Project Proposal

Oliver Grudzinski

Dataset

The dataset selected for this project is the Pima Indians Diabetes Database, originally provided by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset contains medical diagnostic measurements aimed at predicting the onset of diabetes. The dataset includes 768 instances, each representing a female patient of Pima Indian heritage, aged at least 21 years. It contains eight predictor variables (such as glucose levels, BMI, insulin levels, and number of pregnancies) and one target variable, "Outcome," which indicates whether a patient has diabetes (1) or not (0).

The primary goal of this analysis is to develop a predictive model using machine learning techniques to determine whether a patient has diabetes based on the provided diagnostic measurements. In this case, classification techniques are more relevant, as I aim to predict whether patients will develop diabetes based on their diagnostic data.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Variables: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age

Hypotheses:

Null Hypothesis Example - A given variable does not influence whether a patient may have diabetes.

Alternative Hypothesis Example - A given variable does influence whether a patient may have diabetes.

Most Likely Examples -

H0 (Glucose) – Glucose levels do not influence whether a patient may have diabetes.

HA (Glucose) – Glucose levels do have an influence on whether a patient may have diabetes.

H0 (BMI) – The patient's BMI does not influence whether a patient may have diabetes.

HA (BMI) – The patient's BMI does influence on whether a patient may have diabetes.

H0 (Age) – The patient's age does not influence whether a patient may have diabetes.

HA (Age) – The patient's age does influence on whether a patient may have diabetes.

H0 (Diabetes Pedigree Function) – The patient's DPF levels do not influence whether a patient may have diabetes.

HA (Diabetes Pedigree Function) – The patient's DPF levels do have an influence on whether a patient may have diabetes.

Tests:

Test: Kruskal-Willis will be run 8 times because there are 8 variables compared with Binary categorical. But our main goal is making sure that Bonferroni Correction value must be lower than threshold at all costs. Bonferroni Correction should be used when running multiple testing.

Model:

Should be a classification model because we are predicting whether the patient doesn't have diabetes (0) or does have diabetes (1).

1st Option-Random Forest lets us visualize why the model made the decision that it has.

2nd Option-Logistic Regression allows for easy binary categorization too.