Section A

a)

$\overline{X}$ → mean of independent variable

$\overline{Y}$ → mean of dependent variable

To proof : $(\overline{X}, \overline{Y})$ point lies on linear regression line.

In least squares we minimze the L2 loss.

i.e

$$J(w) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$\hat{y}_i$ → predicted label

y_i → true label

$$\hat{y}_i = w^T x_i$$

$$\frac{J(w)}{\partial w_0} = \frac{\partial \sum (w_0 + w_1 x_i - y_i)^2}{\partial w_0} = 0$$

$$\Rightarrow 2\sum(w_0 + w_1 x_i - y_i) = 0$$

$$\Rightarrow \sum(w_0 + w_1 x_i - y_i) = 0$$

$$\Rightarrow w_0 \sum_{i=1}^{n} + w_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0$$

$$\Rightarrow w_0 n = \sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i$$

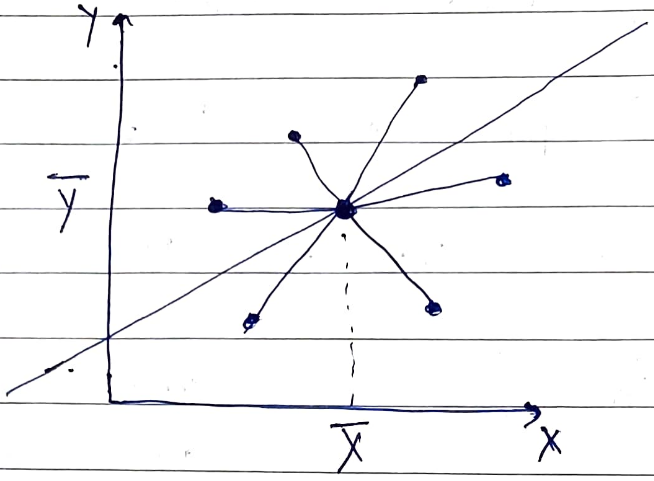
$$\Rightarrow w_0 = \frac{\sum y_i}{n} - w_1 \frac{\sum x_i}{n}$$

$$\Rightarrow \boxed{w_0 = \overline{Y} - w_1 \overline{X}}$$ ← This proves that the line pass through $(\overline{X}, \overline{Y})$

where $\overline{X} = \frac{\sum x_i}{n}$ ⇒ arithmetic mean of x_i

$\overline{Y} = \frac{\sum y_i}{n}$ ⇒ arithmetic mean of y_i

Similarly, we can prove for any no. of feature / degree



b)

$$\text{Correlation}(X,Y) = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sigma_x \sigma_y}$$

→ Let A, B, C be 3 Random variable A is positively related to B & B is positively related to C. Does that imply A is the related to C.

$$C(A,B) = \frac{\sum (a_i-\bar{a})(b_i-\bar{b})}{\sigma_a \sigma_b} \quad -(\text{I})$$

$$C(B,C) = \frac{\sum (b_i-\bar{b})(c_i-\bar{c})}{\sigma_b \sigma_c} \quad -(\text{II})$$

$$C(A,C) = \frac{\sum (a_i-\bar{a})(c_i-\bar{c})}{\sigma_a \sigma_c} \quad -(\text{III})$$

Using (I) (II) & (III)

$$C(A,B) = C(B,C) - C(A,C) - \sqrt{(1-C(B,C)^2) * (1-C(A,C)^2)}$$

• $C(A,B) > 0$, we have assumed that A & B are positively related.

$\Rightarrow$

$$C(B,C) * C(A,C) > \sqrt{(1-C(B,C)^2)(1-C(A,C)^2)}$$

Let, $p = C(B,C)$

$q = C(A,C)$

$$pq > \sqrt{(1-p^2)(1-q^2)}$$

Sq. both side : $p^2q^2 > (1-p^2)(1-q^2)$

: $p^2q^2 > 1 - p^2 - q^2 + p^2q^2$

: $\boxed{p^2 + q^2 > 1}$

$p = C(B,C) > 0$

Let $p = 1$

$\boxed{-1 \le q \le 1}$ true

$1 + q^2 > 1$

$\Rightarrow \boxed{q^2 > 0}$

q is $C(A,C)$

This eq. is true even if $q < 0$

q can belong to any value b/w -1 to 1.

$-1 \le q \le 1$: This prove that even if $C(A,B)$ & $C(B,C)$ is 1 $C(A,C)$ does not necessary be positive it can be $-1, 0, 1$ anything.

Example:

The more amount of time a person sit on his study table the more no. of pages of a book he can read.

The better eye sight a person has then also he ~~will~~ can read more no. of pages of the book.

But the amount of time he spend on study table does not necessarily improve his eye-sight.
i.e. they are not positively related.

c) Proof WLLN (Weak law of large Numbers)

- Suppose $Y_1, Y_2 \dots Y_n$ are independent and identical Random variables, each with finite mean. $E[Y_i] = \mu$ and $Var(Y_i) = \sigma^2$.

- WLLN states that the Sample average X_n of $Y_1, Y_2 \dots Y_n$ random variables converges in probability to μ.

i.e $X_n \to \mu$ $n \to \infty$

or $\lim_{n \to \infty} [P(|X_n - \mu| \geq \varepsilon)] = 0$, $\forall$ $\varepsilon > 0$

$$X_n \to \mu$$

$$\boxed{X_n = \frac{\sum Y_i}{n}}$$

- Chebyshev's Inequality

$$P(|T - E[T]| \geq \varepsilon) \leq \frac{Var(T)}{\varepsilon^2}, \quad \forall \varepsilon > 0$$

$$\Rightarrow P(|X_n - E[X_n]| \geq \varepsilon) \leq \frac{Var(X_n)}{\varepsilon^2}, \quad \forall \varepsilon > 0$$

$$E[X_n] = E\left[\frac{1}{n} \sum Y_i\right] = \frac{1}{n} \sum E[Y_i] = \frac{1}{n} \times n\mu = \mu$$

$$Var(X_n) = Var\left[\frac{1}{n}\sum Y_i\right] = \frac{1}{n^2} Var\left(\sum Y_i\right) = \frac{1}{n^2}\sum Var(Y_i) = \frac{n\sigma^2}{n^2}$$

$$Var(X_n) = \frac{\sigma^2}{n}$$

$$P(|X_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}, \quad \forall\ \varepsilon > 0$$

$$\lim_{n\to\infty} P(|X_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty}\left(\frac{\sigma^2}{n\varepsilon^2}\right) = 0$$

$$\Rightarrow \quad \lim_{n\to\infty} P(|X_n - \mu| \geq \varepsilon) = 0$$

this implies X_n converges to μ as n tends to ∞

$$\boxed{X_n \longrightarrow \mu}$$

Example: Bernoulli distribution

$$P(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

Expected value $\Rightarrow$ $p(1) + (1-p)(0) = p$

Let X: be a random that counts no of heads in a coin toss

Sample space $X = \{0, 1\}$

PMF $=$ $P(X) = \begin{cases} \frac{1}{2}, & X = 0 \\ \frac{1}{2}, & X = 0 \end{cases}$

Expected value $= \frac{1}{2}(0) + \frac{1}{2}(1) = \frac{1}{2}$

Psedo code / Code.

→ Code submitted in a collab notebook.

d) MAP Estimate for Linear Regression

In MAP we maximise the posterior function $P(w/D) \propto P(w/x,y)$
This encodes the info from likelihood & prior distribution.

$$P(w/D) = \frac{\overbrace{P(D/\omega)}^{\text{likelihood}} P(\omega)}{\underbrace{P(D)}_{\text{Constt. can be ignored.}}}$$

Log transformation

$$\log P(w/D) = \log(P(D/w)) + \log P(\omega) \quad -①$$

Prior $P(\omega)$ follow gaussian distribution with 0 mean

i.e $\quad P(\omega) = N(0, \alpha^2 I) \quad$ where $\alpha^2 I$ Covariance matrix

$P(D/\omega)$ or $P(X,Y/\omega)$ assume follows a gaussian distribution

$$P(X,Y/w) = N(y, w^T x, \sigma^2 I)$$

Sample x_i, y_i are independent of each other.

$$P(D/\omega) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \exp\left\{ \frac{-1}{2} \frac{(y_i - w^T x_i)^2}{\sigma^2} \right\}$$

Plugging the distributions into log equation ①

$$\max(\log P(w/D)) = -\min(\log P(w/D))$$

$$-\min(\log P(w/D)) = -\log P(D/\omega) - \log P(\omega)$$

$$\Rightarrow -\sum \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}}\right) - \sum \ln\left(\frac{1}{\sqrt{2\pi\alpha^2}} e^{-\frac{(w^T w)}{2\alpha^2}}\right)$$

$$= -\sum \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum \frac{(y_i - w^T x)}{2\sigma^2} + \frac{\|w\|^2}{\alpha^2}$$

$$W_{MAP} = \min_{w} \left[\sum_{i=1}^{n} \frac{(y - w^T x)^2}{\sigma^2} + \frac{w w^T}{\alpha^2} \right]$$

- We can get the optimal weigths by minimising this cost. As we can see this cost function resembles, L2 regularisation.

We can further solve the eq by differentiating wrt to w.

$$\frac{\partial(\ln P(w/D))}{\partial w} = \frac{-1}{\sigma^2}(w^T X^T x - y^T X) + \frac{1}{\alpha^2} w^T = 0$$

$\Rightarrow$

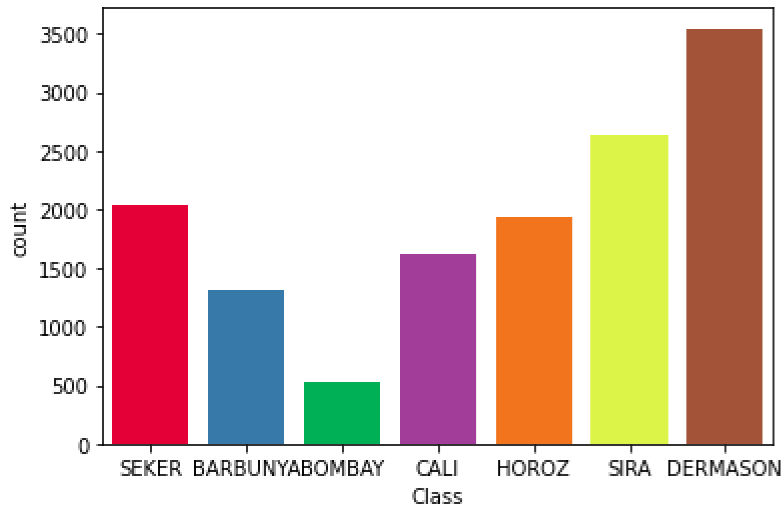
$$\boxed{W_{MAP} = \left(X^T X + \frac{\sigma^2}{\alpha^2} I\right)^{-1} X^T y}$$

Normal equation for MAP.

SECTION - C

a) Class distribution

<matplotlib.axes._subplots.AxesSubplot at 0x7fbe31799e90>



Analysis:-
Bean DERMASON has high-class distribution than all the other classes. Whereas the BOMBAY class has the lowest distribution.

b) Data Insights:-

- There is a total of 16 features in the data and 7 classes in the given data.
  16 features:- { 'Area', 'Perimeter', 'MajorAxisLength', 'MinorAxisLength','AspectRation', 'Eccentricity', 'ConvexArea', 'EquivDiameter', 'Extent','Solidity', 'roundness', 'Compactness', 'ShapeFactor1', 'ShapeFactor2','ShapeFactor3', 'ShapeFactor4' }

  7 Classes :- {'SEKER', 'BARBUNYA', 'BOMBAY', 'CALI', 'HOROZ', 'SIRA', 'DERMASON'}

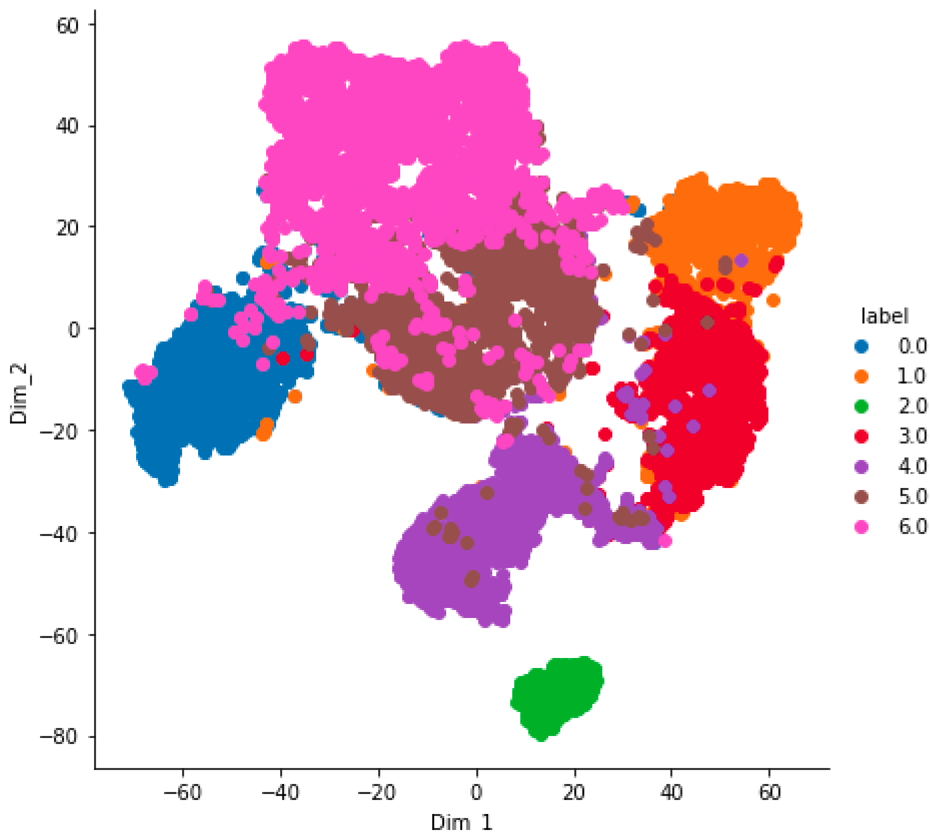
- There are no null values in the data. I.e there is no rows and column that has null values

- Class distribution is not equal. The highest class count is 3500 whereas, the lowest class count is 500, which means there is very high variation in class distribution.

- Some features are highly correlated and some are negatively correlated. For E.g. Compactness and Aspect Ratio are most negatively related with a correlation of -0.984687, and Area and Convex are positively related with a correlation of +0.9999

- Box plot show that most of the beans have Area in the range of 50000 and aspect ratio are in the 0.0006 to 0.0007

c)

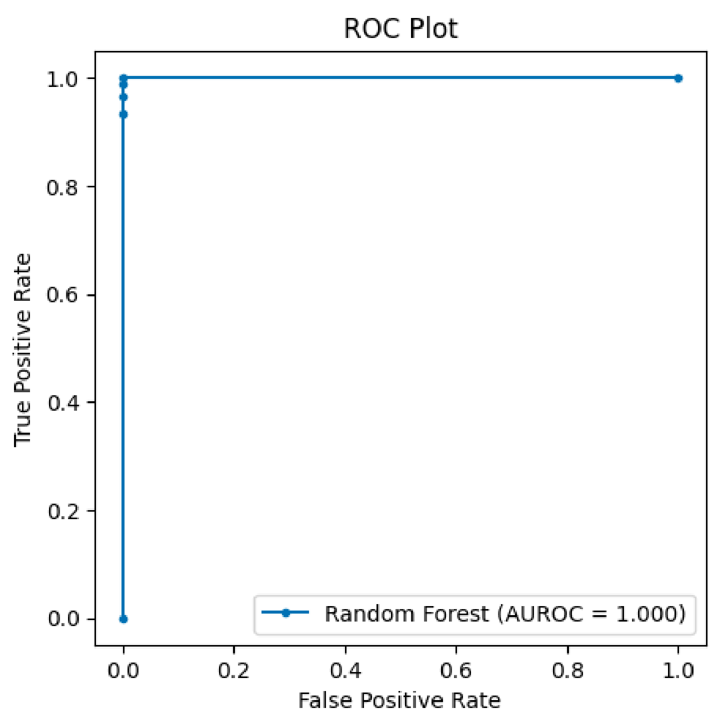


TSNE has reduced the no of features from 16 to 2 dimensions. As we can see from the plot above the data has been successfully separated successfully. There are a few overlapping but mostly data are well separated.

d) With the Naive Byes Model we have got an accuracy of 89.7 per cent whereas with the Mulitnomila Byes Model we have got an accuracy of 56.12 per cent.
Naive works better because it’s continuous data and naive model works well on continuous data, whereas Mulitnominal Byes model is mostly used in NLP for Text feature Classification.
Here the nature of continuous data and classification problem suits well for Naive Byes.

e) As we are preserving more and more variance for the training, the accuracy of the model is also increasing. This is because with more variance our model is capturing more information about the data and is able to predict well. As we can see in the code we got an accuracy of 87 % by preserving 90% variance and an accuracy of 90% by preserving 99% variance.

f) As this is a multiclass classification problem, I have used the One vs Rest approach for plotting the ROC-AUC curve. There are seven ROC-AUC curves because there are seven classes.
We have got the perfect model for when we have plotted for BOMBAY vs REST



g) With the logistics Regression model, we have got an accuracy of 93% whereas with Naive byes we have got an accuracy of 89%.

Parameters:- `LogisticRegression(multi_class='multinomial', solver='lbfgs')`

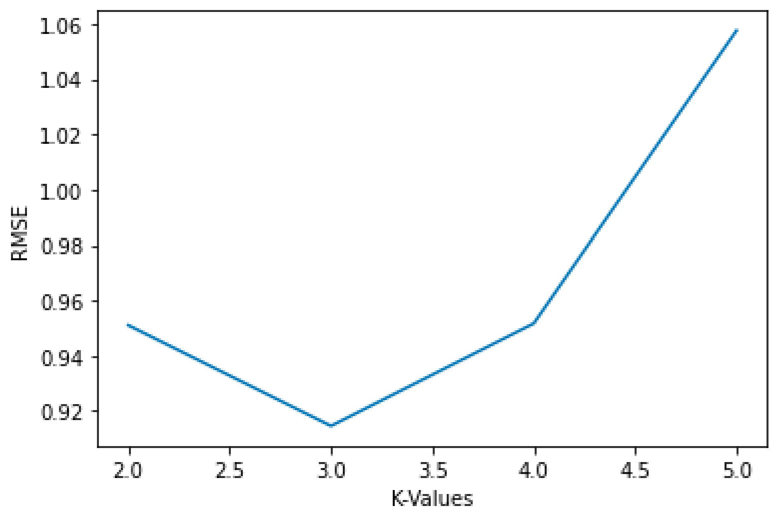
I have used mulit_class = “multinomial” because this is a multiclass classification model not a binary class. There are 7 classes. Logistic Regression is basically binary classification but for multiclass we need to use appropriate parameters. **Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm (LBFGS) as the solver this is basically a gradient decent technique to find the optimal minima. Similar to Newton Gradient decent.**

SECTION - B

a) K - FOLD CROSS VALIDATION
TABLE FOR RMSE VS K VALUES

K_VALUES	RMSE
2	0.9510598189251254
3	0.9146036624484157
4	0.9517185450481481
5	1.0576021326153915

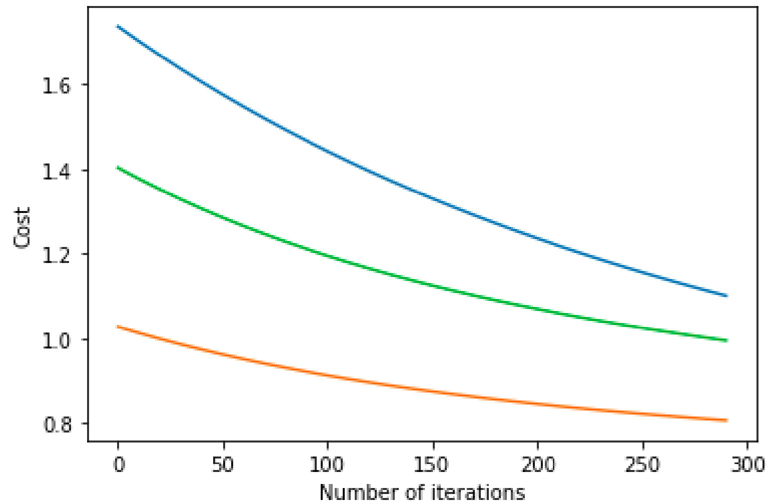
Text(0, 0.5, 'RMSE')



In this case, the best value of k is 3. Because RMSE is lower when k is 3.

b)

RMSE V/s iteration graph for all models trained with the optimal value of K for K-Fold cross-validation.

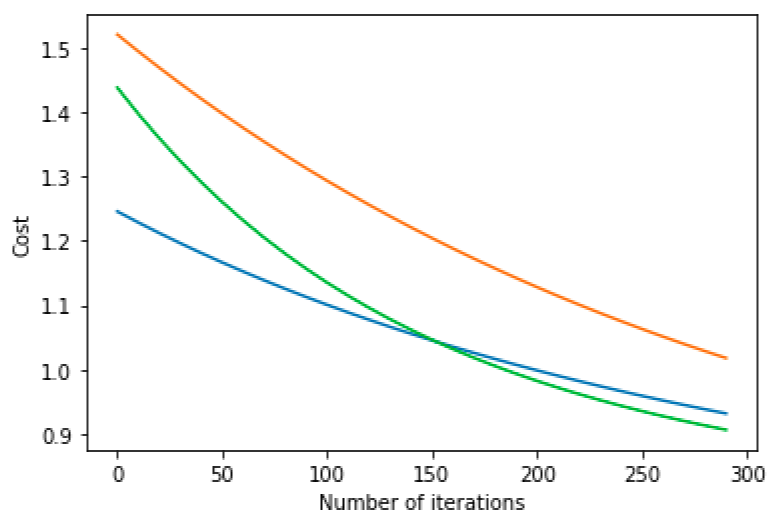


c)

- **L2 REGULARISATION**

I have tried L2 Regularisation for 5 different parameters of lamda
lamda_L2 = [0.001, 0.01, 0.1, 1, 2]

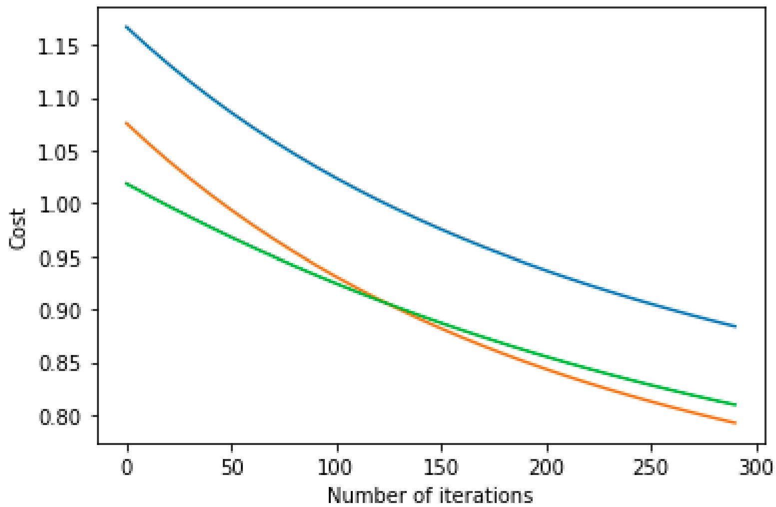
In this case lmd = **0.001** is the giving minimum rmse
As the value of lamda starts increasing the rmse value also increases.
This is because as lamda increases the value to weights and complexity of the model also increases leading to overfitting and more loss



- **L1 REGULARSATION**

I have tried L1 Regularisation for 5 different parameters of lamda
lamda_L1 = [0.001, 0.01, 0.1, 1, 2]

The best lmd = **0.001** is the giving minimum rmse
Giving minimum rmse of 0.8



d) Applying NORMAL EQUATION FOR LINEAR REGRESSION DIRECTLY(MLE)

Tr AS we can see from the table and graph RMSE for the validation set.
This graph is for the best_k, for each validation set of best_k

```
Rmse on validation set  0.5983683841153545
Rmse on validation set  0.7332447079473199
Rmse on validation set  0.5754223830290294
Text(0, 0.5, 'Cost / RMSE')
```