1.a;1, The Likelihood function of Beta function with α unknown and $\beta = 1$, is given by

Likelihood $L(\alpha) = \prod_{i=1}^{N} x_i^{\alpha-1} * \alpha$

Taking Log likelihood , $l(\alpha) = \sum_{i=1}^{N} \log (x^{\alpha-1} * \alpha)$

Taking first derivative of $l(\alpha)$ we get $\sum_{i=1}^{N} (\log x_i) + \frac{N}{\alpha}$

Equating this to zero we get $\alpha = -\frac{N}{\sum_{i=1}^{n} \log x_i}$

1.a;2, Normal distribution of $N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\ e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

So, $N(\theta, \theta) = \frac{1}{\sqrt{2\pi\theta}}\ e^{\frac{-(x-\theta)^2}{2\theta}}$

Maximum Likelihood Estimation, $L\big(N(\theta, \theta)\big) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\theta}}\ e^{\frac{-(x-\theta)^2}{2\theta}}$

Taking Log likelihood we get, $l(\theta, \theta) = \sum_{i-1}^{N}(-\log\sqrt{2\pi\theta}\,) + \frac{-(x_i-\theta)^2}{2\theta}$

Simplifying this we get, $-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\theta - \sum_{i=1}^{N}(\frac{x_i^2}{2\theta} + \frac{\theta}{2} - x_i\,)$

Taking the first derivative and equating it to zero we get, $n\theta - \sum_{i-1}^{N} x_i^2 + n\theta^2 = 0$

Solving this we get, $\theta = \frac{-n \pm \sqrt{(n^2 + 4n\sum_{i=1}^{N} x_i^2)}}{2n}$

1.b, Given $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{N} \frac{1}{h} K(\frac{x - X\_i}{h})$

Taking Estimation on both sides we get,

$$E\left(\hat{f}(x)\right) = E\left(\frac{1}{n}\sum_{i=1}^{N}\frac{1}{h}K(\frac{x - X_i}{h})\right)$$

Considering the linearity of Expectation we get R.H.S. = $\frac{1}{hn}\sum_{i=1}^{N} E\left(K\left(\frac{x-X_i}{h}\right)\right)$

So finally we get R.H.S= $\frac{1}{h}\int f(x) * K\left(\frac{x-X_i}{h}\right)$. Hence Proved.

Using Taylor's series we expand, $f\ (x - hz)$ so we get

$$f(x-hz) = f(x) - hzf'(x) + \frac{1}{2}h^2z^2f''(x) + o(h^2)$$
$$E\left(\hat{f}(x)\right) = f(x) + \frac{1}{2}h^2f''(x)\int z^2k(z)dz + o(h^2)$$
$$where\ \int k(z)dz = 1, \int zk(z)dz = 0$$
$$bias = E\left(\hat{f}(x)\right) - f(x) = \frac{1}{2}\ h^2\mu_2(K)f''(x) + o(h^2), where\ \mu_2(K) = \int z^2k(z)dz\ .$$

2.a, Using the equation = $\frac{x_i - \bar{x}}{\sigma}$ we normalized the data.

Normalized data

Major	X-coordinate	Y-coordinate
Mathematics	0.0623	1.1070
Mathematics	0.9163	0.6928
Mathematics	-0.7829	1.0317
Electrical Engineering	0.9074	-0.0226
Electrical Engineering	1.0409	0.4292
Electrical Engineering	1.2633	0.6928
Computer Science	-0.0267	-0.3991
Computer Science	0.9519	-1.7922
Computer Science	-1.1832	-1.4534
Computer Science	-1.3167	-0.2862

Student coordinate classification [9 18], so normalizing that we get [0.0178 -0.0602]
Classification using L2 norm
For K=1,classified major is Computer Science.
For K=3,classified major is Electrical Engineering.
Classification using L1 norm
For K=1 classified major is Computer Science.
For K = 3 classified major is Computer Science. We selected computer science for this this because we got three unique values for major and computer science had the highest probability among them.

2.b;1, $P(x) = \sum_c P(x|y=c) * P(y=c)$

so, $P(x) = \sum_c \frac{K_C}{N_C * V} * \frac{N_c}{N}$
Therefore, $P(x) = \frac{K}{N*V}$ as $\sum_c K_c = K$.

2b;2, Using Bayes rule, $P(y=c|x) = \frac{P(x|y=c)*P(y=c)}{P(x)}$

$$=> p(Y=c|x) = \frac{K_c}{NcV} * \frac{N_c}{N} * \frac{NV}{K} = \frac{K_c}{K}$$

3a; $H(\#\text{Rainy by }\#\text{Observations}) = E(36,44)$

$$-\frac{36}{80}\log\left(\frac{36}{80}\right) - \frac{44}{80}\log\left(\frac{44}{80}\right) = 0.15605 + 0.1428 = 0.2988$$

$$H(Temperature = hot) = -\left(\frac{23}{40}\log\frac{23}{40} + \frac{17}{40}\log\frac{17}{40}\right) = 0.2960$$

$$H(Temperature = cool) = -\left(\frac{13}{40}\log\frac{13}{40} + \frac{27}{40}\log\frac{27}{40}\right) = 0.2738$$

$$Weighted\ Average = \frac{4}{8}H(Temperature = hot) + \frac{4}{8}H(Temperature = cool)$$

$$Weighted\ Average = H(Temperature) = \frac{4}{8} * 0.2960 + \frac{4}{8} * 0.2738 = 0.2849$$

$Information\ Gain_{Temperature}$ =
$H(\#\text{Rainy by }\#\text{Observations}) - H(Temperature)$

$$Information\ Gain_{Temperature} = 0.2988 - 0.2849 = 0.0139$$

$$H(Humidity = high) = -\left(\frac{23}{40}\log\frac{23}{40} + \frac{17}{40}\log\frac{17}{40}\right) = 0.2960$$

$$H(Humidity = low) = -\left(\frac{13}{40}\log\frac{13}{40} + \frac{27}{40}\log\frac{27}{40}\right) = 0.2738$$

$$Weighted\ Average = \frac{4}{8}H(Humidity = high) + \frac{4}{8}H(Humidity = low)$$

$$Weighted\ Average = H(Humidity) = \frac{4}{8} * 0.2960 + \frac{4}{8} * 0.2738 = 0.2849$$

$$Information\ Gain_{Humidity} = H(\#\text{Rainy by }\#\text{Observations}) - H(Humidity)$$

$$Information\ Gain_{Humidity} = 0.2988 - 0.2849 = 0.0139$$

$$H(Sky\ condition = cloudy) = -\left(\frac{25}{40}\log\frac{25}{40} + \frac{15}{40}\log\frac{15}{40}\right) = 0.2873$$

$$H(Sky\ condition = clear) = -\left(\frac{11}{40}\log\frac{11}{40} + \frac{29}{40}\log\frac{29}{40}\right) = 0.25543$$

$$Weighted\ Average = \frac{4}{8}H(Sky\ condition = cloudy) + \frac{4}{8}H(Sky\ condition = clear)$$

$$Weighted\ Average = H(Sky\ condition) = \frac{4}{8} * 0.2873 + \frac{4}{8} * 0.2554 = 0.2713$$

$$Information\ Gain_{Humidity} = H(\#\text{Rainy by }\#\text{Observations}) - H(Sky\ condition)$$

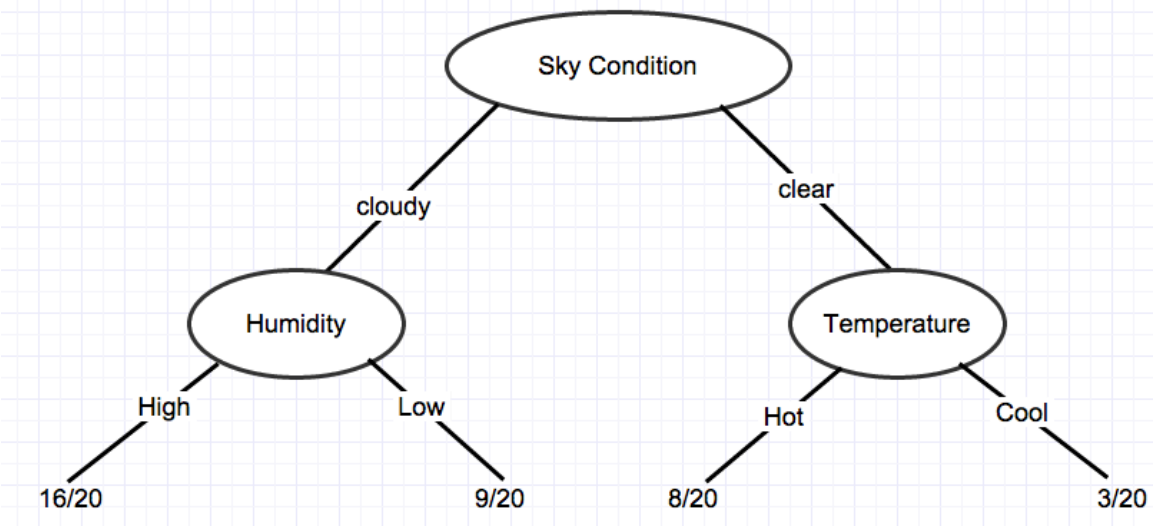
$$Information\ Gain_{Humidity} = 0.2988 - 0.2713 = 0.0275$$

Since the Information gain of 'Sky Condition' is highest, we select 'Sky Condition' as first feature to split.

For 'Sky Condition = cloudy' , Information gain of 'Humdity' is highest, hence we select 'Humidty' feature to split.

For 'Sky Condition = clear' Information gain of 'Temperature is highest, hence we select 'Temperature feature to split.

So, the final decision tree we got looks like,



3b;Given, $Gini\ Indiex = \sum_{k=1}^{K} p_k(1-p_k)$, $Cross\ Entropy = -\sum_{k=1}^{K} p_k log p_k$

We need to prove that Gini Index <= Cross Entropy i.e. $\sum_{k=1}^{K} p_k(1-p_k) \leq -\sum_{k=1}^{K} p_k log p_k$

Using the concept of calculus we can just prove that $p*(1-p)+p\log p \leq 0$ to prove the above inequation.
Since the probability is greater than 0 we can divide by p on both sides.

Let,

$$f(p) = (1-p+\log p)$$

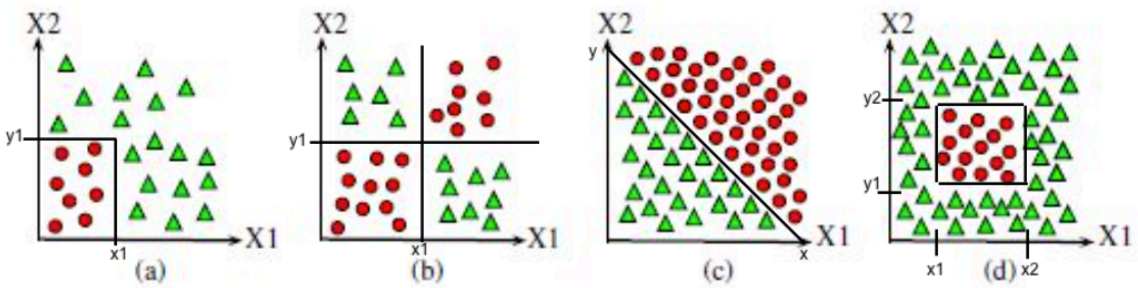
differentiating we get $f'(p) = -1 + \frac{1}{p}$

equating it to 0 we get p=1

$$f(p=1) = 0$$

So for every other value of p $f(p)$ will be negative.
Hence $f(p) \leq 0$ for all 0<p<1. This proves the above inequation.

3.c;



Using less than and greater than inequalitites we could classify the graphs a, b, d graphs with a depth less than 6.

4.a The likelihood function is given by

$$L = \prod_{i=1}^{n} p(X_i|Y_i) * p(Y_i)$$

$p(X_i|Y_i) = \prod_{k=1}^{D} p(x_k|y_i) * p(y_i), where\ p(y_i) = p_i,\ p(x_k|y_i) = N(\mu_{ik}, \sigma_{ik})$

$L(p_i, \mu_{ik}, \sigma_{ik}) = \prod_{i=1}^{n} \prod_{k}^{D} p(x_k|y_i) * p(y_i)$

Taking log on both sides we get

$l(p_i, \mu_{ik}, \sigma_{ik}) = \sum_i \sum_k \log\left(p(x_{ik}|y_i)\right) + \sum_i log p_i\ , where\ p(x_{ik}|y_i) = N(\mu_{ik}, \sigma_{ik})$

$$\sum_i log p_i$$

$$= \sum_k c_k p_k\ , where\ c_k is\ the\ number\ of\ times\ k\ th\ class, appears\ in\ sample$$

We have to use Lagrange's multiplier with the constraint that $\sum_i p_i = 1$

$\frac{dl(p_i, \mu_{ik}, \sigma_{ik})}{dp_i} = c_k \sum_k \frac{1}{p_k} + \lambda * k = 0$

$\sum_k p_k = 1, we\ get\ \lambda = -\frac{n}{k}, substituting\ this\ in\ the\ above\ equation\ we\ get\ p_k = \frac{c_k}{n}$

4.b; (b) $P(Y = 1|X) =$

$$\frac{P(X|Y = 1) * P(Y = 1)}{P(X|Y = 1) * P(Y = 1) + P(X|Y = 0) * P(Y = 0)}$$

$$= \frac{p(Y = 1) * \prod_i p(x_i|Y = 1) * p(x_i)}{p(Y = 1) * \prod_i p(x_i|Y = 1) * p(x_i) + p(Y = 0) * \prod_i p(x_i|Y = 0) * p(x_i)}$$

Dividing by $p(Y=1) * \prod_i p(x_i|Y=1) * p(x_i)$ *in* numerator and denominator we get,

$$= \frac{1}{1 + \left(\prod_i \frac{p(x_i|y=0) * p(y=0)}{p(x_i|y=1) * p(y=1)}\right)}$$

$$Let, \prod_i \frac{p(x_i|y=0) * p(y=0)}{p(x_i|y=1) * p(y=1)} = f(\theta) = \left(\frac{1-\pi}{\pi}\right) \prod_{i=1}^{D} \frac{\theta_{j0}^{x_i} \left(1-\theta_{j0}\right)^{1-x_i}}{\theta_{j1}^{x_i} \left(1-\theta_{j1}\right)^{1-x_i}}$$

$$\log f(\theta) =$$

$$\left(-\log\left(\frac{\pi}{1-\pi}\right) - \sum_{i=1}^{D} \log\left(\frac{1-\theta_{j1}}{1-\theta_{j0}}\right)\right) + \sum_{i=1}^{D} \left(x_i \left(\log\left(\left(\frac{\theta_{j0}}{\theta_{j1}}\right)\left(\frac{1-\theta_{j1}}{1-\theta_{j0}}\right)\right)\right)\right)$$

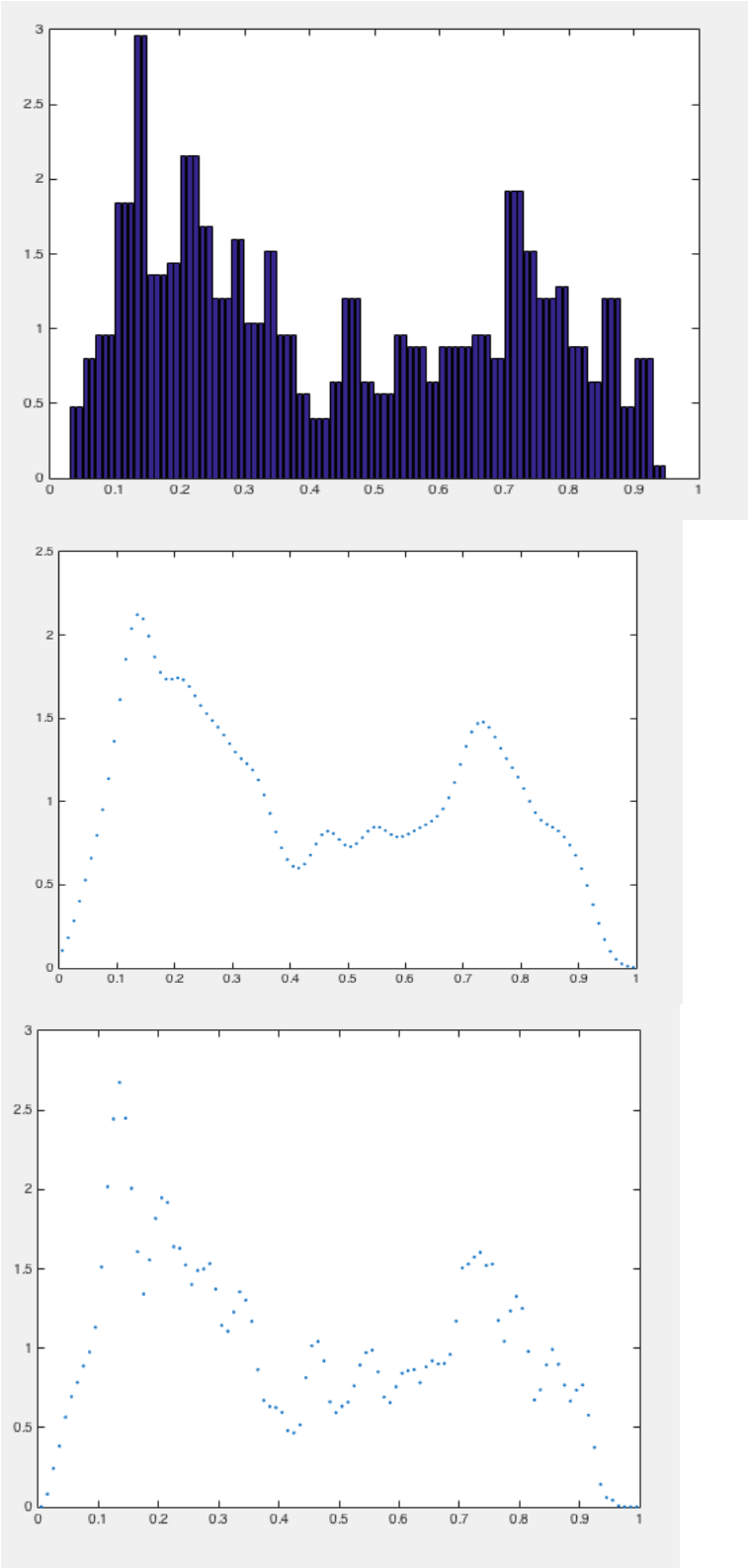
$$substituting\ w_0 = -\log\left(\frac{\pi}{1-\pi}\right) + \sum_i \log\left(\frac{1-\theta_{j1}}{1-\theta_{j0}}\right) \quad -(1)$$

$$and\ w = \left[\log\left(\frac{\theta_{11}}{\theta_{10}}\right)\left(\frac{1-\theta_{10}}{1-\theta_{11}}\right) \ldots \log\left(\frac{\theta_{D1}}{\theta_{D0}}\right)\left(\frac{(1-\theta_{D0})}{(1-\theta_{D1})}\right]\right. \quad -(2)$$

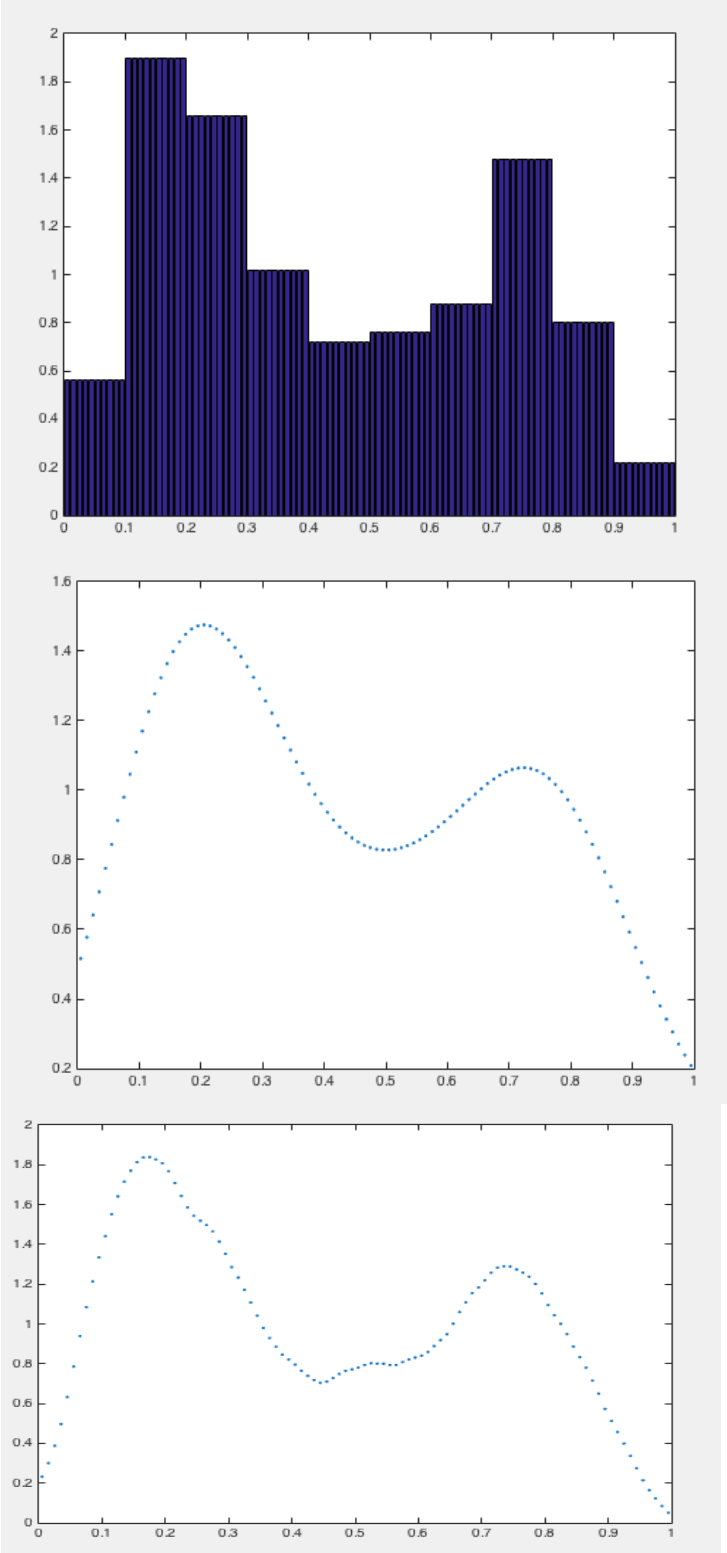
$$=> f(\theta) = \exp\ \hat{}(-w_0 + w^T X)$$

$$=> P(Y=1|X) = \frac{1}{1+\exp\ \hat{}(w_0 + w^T X)}\ \ where\ w_0\ given\ in\ (1)\ and\ w\ in\ (2).$$

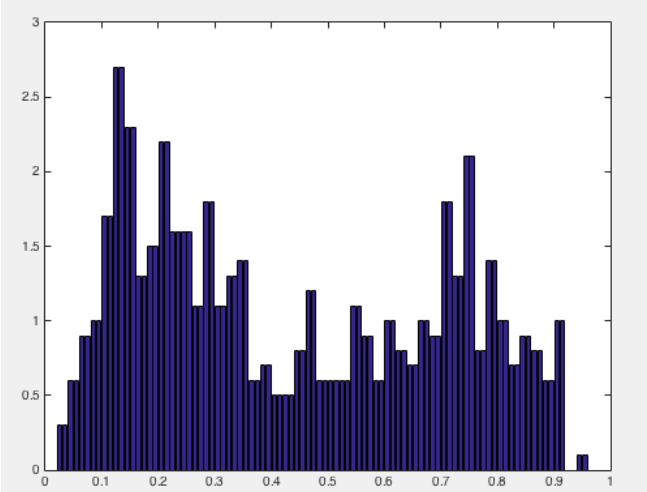
5.1 For each H value I have plotted 1.histogram, 2.Gaussian and 3.Epanechnikov kernel in the same order.

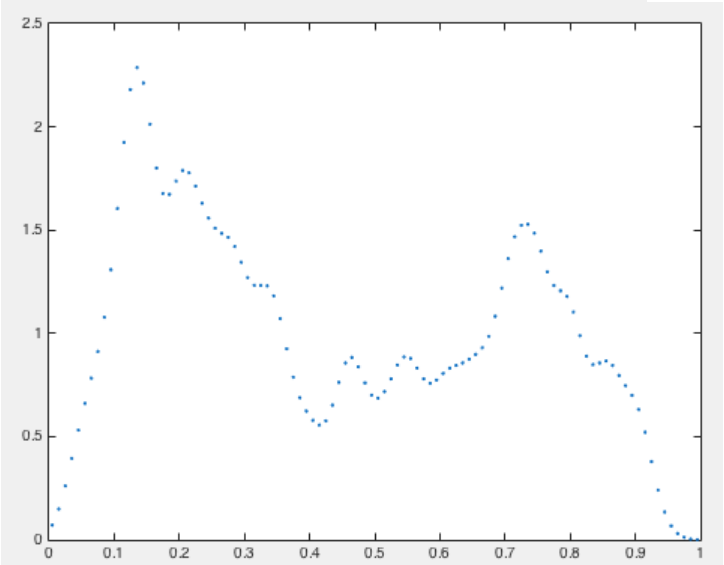


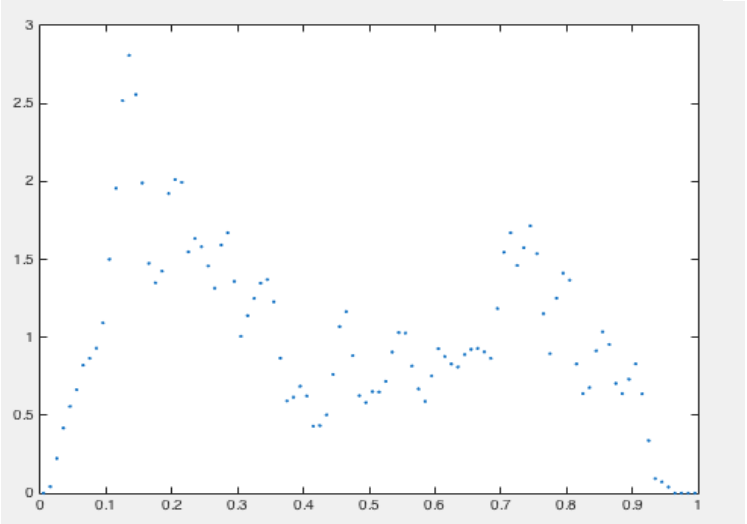
h=0.01



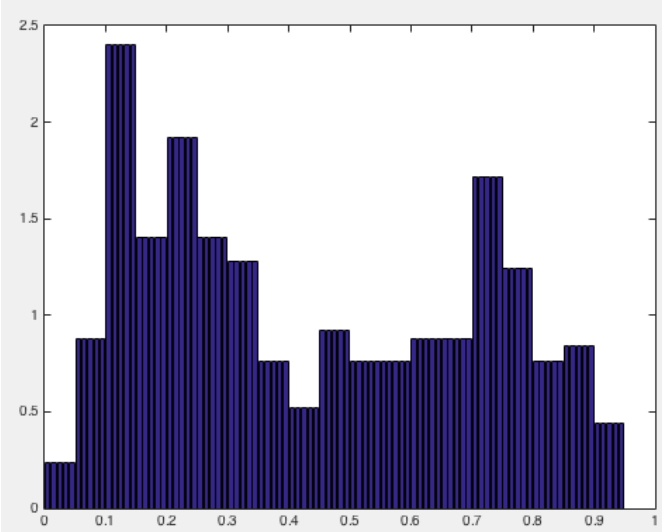
h=0.05

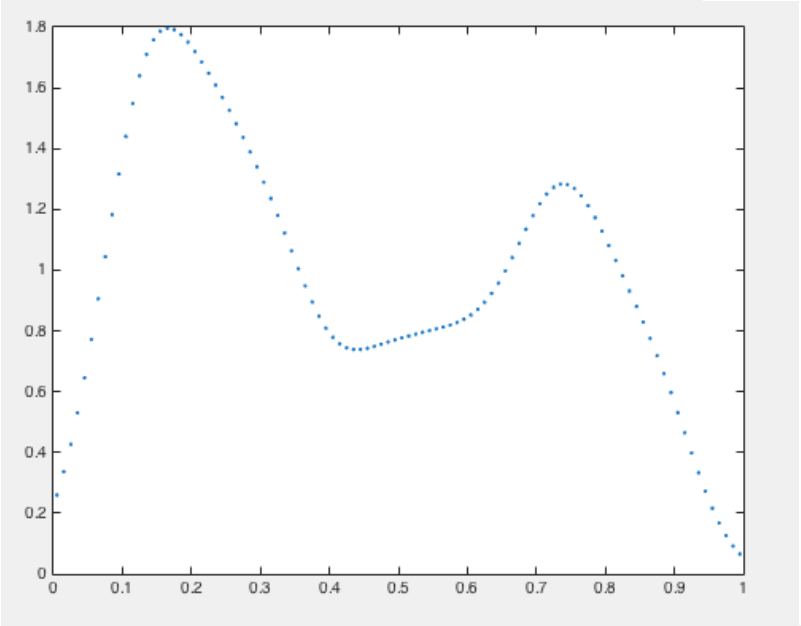

3
2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

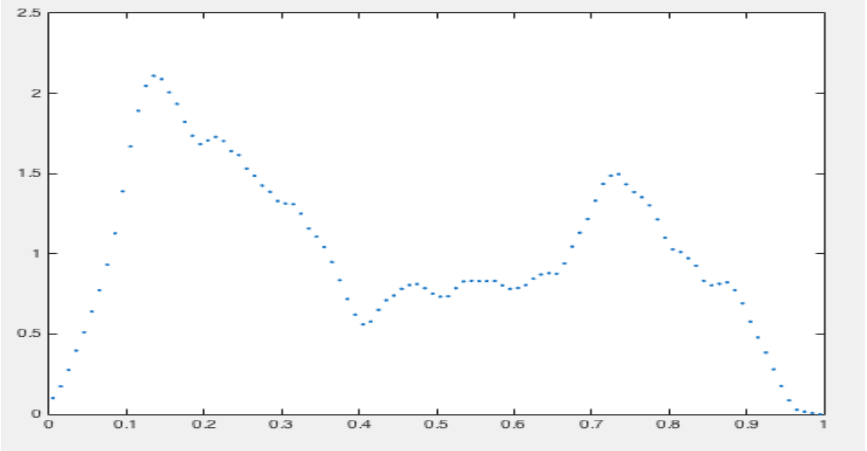

2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


3
2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

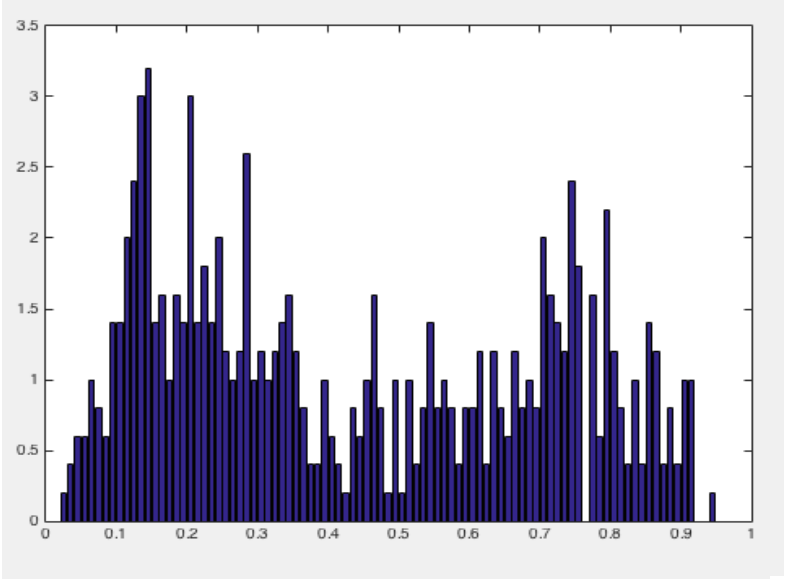
h=0.02

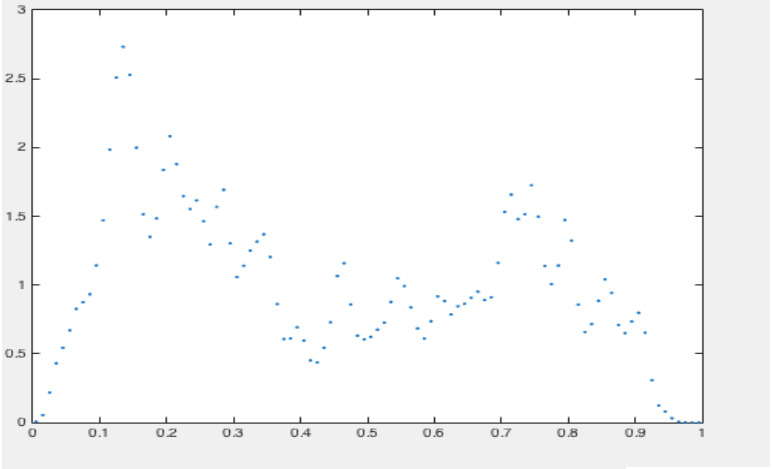

2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

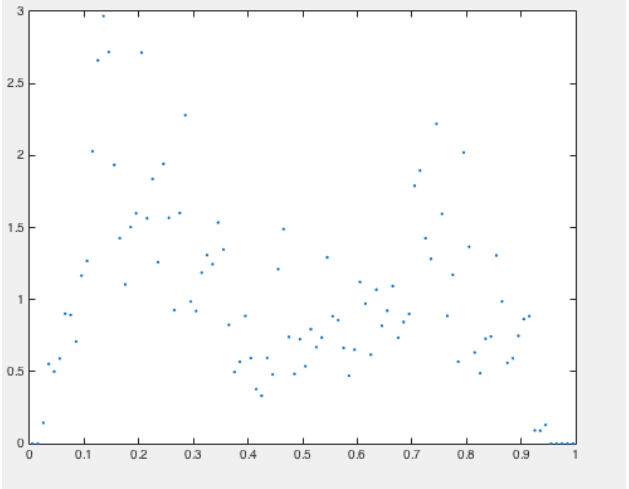

1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

h=0.1

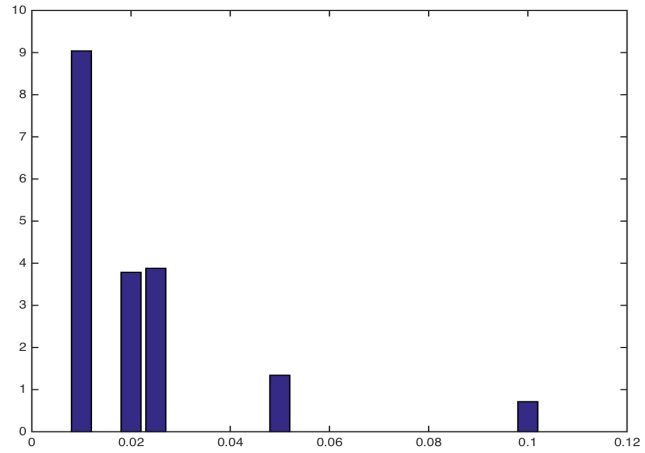

3.5
3
2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


3
2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

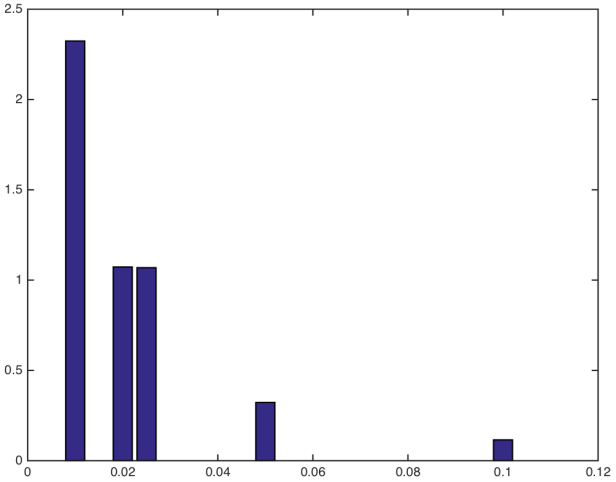

3
2.5
2
1.5
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

h=0.025

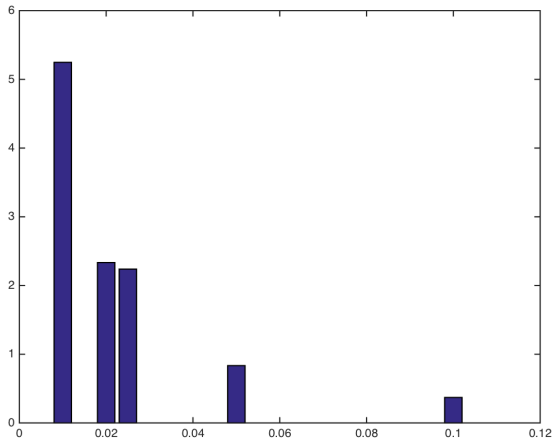
5.1 Integrated square error Graph for Histogram ,Gaussian and Epanechnikov kernel and among them we can see that Gaussian kernel give the least error.



Histogram



Gaussian



Epanechnikov

5.2d,
DECISION TREE
SplitCriterion gdi , MinLeaf 1new_accu=0.795349 train_accu=0.907021
SplitCriterion deviance , MinLeaf 1new_accu=0.776744 train_accu=0.897533

SplitCriterion gdi , MinLeaf 2new_accu=0.795349 train_accu=0.907021
SplitCriterion deviance , MinLeaf 2new_accu=0.776744 train_accu=0.897533

SplitCriterion gdi , MinLeaf 3new_accu=0.767442 train_accu=0.901328
SplitCriterion deviance , MinLeaf 3new_accu=0.786047 train_accu=0.899431

SplitCriterion gdi , MinLeaf 4new_accu=0.772093 train_accu=0.888046
SplitCriterion deviance , MinLeaf 4new_accu=0.776744 train_accu=0.893738

SplitCriterion gdi , MinLeaf 5new_accu=0.772093 train_accu=0.878558
SplitCriterion deviance , MinLeaf 5new_accu=0.762791 train_accu=0.878558

SplitCriterion gdi , MinLeaf 6new_accu=0.772093 train_accu=0.870968
SplitCriterion deviance , MinLeaf 6new_accu=0.758140 train_accu=0.869070

SplitCriterion gdi , MinLeaf 7new_accu=0.790698 train_accu=0.855787
SplitCriterion deviance , MinLeaf 7new_accu=0.767442 train_accu=0.853890

SplitCriterion gdi , MinLeaf 8new_accu=0.781395 train_accu=0.836812
SplitCriterion deviance , MinLeaf 8new_accu=0.767442 train_accu=0.838710

SplitCriterion gdi , MinLeaf 9new_accu=0.767442 train_accu=0.825427
SplitCriterion deviance , MinLeaf 9new_accu=0.758140 train_accu=0.829222

SplitCriterion gdi , MinLeaf 10new_accu=0.795349 train_accu=0.823529
SplitCriterion deviance , MinLeaf 10new_accu=0.786047 train_accu=0.827324

SplitCriterion gdi , MinLeaf 11new_accu=0.767442 train_accu=0.819734
SplitCriterion deviance , MinLeaf 11new_accu=0.767442 train_accu=0.819734

SplitCriterion gdi , MinLeaf 12new_accu=0.758140 train_accu=0.819734
SplitCriterion deviance , MinLeaf 12new_accu=0.758140 train_accu=0.819734

SplitCriterion gdi , MinLeaf 13new_accu=0.758140 train_accu=0.819734
SplitCriterion deviance , MinLeaf 13new_accu=0.758140 train_accu=0.819734

SplitCriterion gdi , MinLeaf 14new_accu=0.762791 train_accu=0.815939
SplitCriterion deviance , MinLeaf 14new_accu=0.762791 train_accu=0.815939

SplitCriterion gdi , MinLeaf 15new_accu=0.767442 train_accu=0.804554
SplitCriterion deviance , MinLeaf 15new_accu=0.767442 train_accu=0.804554

SplitCriterion gdi , MinLeaf 16new_accu=0.753488 train_accu=0.793169
SplitCriterion deviance , MinLeaf 16new_accu=0.753488 train_accu=0.793169

SplitCriterion gdi , MinLeaf 17new_accu=0.753488 train_accu=0.789374
SplitCriterion deviance , MinLeaf 17new_accu=0.753488 train_accu=0.789374

SplitCriterion gdi , MinLeaf 18new_accu=0.762791 train_accu=0.787476
SplitCriterion deviance , MinLeaf 18new_accu=0.762791 train_accu=0.787476

SplitCriterion gdi , MinLeaf 19new_accu=0.748837 train_accu=0.783681
SplitCriterion deviance , MinLeaf 19new_accu=0.748837 train_accu=0.783681

SplitCriterion gdi , MinLeaf 20new_accu=0.716279 train_accu=0.772296
SplitCriterion deviance , MinLeaf 20new_accu=0.716279 train_accu=0.772296

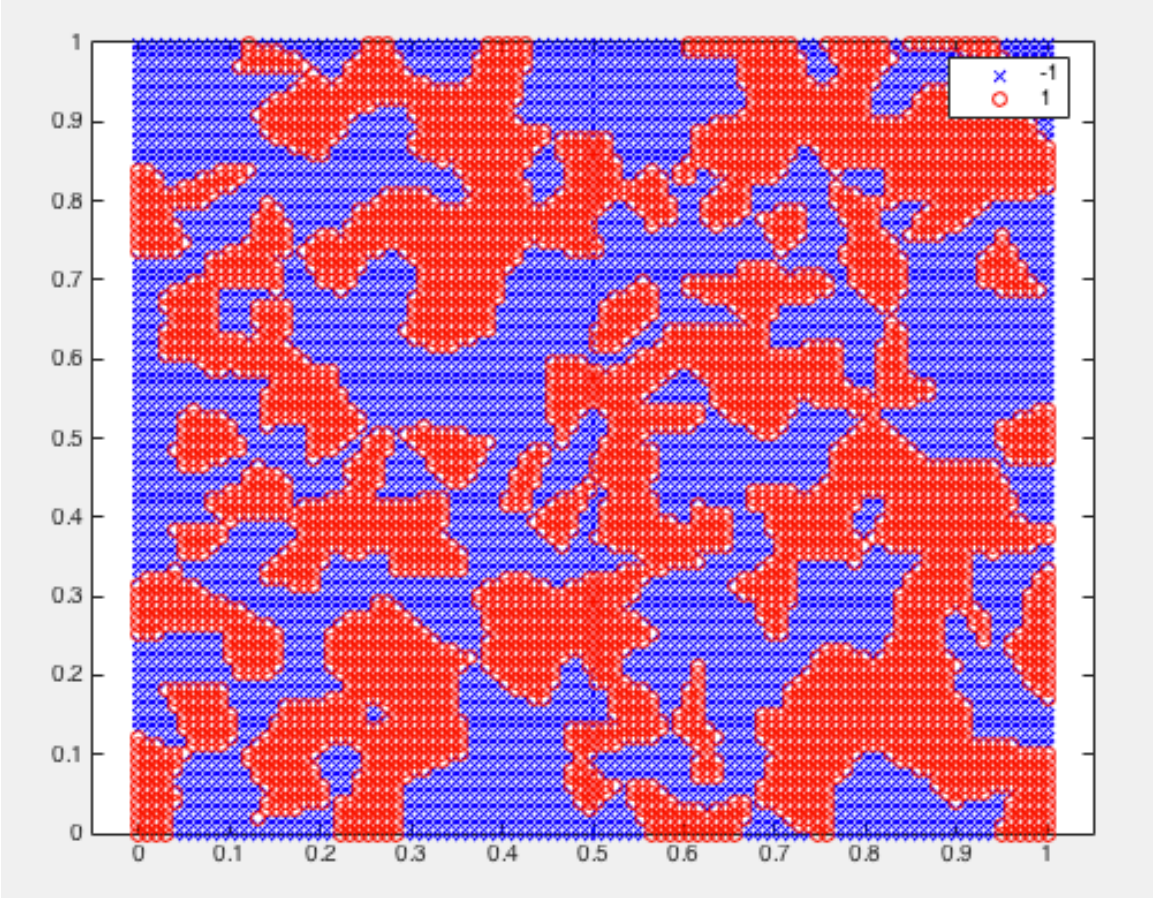
5.2.d KNN- From the observation below we can see that highest prediction is for k=9 and k=11

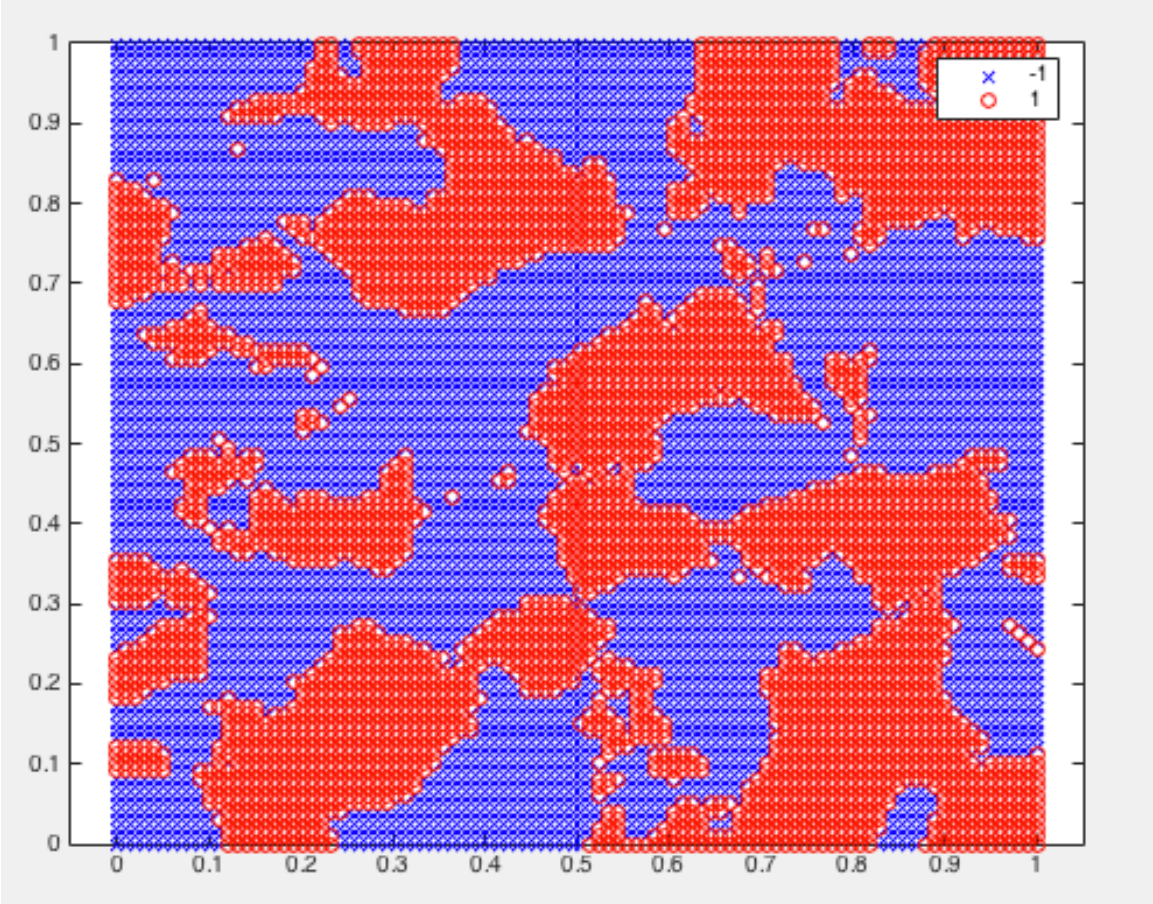
k=1	new_accuracy=0.651163	train_accuracy=0.686907
k=3	new_accuracy=0.688372	train_accuracy=0.724858
k=5	new_accuracy=0.734884	train_accuracy=0.804554
k=7	new_accuracy=0.832558	train_accuracy=0.869070
k=9	new_accuracy=0.888372	train_accuracy=0.908918
k=11	new_accuracy=0.888372	train_accuracy=0.908918
k=13	new_accuracy=0.832558	train_accuracy=0.870968
k=15	new_accuracy=0.781395	train_accuracy=0.808349

5.2.d
Naive Bayes for Nursery data New accuracy=0.896991 Train accuracy=0.906507
Naive Bayes for Tic-Tac-Toe data New accuracy=0.730233 Train accuracy=0.728653

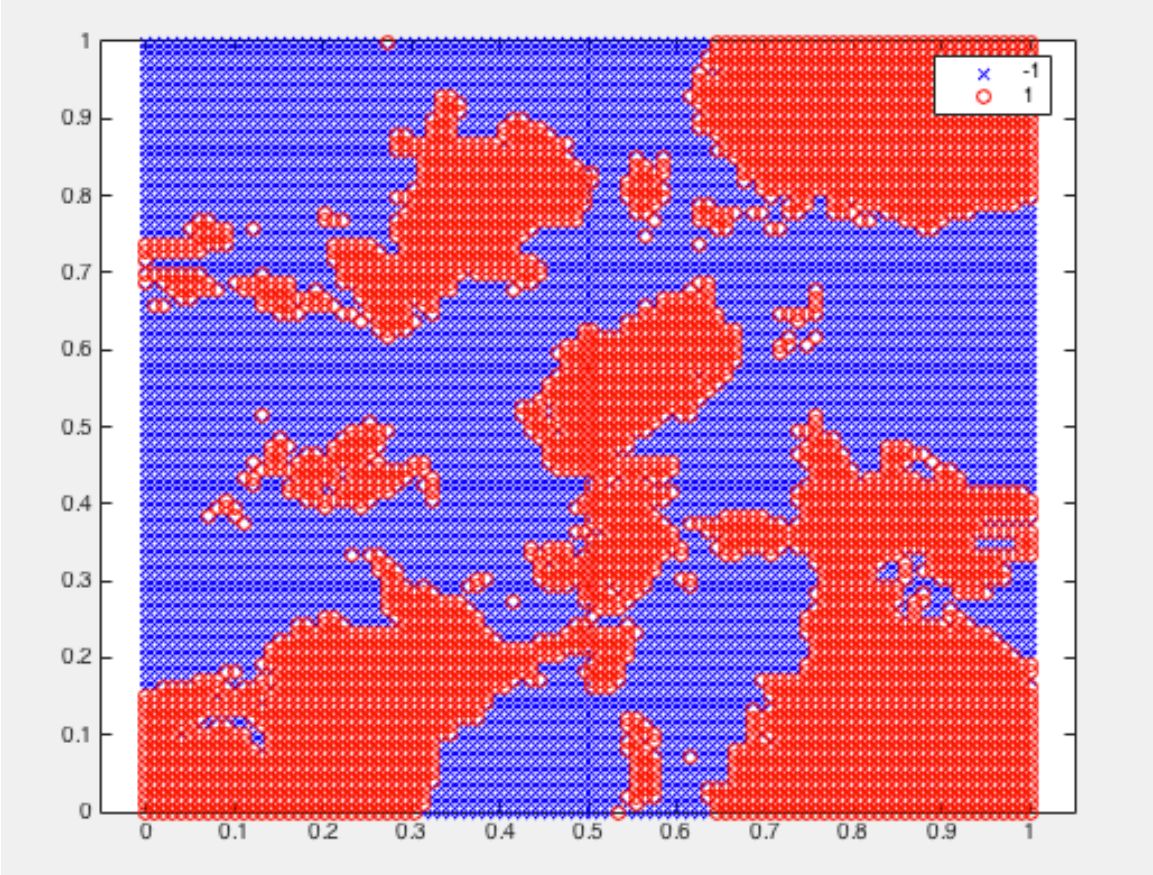
Conclusion---So from all the observations we can see that KNN with k= 9 and k=11 gives the best prediction. Also, decision tree with Gini-index performs better than Naïve bayes.

5.2(e)
k=1

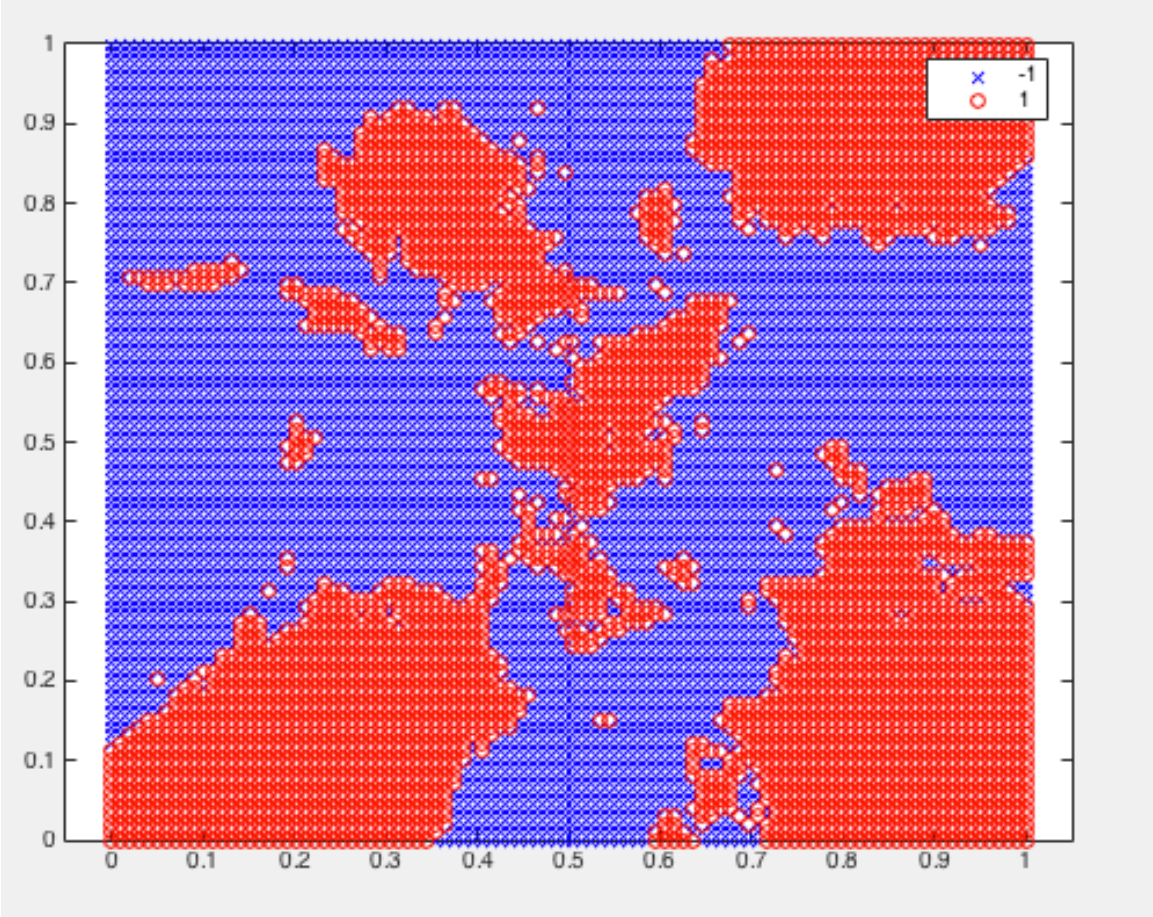




K=5



K=15



K=25

As the value of k increases smoothness increases.

I have discussed with Swaroop , Ismail, Amndeep and Vaibhav.