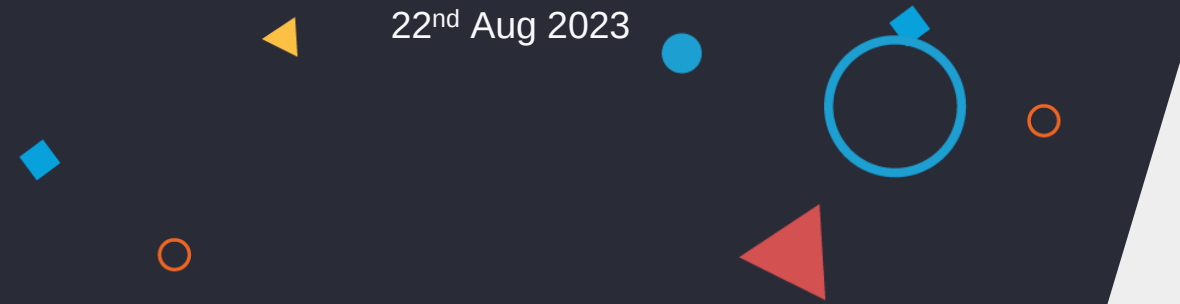
Supervised Learning Project

by
Abi Afolabi
22nd Aug 2023

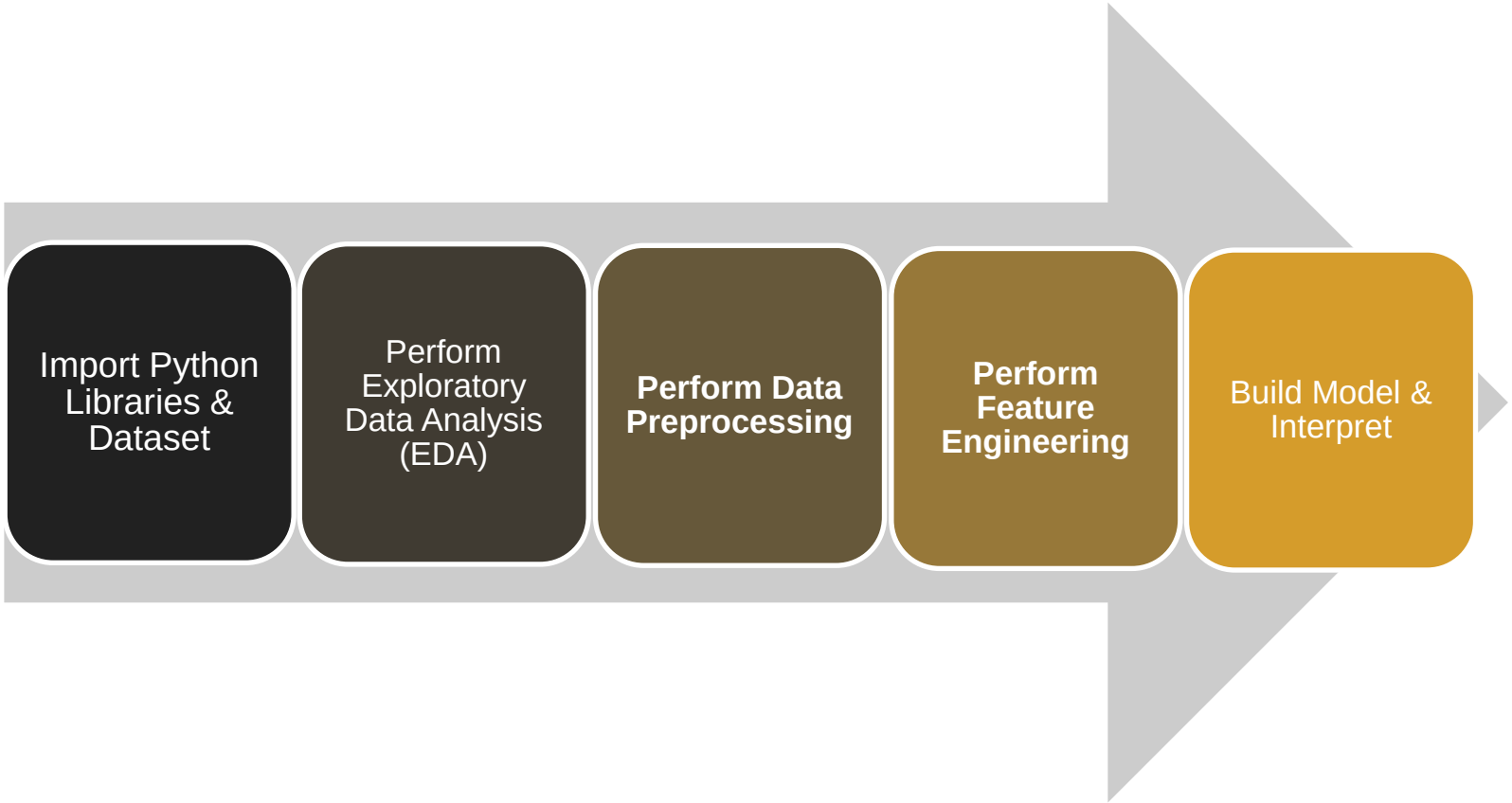


LIGHTHOUSE LABS

Project Overview & Goals

- Reinforce learning with hands-on experience
    - Use supervised learning techniques to build a machine learning model that can predict whether a patient has diabetes or not, based on certain diagnostic measurements.

Project Execution Steps

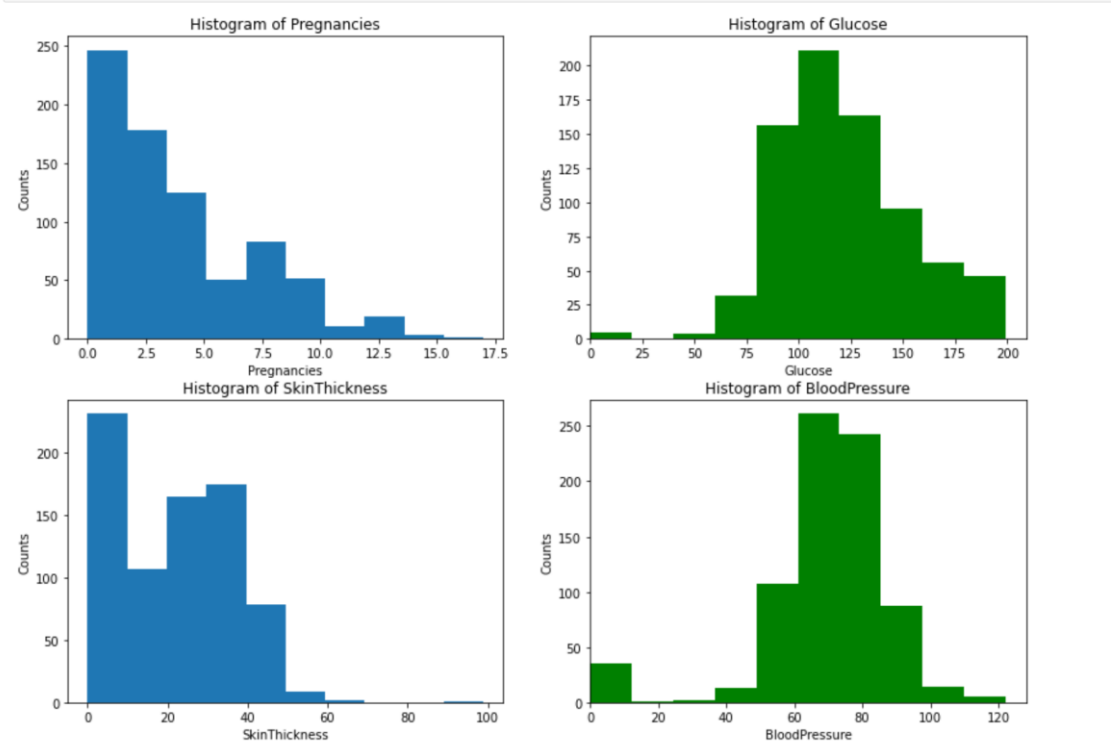


Dataset in Pandas DataFrame showing table columns and dimensions

	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

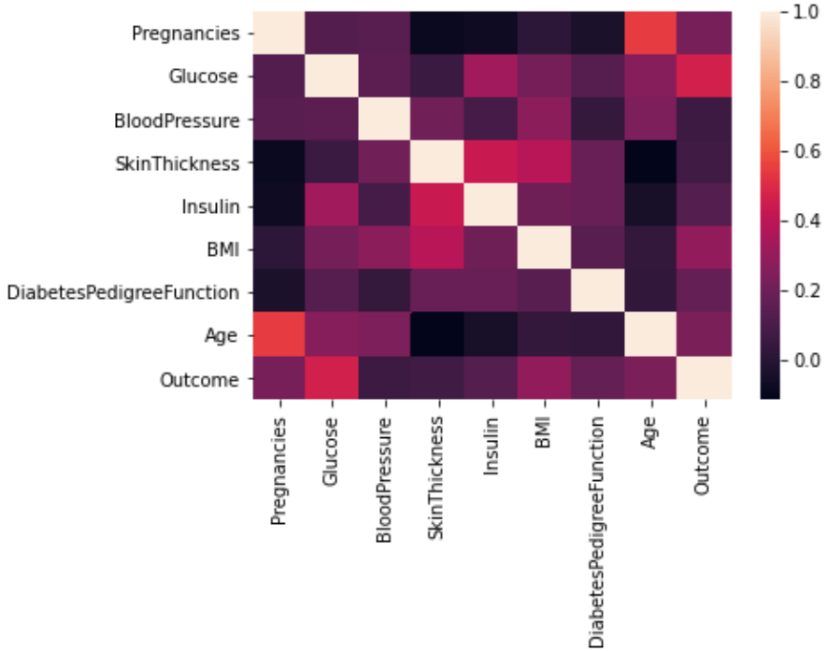
768 rows × 9 columns

Discoveries from EDA – Histogram



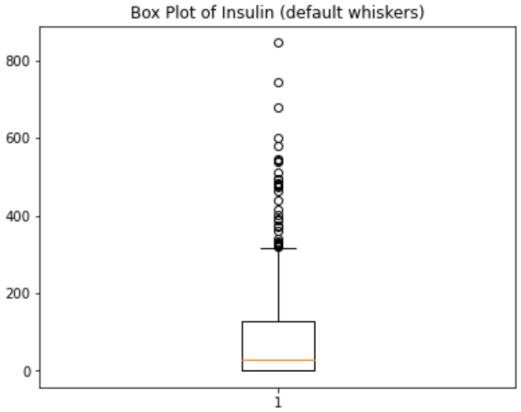
Some of the dataset is mostly symmetrical around the mean and some right skewed

Discoveries from EDA – Correlation Matrix

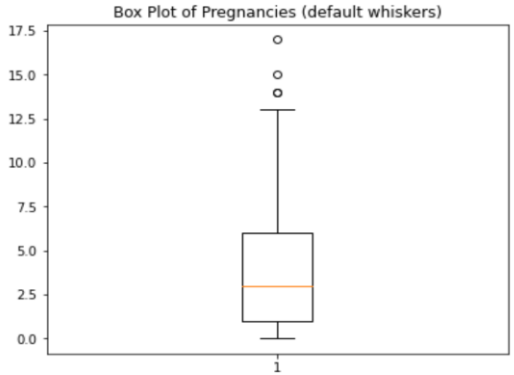


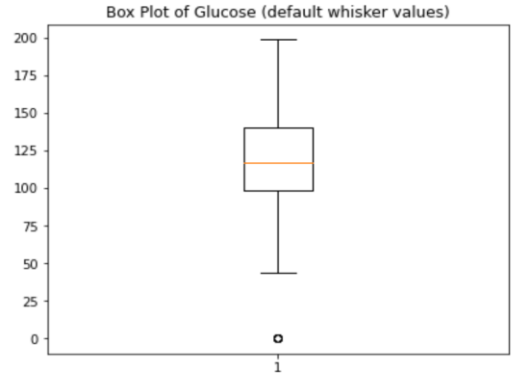
Strong correlation Glucose levels and Outcome

Discoveries from EDA – Box plots



Presence of outliers in dataset



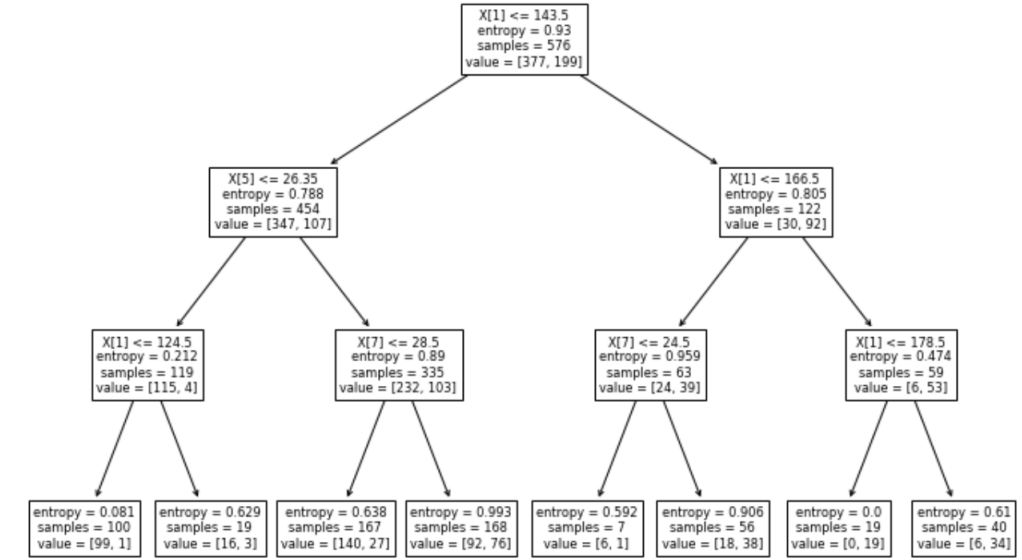


Main Challenge – Very Little Time

With more time I will….

- Explore the dataset with the EDA process.
- Investigate if outliers in dataset are real or not, then address them accordingly.
- Build a third model for further comparison

Decision Tree Model



Conclusion

- Glucose levels in patients with corelation coefficient of ~0.7 is a strong indicator of being diabetic while BloodPressure & SkinThickness with correlation coeficient of ~0.1 are low indicators of diabetes

- Adults in the age bracket between 30 and 44 years are mostly diabetic.

- Interaction effect between the predictor variables is insignificant since the Variance Inflation Factor (VIF) is close to 1 (far away from 5). Thus we can conclude that these parameters are good for model building.

- Result of accuracy from Random Forest (75%) is higher than Decison Trees (71.3%) proving that it is a viable ensemble technique for improving model performance.

https://github.com/AbiAfolabi/ml-project-supervised-learning