1) a) Data in olden days are not in sufficient quantity. So dividing the data as 60% train, 20% test and 20% validation would have made sense. But with modern day, data is abundant and splitting as 20% for each validation and test might not be a good fit as that itself might be a huge chunk.

So, we can decide on the split based on the availability of data.

b) i) Overfitting - C

ii) Underfitting - A

iii) Ideal model capacity - B.

c) iii) Increase the depth of decision tree.

3) a) a) $P(\text{Malaria}) = 0.01$, $P(\text{Not Malaria}) = 1 - P(\text{Malaria}) = 0.99$
(M) (NM)

$P(\text{Test positive} \mid \text{Malaria}) = 0.95$
(TP) (M)

$P(\text{Test Positive} \mid \text{Not Malaria}) = 0.05$
(TP) (NM)

a) $P(\text{Test Positive}) = P(\text{Malaria}) * P(TP/M) + P(\text{not malaria}) * P(TP/NM)$

$= 0.01 * 0.95 + 0.99 * 0.05$

$= 0.0095 + 0.0495$

$P(\text{Test positive}) = 0.059$

3) a)
b)

$$P(\text{Malaria} / TP) = \frac{P(M) * P(TP/M)}{P(TP)}$$

$$= \frac{0.95 \times 0.01}{0.059}$$

$$= \frac{0.0095}{0.059}$$

$$P(\text{Malaria}/TP) = 0.161$$

3) b) $P(\text{rain today}) = 0.30$

$P(\text{rain tomorrow}) = 0.60$

$P(\text{rain today and tomorrow}) = 0.25$

$$P(\text{rain tomorrow} \mid \text{rain today}) = \frac{P(\text{rain today} \,\&\, \text{tomorrow})}{P(\text{rain today})}$$

$$= \frac{0.25}{0.30}$$

$$P(\text{rain tomorrow} \mid \text{rain today}) = 0.833$$

3) c)

i) $P(odd) = P(1) + P(3) + P(5)$

$= 0.2 + 0.1 + 0.1$

$P(odd) = 0.4$

In a fair die with equal probabilities for each face, the probabilities will be

| face | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| P(face) | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

In this case,

$P(odd) = P(1) + P(3) + P(5)$

$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$

$= \frac{3}{6}$

$P(odd) = \frac{1}{2} = 0.5$

Hence for a fair die with equal probability on each face, the probability is exactly half of actual probability ($\frac{1}{2} = 0.5$). But for a biased die, this probability for each face varies, hence the probability is less than a fair die, which is 0.4.

3) c) ii)

$$\text{Entropy}[H(X)] = -\sum_{x \in \Omega_X} P(x) \log_e(P(x))$$

$$H(X) = -\Big[P(1)\log_e(P(1)) + P(2)\log_e(P(2)) + P(3)\log_e(P(3)) + P(4)\log_e(P(4)) + P(5)\log_e(P(5)) + P(6)\log_e(P(6))\Big]$$

$$= -\Big[0.2 \times \log_e(0.2) + 0.1 \times \log_e(0.1) + 0.1 \times \log_e(0.1) + 0.2 \times \log_e(0.2) + 0.1 \times \log_e(0.1) + 0.3 \times \log_e(0.3)\Big]$$

$$= -\Big[0.2(-1.609) + 0.1(-2.30) + 0.1(-2.30) + 0.2(-1.609) + 0.1(-2.30) + 0.3(-1.20)\Big]$$

$$= -[0.3218 - 0.23 - 0.23 - 0.3218 - 0.23 - 0.36]$$

$$= -[-1.69]$$

$$H(X) = 1.69$$

3) For a fair die, the probability for each face is $1/6$

$$H(X) = -\Big[P(1)\log_e(P(1)) + P(2)\log_e(P(2)) + P(3)\log_e(P(3)) + P(4)\log_e(P(4)) + P(5)\log_e(P(5)) + P(6)\log_e(P(6))\Big]$$

$$= -\Big[\cancel{6} \times \log_e\left(\tfrac{1}{6}\right) \times \tfrac{1}{\cancel{6}}\Big]$$

$$H(X) = -[-1.789] = 1.789$$

**2.1**

**i)**

**Niger had the highest child mortality rate in 1990. The rate was 313.7**

**Let us assume data contains the unicef's data**

**data[data['Under-5 mortality rate (U5MR) 1990']== [data['Under-5 mortality rate (U5MR) 1990'].max()]**

**ii)**

**Sierra Leone had the highest child mortality rate in 2011. The rate was 185.3**

**Let us assume data contains the unicef's data**

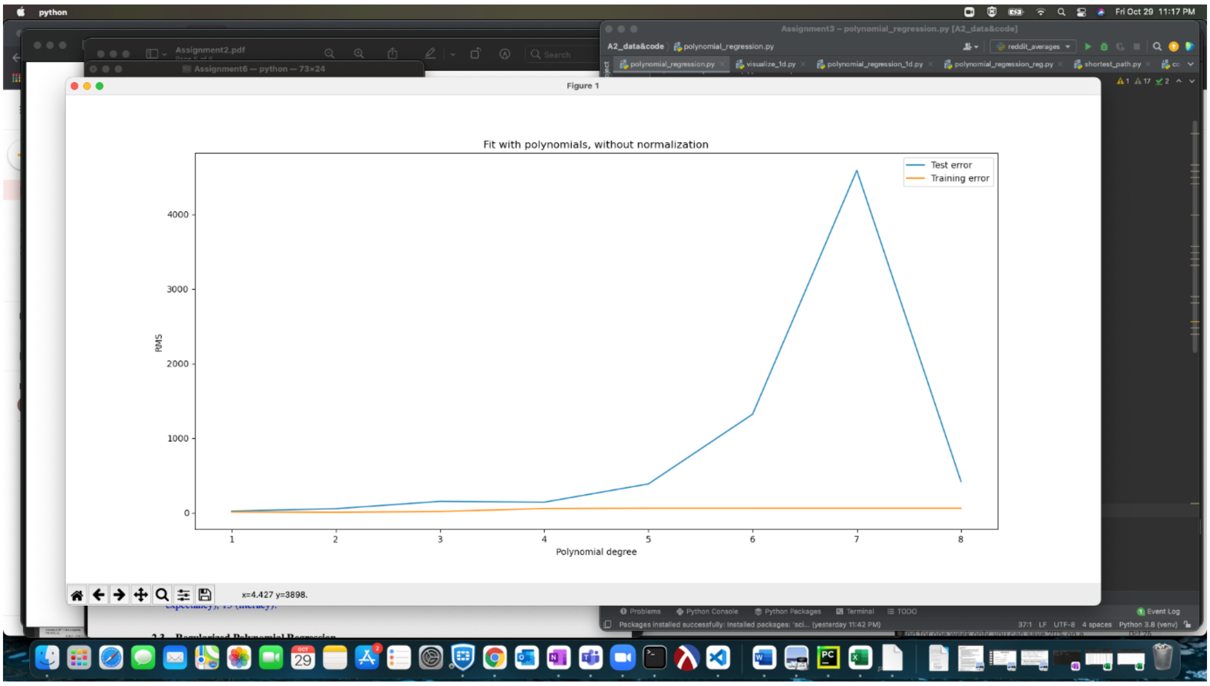
**data[data['Under-5 mortality rate (U5MR) 2011 ']== [data['Under-5 mortality rate (U5MR) 2011'].max()]**

**iii)**

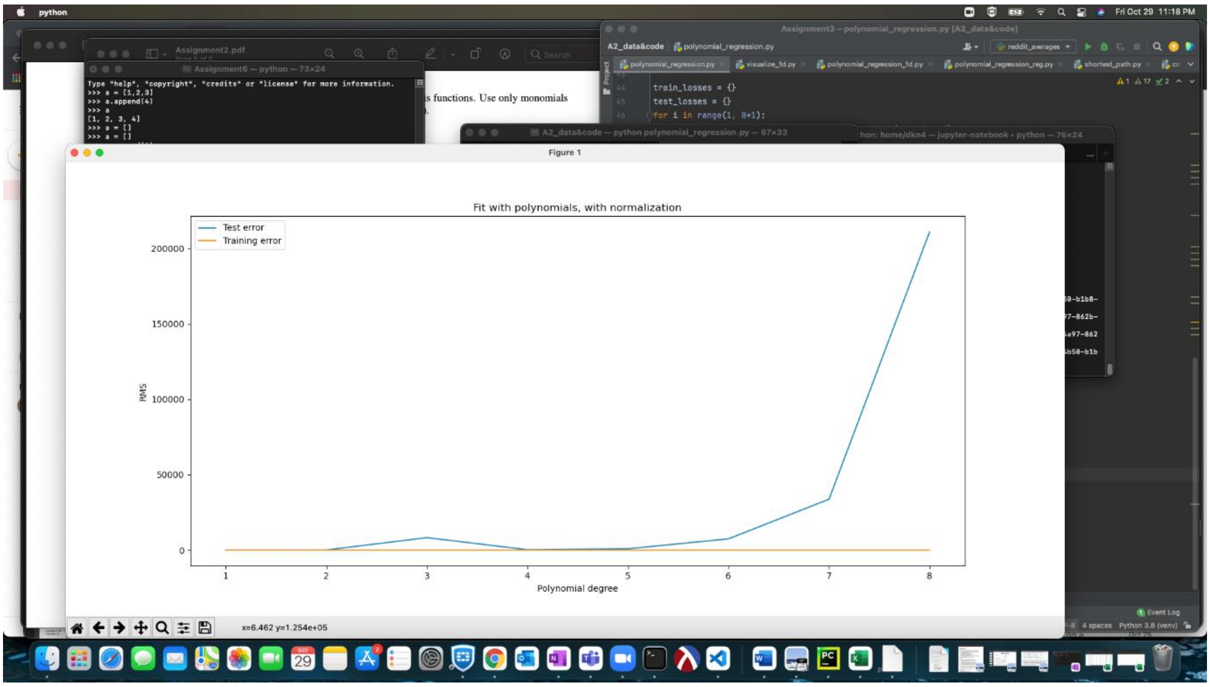
**Yes, there are some missing features in the original csv spreadsheet. This is handled in the function assignment2.load_unicef_data() by calculating the mean values for each column and then replacing the missing values with the mean value**

## 2.2

### a)

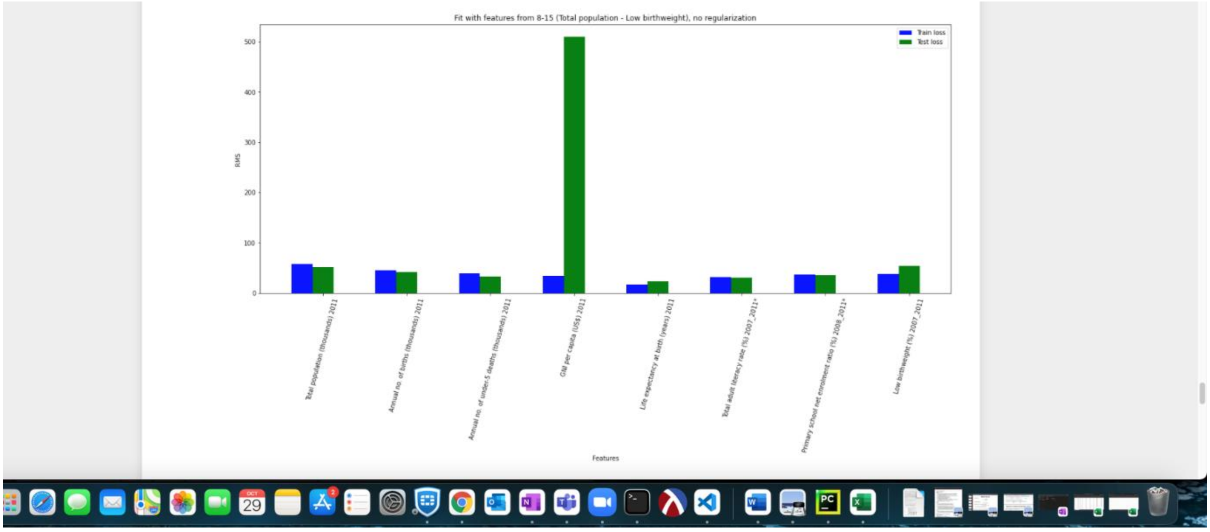



**The input data(features) x is not normalized, so the output produced will not be accurate.**

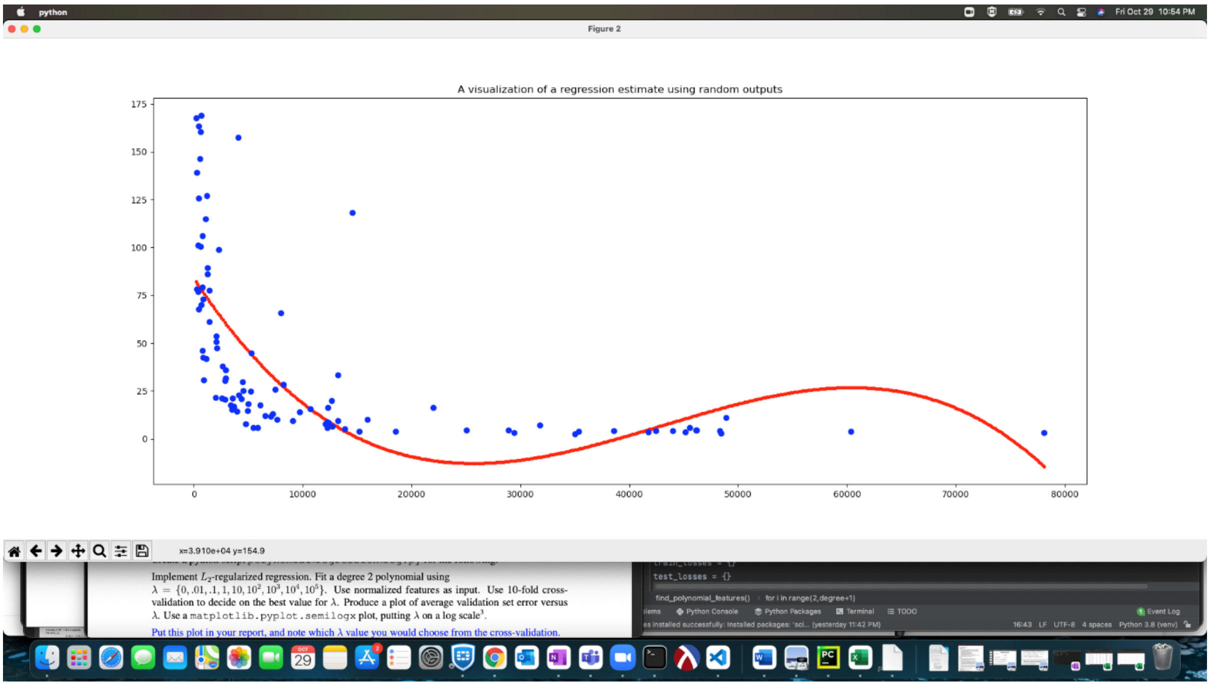


## 2.2

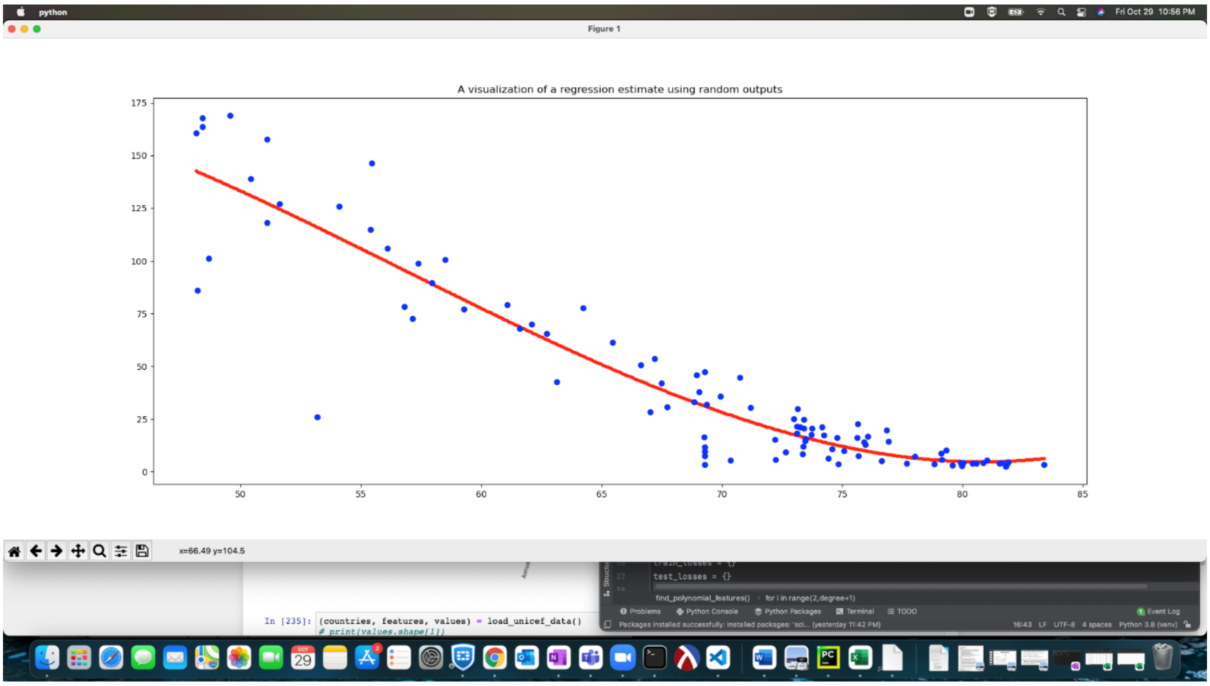
## b) Bar chart



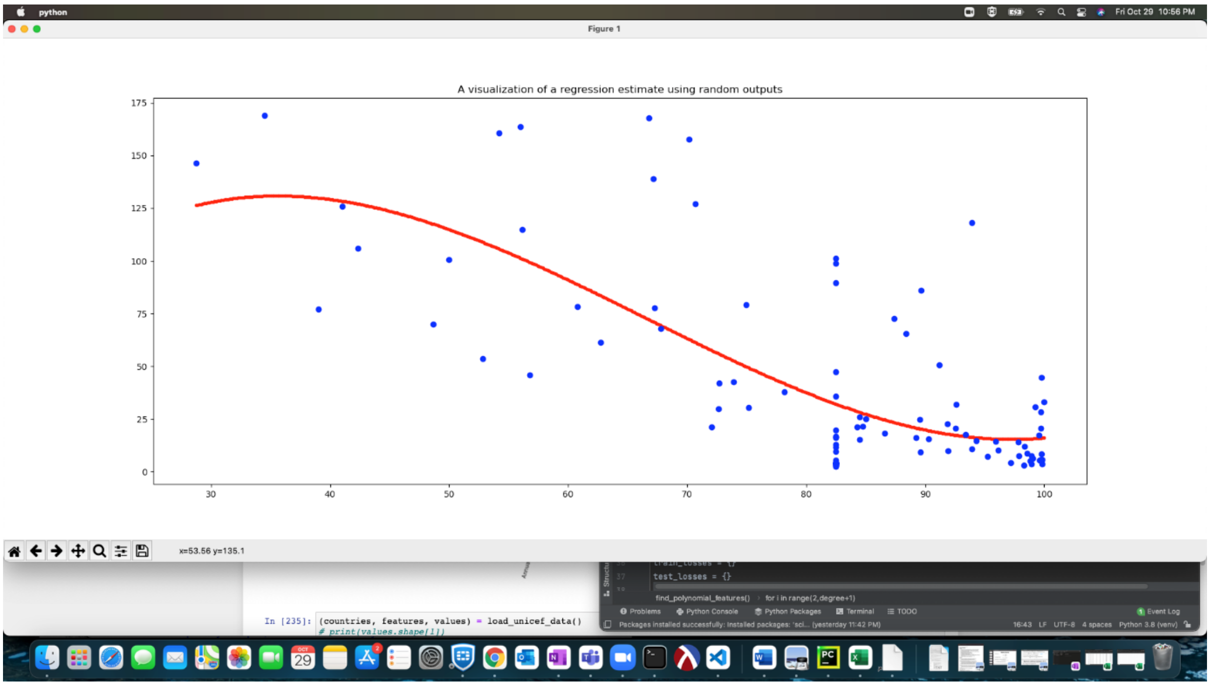

## Feature 11 GNI



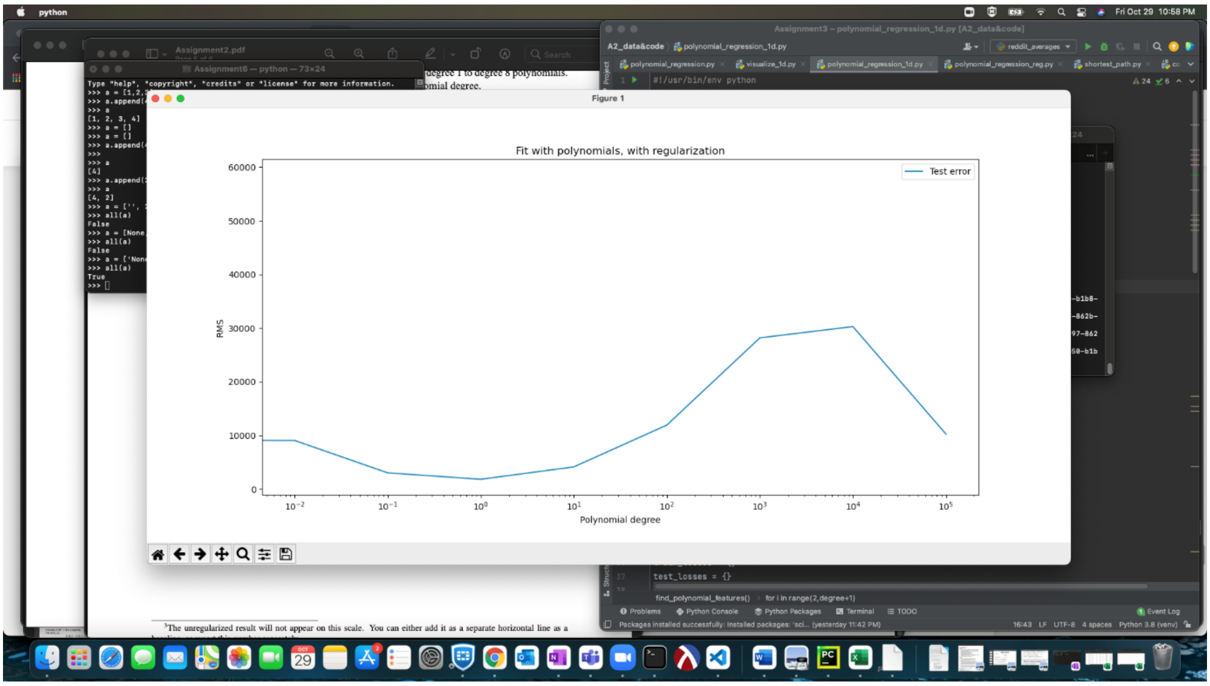
## Feature 12 Life expectancy




## Feature 13 Literacy



## 2.3




**I will choose a lambda value of 10^0 from the cross validation as the RMS value is low.**