Hsu, and Brian Kim

Project Goal

1. Build a model that accurately predicts an individual's risk factor for diabetes to aid in early detection of diabetes in the public.

2. Implement the model in public health screenings to identify at risk individuals and detect diabetes in patients at earliest possible stage.

3. Individuals made aware with early detection lead healthier lives with the disease and avoid blindness, amputations, heart disease, stroke, and kidney failure.

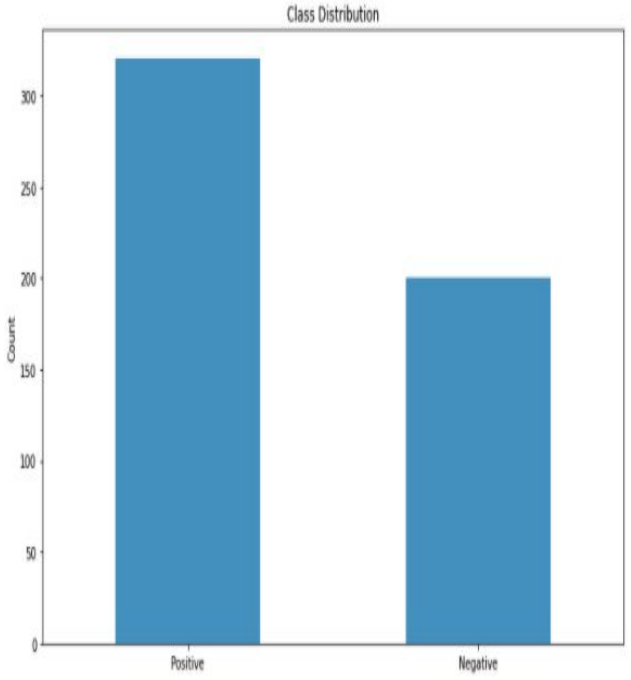
EDA Dataset

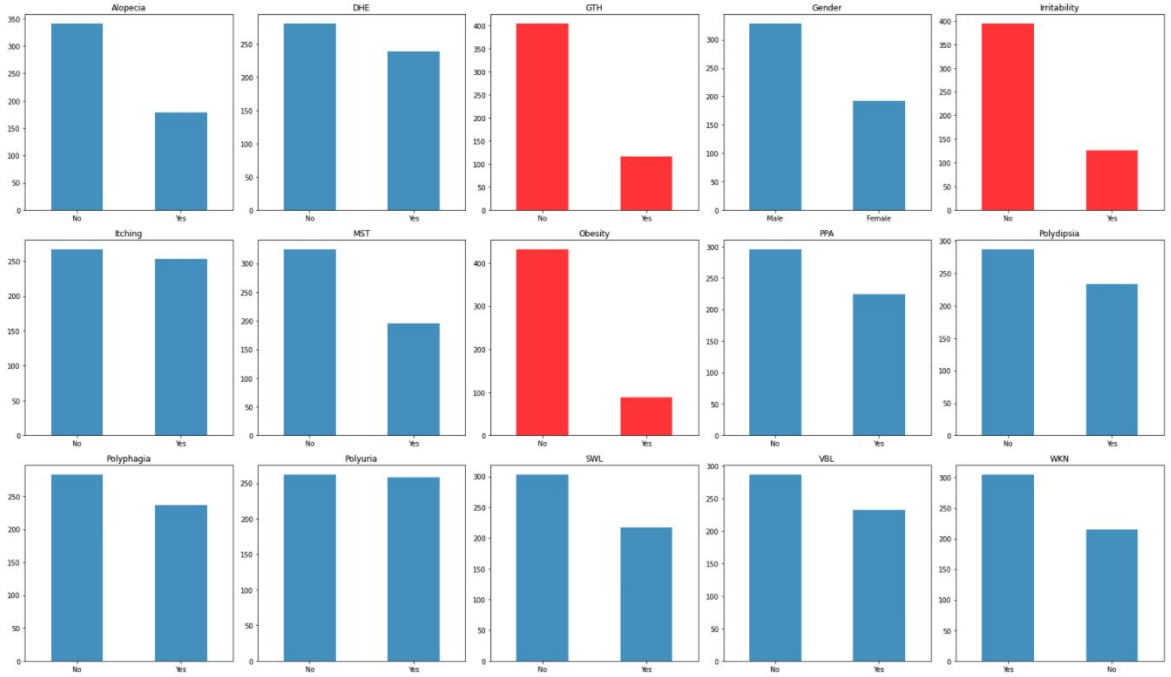
	Age	Gender	Polyuria	Polydipsia	SWL	WKN	Polyphagia	GTH	VBL	Itching	Irritability	DHE	PPA	MST	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

```
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 520 non-null    int64
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
```

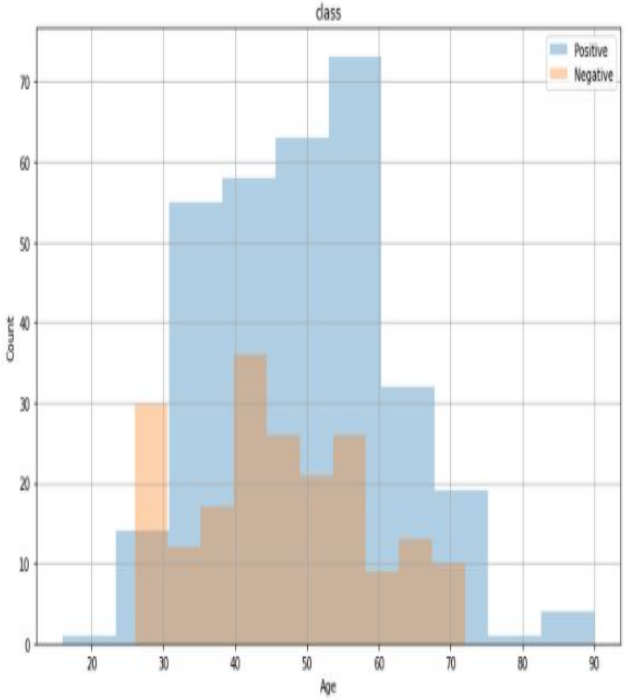
- 16 predictors, 1 target
- 1 numeric predictor, 15 categorical predictors
- Target is binary [Positive, Negative]
- No missing values in the dataset

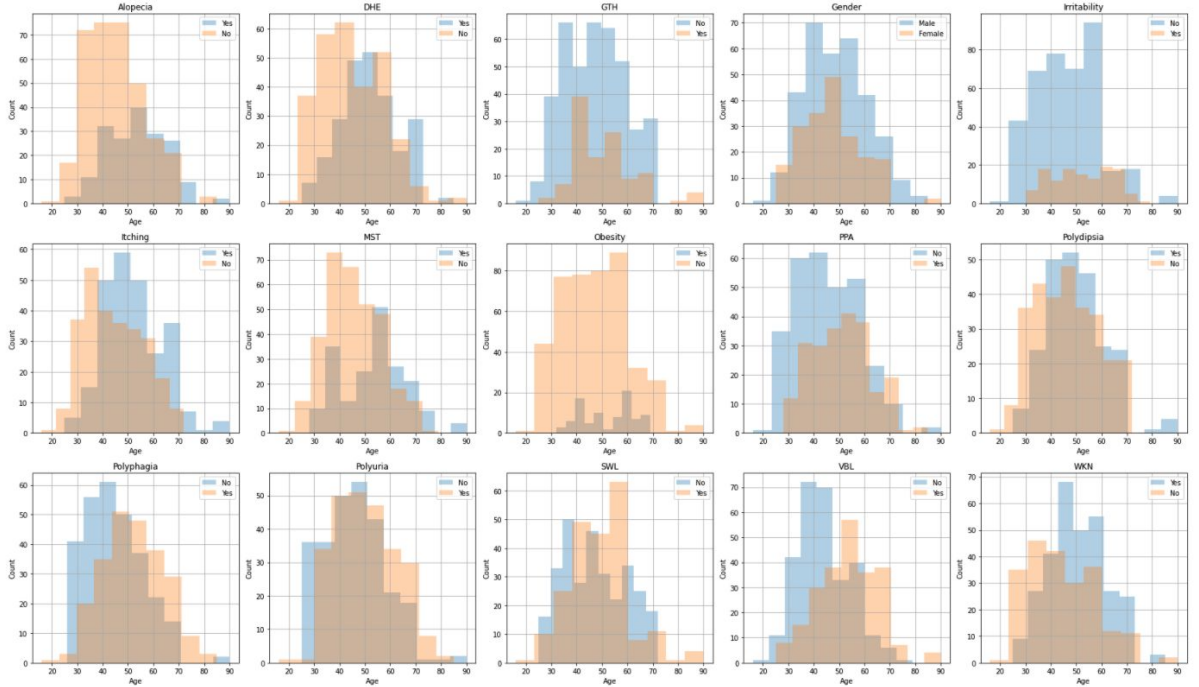
EDA Imbalance distribution





EDA Distributions by age





EDA Pearson correlation

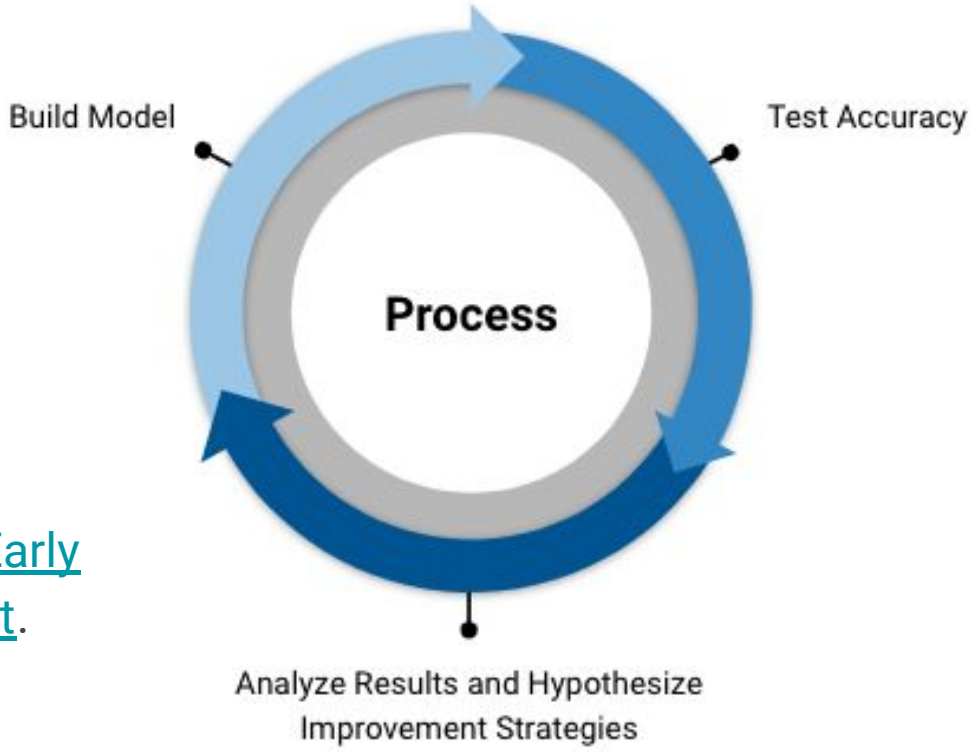
About the Model

Data Source

The dataset is a collection of records from 520 patients at the Sylhet Diabetes Hospital in Syhlet, Bangladesh. Data was collected from a questionnaire approved and administered by doctors at the hospital. In each patient record, indications were made on whether or not the patient had a list of symptoms and also whether or not the patient has or does not have diabetes.

Link to Data

https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.



About the Model

Data Variables

Age - Patients age

Sex- Is the patient Male or Female

Polyuria - Yes or No does the patient have large amounts of dilute urine

Polydipsia - Yes or No does the patient have abnormally high levels of thirst

Sudden Weight Loss 'SWL' - Yes or No has the patient experienced sudden weight loss

Weakness 'WKN' - Yes or No has the patient felt abnormally weak

Polyphagia - Yes or No does the patient have excess hunger

Genital thrush 'GTH'- Yes or No does the patient have genital thrush, a genital yeast infection found in both men and women

Visual Blurring 'VBL' - Yes or No does the patient experience less sharpness in vision

Itching - Yes or No does the patient experience excess itching

Irritability - Yes or No does the patient experience irritability

Delayed Healing 'DHE' - Yes or No does the patient experience delayed healing

Partial Paresis 'PPE' - Yes or No has the patient lost feeling or control in their muscles

Muscle Stiffness 'MST' - Yes or No does the patient have muscle stiffness

Alopecia - Yes or No has the patient experienced baldness in spots where hair normal grows

Obesity - Yes or No is the patient obese

Class 1.Positive, 2.Negative - does the patient have diabetes

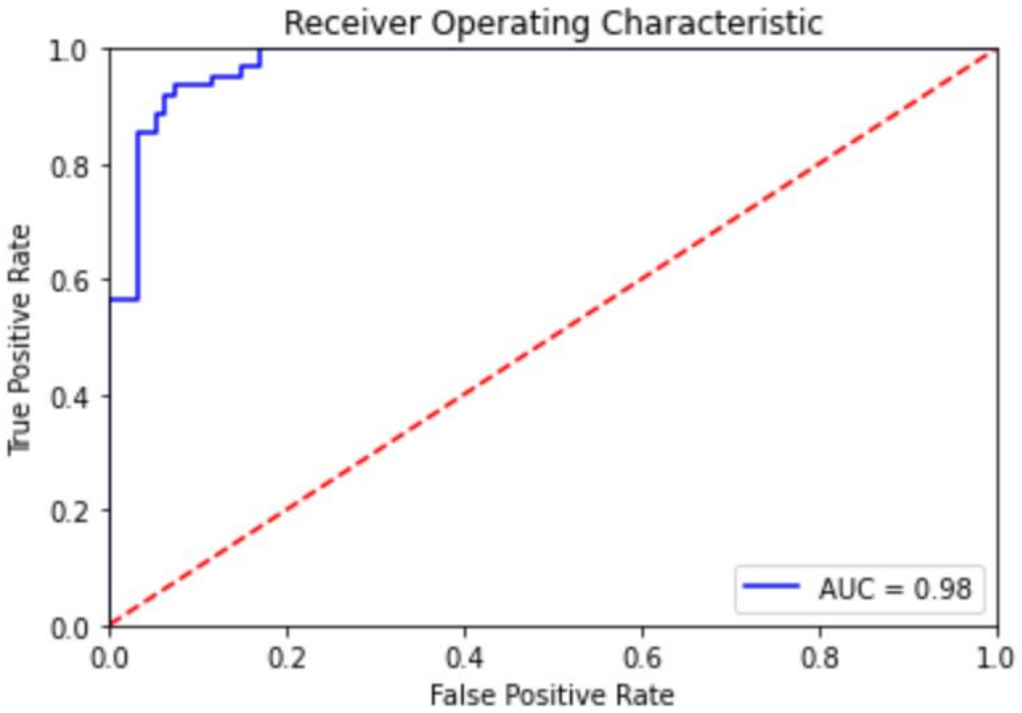
Results and Accuracy

Data was split into training and testing sets. Training data formed the model and test data determined the accuracy of the model's predictions.

We measured accuracy by an accuracy score, which is the percentage of correct predictions performed by the model using the test data. Accuracy score was 93%.

Area under the curve is another useful test of accuracy. The higher AUC the more useful the model. Our best model had a significant AUC of .98.

Most important variables in prediction were Polydipsia, Polyuiria, Gender, Age, and Irritability



Baseline Model

Baseline Model

Dep. Variable:	class	No. Observations:	
Model:	Logit	Df Residuals:	
Method:	MLE	Df Model:	
Date:	Wed, 17 Feb 2021	Pseudo R-squ.:	0.7
Time:	03:17:03	Log-Likelihood:	-60.
converged:	True	LL-Null:	-241
Covariance Type:	nonrobust	LLR p-value:	5.293e

	coef	std err	z	P>|z|	[0.025	0
Age	0.0872	0.015	5.941	0.000	0.058	
Gender	-4.4376	0.721	-6.153	0.000	-5.851	-
Polyuria	-4.4387	0.807	-5.501	0.000	-6.020	-
Polydipsia	-5.2667	1.005	-5.239	0.000	-7.237	-
SWL	-0.3365	0.683	-0.493	0.622	-1.675	
WKN	-0.3068	0.681	-0.451	0.652	-1.641	
Polyphagia	-1.0227	0.621	-1.647	0.100	-2.240	
GTH	-1.7942	0.653	-2.747	0.006	-3.074	-
VBL	-0.3050	0.788	-0.387	0.699	-1.849	
Itching	2.3073	0.766	3.012	0.003	0.806	
Irritability	-2.6303	0.734	-3.583	0.000	-4.069	-
DHE	0.3592	0.669	0.537	0.592	-0.953	
PPA	-1.0099	0.713	-1.417	0.156	-2.407	
MST	0.0190	0.672	0.028	0.977	-1.297	
Alopecia	-0.3214	0.698	-0.461	0.645	-1.689	
Obesity	0.2235	0.645	0.347	0.729	-1.040	

Best Model

Improve Baseline Model by Dropping SWL

Dep. Variable:	class	No. Observations:	
Model:	Logit	Df Residuals:	
Method:	MLE	Df Model:	
Date:	Wed, 17 Feb 2021	Pseudo R-squ.:	0.7
Time:	03:17:04	Log-Likelihood:	-60.
converged:	True	LL-Null:	-241
Covariance Type:	nonrobust	LLR p-value:	1.145e

	coef	std err	z	P>|z|	[0.025	0
Age	0.0862	0.015	5.934	0.000	0.058	
Gender	-4.4718	0.717	-6.233	0.000	-5.878	-
Polyuria	-4.5354	0.791	-5.734	0.000	-6.086	-
Polydipsia	-5.4035	0.983	-5.496	0.000	-7.331	-
WKN	-0.4750	0.585	-0.812	0.417	-1.622	
Polyphagia	-1.0076	0.615	-1.639	0.101	-2.212	
GTH	-1.8445	0.647	-2.852	0.004	-3.112	-
VBL	-0.2132	0.765	-0.279	0.781	-1.713	
Itching	2.3672	0.760	3.114	0.002	0.877	
Irritability	-2.5559	0.707	-3.616	0.000	-3.941	-
DHE	0.4002	0.661	0.606	0.545	-0.895	
PPA	-1.0814	0.692	-1.562	0.118	-2.438	
MST	0.0724	0.657	0.110	0.912	-1.215	
Alopecia	-0.3156	0.701	-0.450	0.653	-1.690	
Obesity	0.2131	0.643	0.332	0.740	-1.047	

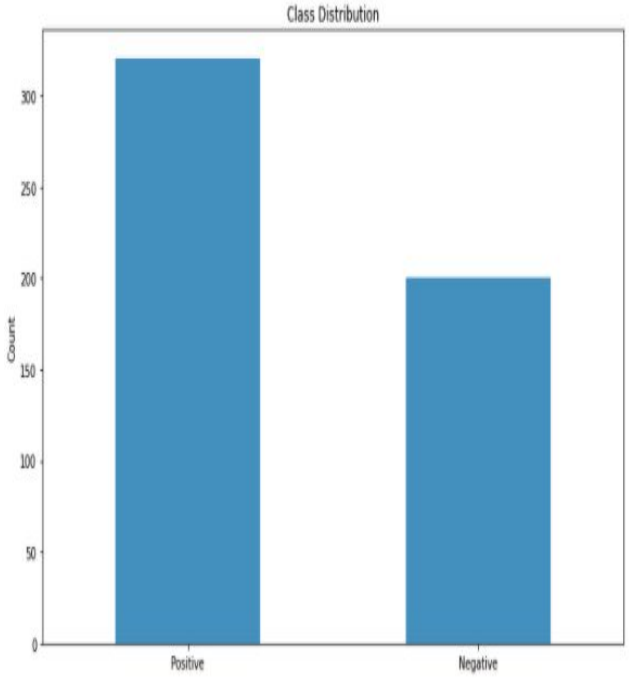
EDA Dataset

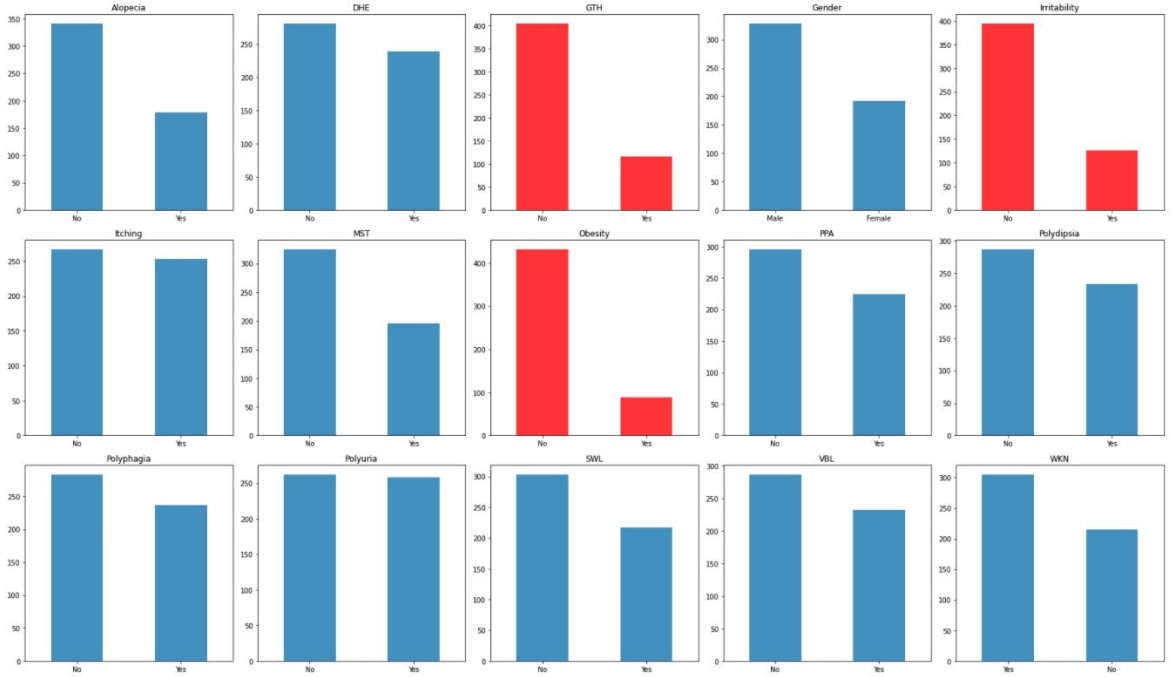
	Age	Gender	Polyuria	Polydipsia	SWL	WKN	Polyphagia	GTH	VBL	Itching	Irritability	DHE	PPA	MST	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

```
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 520 non-null    int64
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
```

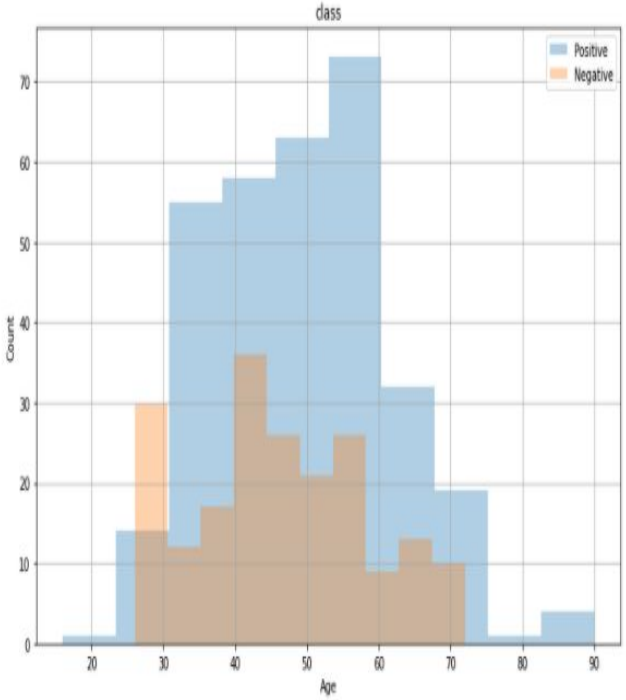
- 16 predictors, 1 target
- 1 numeric predictor, 15 categorical predictors
- Target is binary [Positive, Negative]
- No missing values in the dataset

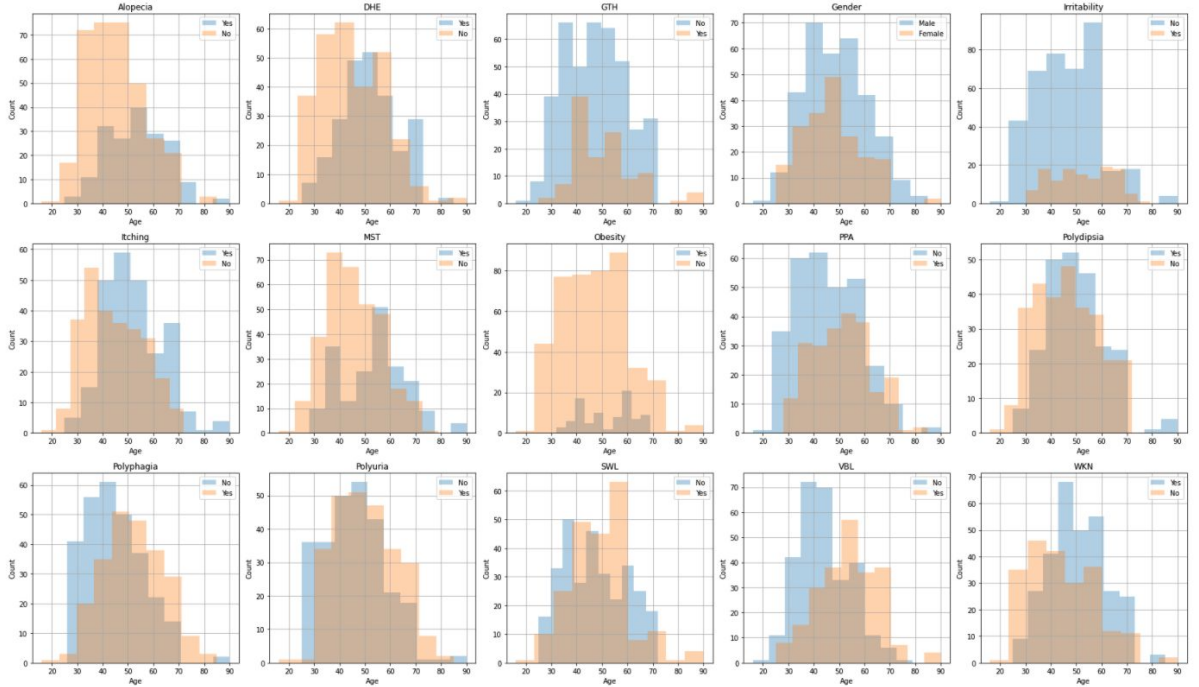
EDA Imbalance distribution





EDA Distributions by age





EDA Pearson correlation

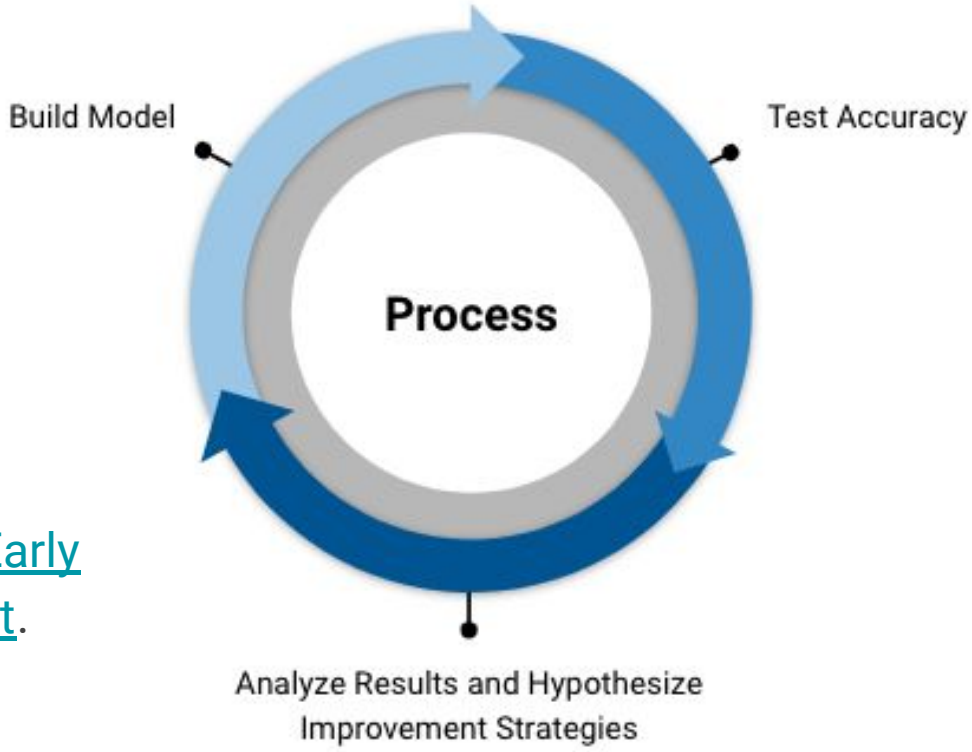
About the Model

Data Source

The dataset is a collection of records from 520 patients at the Sylhet Diabetes Hospital in Syhlet, Bangladesh. Data was collected from a questionnaire approved and administered by doctors at the hospital. In each patient record, indications were made on whether or not the patient had a list of symptoms and also whether or not the patient has or does not have diabetes.

Link to Data

https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.



About the Model

Data Variables

Age - Patients age

Sex- Is the patient Male or Female

Polyuria - Yes or No does the patient have large amounts of dilute urine

Polydipsia - Yes or No does the patient have abnormally high levels of thirst

Sudden Weight Loss 'SWL' - Yes or No has the patient experienced sudden weight loss

Weakness 'WKN' - Yes or No has the patient felt abnormally weak

Polyphagia - Yes or No does the patient have excess hunger

Genital thrush 'GTH'- Yes or No does the patient have genital thrush, a genital yeast infection found in both men and women

Visual Blurring 'VBL' - Yes or No does the patient experience less sharpness in vision

Itching - Yes or No does the patient experience excess itching

Irritability - Yes or No does the patient experience irritability

Delayed Healing 'DHE' - Yes or No does the patient experience delayed healing

Partial Paresis 'PPE' - Yes or No has the patient lost feeling or control in their muscles

Muscle Stiffness 'MST' - Yes or No does the patient have muscle stiffness

Alopecia - Yes or No has the patient experienced baldness in spots where hair normal grows

Obesity - Yes or No is the patient obese

Class 1.Positive, 2.Negative - does the patient have diabetes

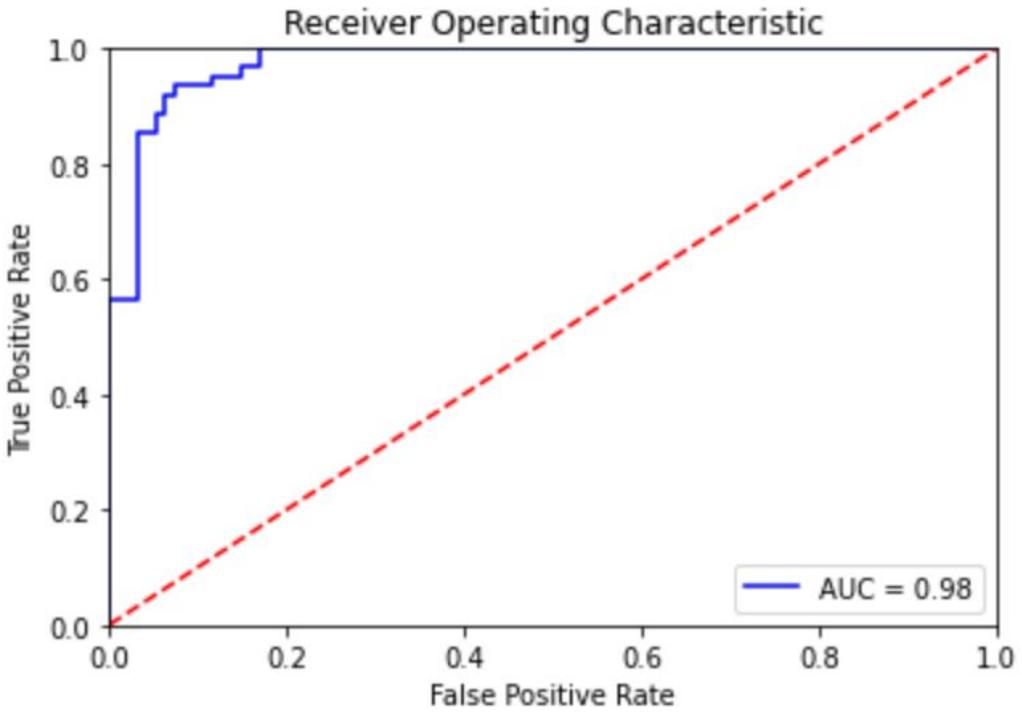
Results and Accuracy

Data was split into training and testing sets. Training data formed the model and test data determined the accuracy of the model's predictions.

We measured accuracy by an accuracy score, which is the percentage of correct predictions performed by the model using the test data. Accuracy score was 93%.

Area under the curve is another useful test of accuracy. The higher AUC the more useful the model. Our best model had a significant AUC of .98.

Most important variables in prediction were Polydipsia, Polyuiria, Gender, Age, and Irritability



Baseline Model

Baseline Model

Dep. Variable:	class	No. Observations:	
Model:	Logit	Df Residuals:	
Method:	MLE	Df Model:	
Date:	Wed, 17 Feb 2021	Pseudo R-squ.:	0.7
Time:	03:17:03	Log-Likelihood:	-60.
converged:	True	LL-Null:	-241
Covariance Type:	nonrobust	LLR p-value:	5.293e

	coef	std err	z	P>|z|	[0.025	0
Age	0.0872	0.015	5.941	0.000	0.058	
Gender	-4.4376	0.721	-6.153	0.000	-5.851	-
Polyuria	-4.4387	0.807	-5.501	0.000	-6.020	-
Polydipsia	-5.2667	1.005	-5.239	0.000	-7.237	-
SWL	-0.3365	0.683	-0.493	0.622	-1.675	
WKN	-0.3068	0.681	-0.451	0.652	-1.641	
Polyphagia	-1.0227	0.621	-1.647	0.100	-2.240	
GTH	-1.7942	0.653	-2.747	0.006	-3.074	-
VBL	-0.3050	0.788	-0.387	0.699	-1.849	
Itching	2.3073	0.766	3.012	0.003	0.806	
Irritability	-2.6303	0.734	-3.583	0.000	-4.069	-
DHE	0.3592	0.669	0.537	0.592	-0.953	
PPA	-1.0099	0.713	-1.417	0.156	-2.407	
MST	0.0190	0.672	0.028	0.977	-1.297	
Alopecia	-0.3214	0.698	-0.461	0.645	-1.689	
Obesity	0.2235	0.645	0.347	0.729	-1.040	

Best Model

Improve Baseline Model by Dropping SWL

Dep. Variable:	class	No. Observations:	
Model:	Logit	Df Residuals:	
Method:	MLE	Df Model:	
Date:	Wed, 17 Feb 2021	Pseudo R-squ.:	0.7
Time:	03:17:04	Log-Likelihood:	-60.
converged:	True	LL-Null:	-241
Covariance Type:	nonrobust	LLR p-value:	1.145e

	coef	std err	z	P>|z|	[0.025	0
Age	0.0862	0.015	5.934	0.000	0.058	
Gender	-4.4718	0.717	-6.233	0.000	-5.878	-
Polyuria	-4.5354	0.791	-5.734	0.000	-6.086	-
Polydipsia	-5.4035	0.983	-5.496	0.000	-7.331	-
WKN	-0.4750	0.585	-0.812	0.417	-1.622	
Polyphagia	-1.0076	0.615	-1.639	0.101	-2.212	
GTH	-1.8445	0.647	-2.852	0.004	-3.112	-
VBL	-0.2132	0.765	-0.279	0.781	-1.713	
Itching	2.3672	0.760	3.114	0.002	0.877	
Irritability	-2.5559	0.707	-3.616	0.000	-3.941	-
DHE	0.4002	0.661	0.606	0.545	-0.895	
PPA	-1.0814	0.692	-1.562	0.118	-2.438	
MST	0.0724	0.657	0.110	0.912	-1.215	
Alopecia	-0.3156	0.701	-0.450	0.653	-1.690	
Obesity	0.2131	0.643	0.332	0.740	-1.047	