**Week 1**

* Outcome is binary so classification or logistic regression can be used
* no missing value
* Pregnancies can be treated as categorical(Need SME for more Knowledge)
* Pregnancies from 0 to 17
* For Glucose Clearly anything less than 50 is missing value Use median

From research we found out 140 is normal

normal < 100

preddeb 100-125

deb >125

* Need SME to guide further: can we converted to 3 categories
* Blood pressure <40 is missing value
* Skin Thickness zero not possible can br replace with mode
* Insulin lvl will be replaced by median to don' disturb distribution
* Bmi with median
* Used MeanMedianImputer as median…..need SME to do the above

**Week 2**

* Dataset is in ratio 2:1 and can be treated as balanced
* Outcome is good correlated with Glucose
* Age and pregnancies are highly correlated
* Insulin and skin thickness is correlated
* BMI and skin thickness is good correlated
* As dataset wasn’t large enough and so many missing value as of now I have omitted to drop feature some libraries that can be used are feature engine: smart\_correlation

**Week 3**

* **under sampling our majority class to increase precicision of calculating minority as thats more imp have choosen RandomUnderSample for this**

**Week 4**

* **\*\*Our main objective was to predict more person that can get diabetic and we could predict 198 out of 268 \*\***