Name: Shreehar Joshi **ID: R00730962**

Q.no.1.

a. A flexible machine learning algorithm will perform **better** than an inflexible method in this case as the flexible algorithm can effectively capture the pattern in the data when the sample size is large and the number of predictors is less and hence fit to the training data better.

b. A flexible machine learning algorithm will perform **worse** than an inflexible method in this case as when there are very few observation and the number of predictors is extremely large, it will overfit to the training data as there will be very few data to learn the pattern from and will not accurately be able to make prediction as compared to an inflexible method. On the other hand, an inflexible method which has less number of model parameters will avoid overfitting and will make better predictions than the flexible machine learning algorithm.

c. A flexible machine learning algorithm will perform **better** than an inflexible method in this case as when the relationship between the predictors and response is highly-non linear then the non-flexible algorithm will be restricted from effectively capturing the pattern in the training data. On the other hand, the flexible machine learning algorithm can have more number of model parameters to be estimated from the data and hence less restriction to capture the pattern in the training data and hence perform better than the inflexible machine learning algorithm.

d. A flexible machine learning method will perform **worse** than the inflexible method in this case as when the variance of the error terms is extremely high, the flexible method, in attempt to capture every pattern in the data, would fit to the irrelevant information, referred to as noise, in the error terms and hence while predicting perform worse than the inflexible method. On the other hand, a less-flexible method will not be affected as much by the high variance as it has less number of model parameters to attempt to learn every pattern in the data.

Q.no.2.

a. Principal Component Analysis (PCA) is an unsupervised machine learning approach which is used for finding the low-dimensional representation of the dataset that contains as much as possible of the variation and hence information in the dataset. In other words, PCA is used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower dimensional data while preserving as much of the data's variation as possible. Each subsequent principal component found from PCA is orthogonal to the previous ones and they point in the directions of the largest variance of the residual subspace and in doing so ensure that as much variance as possible is captured from the high-dimensional dataset. PCA is used to address the curse of dimensionality (which refers to various problems that arise when dealing with high dimensional data) as it reduces the dimension of data with little effect to accuracy as it captures most of the necessary pattern in a small number of components.

   The two objectives of PCA is performing dimensionality reduction for:
   i. Performing visualization with low dimensional data as data in higher dimension are not easier to visualize and hence are not easily interpretable. The visualization of the low dimensional data can be used to obtain important insights such as new meaningful underlying variables and find patterns, such as clusters.
   ii. Speeding up training of the machine learning models in a dataset as PCA captures only the relevant information from the dataset.

b. **False** as the number of output dimensions might be equal to the number of initial dimensions in the dataset. For example, if there are relatively few features (say 3 predictor variables although we would never use PCA in such a less number of variables) in a dataset and none of these have relation, such as correlation, with each other, then the number of principal components needed to capture sufficient amount of variance (say, >80%) in the data might be same as the number of features or variables in the dataset before using PCA. Hence the new representation of the data is not always of lower dimensionality than the original feature representation

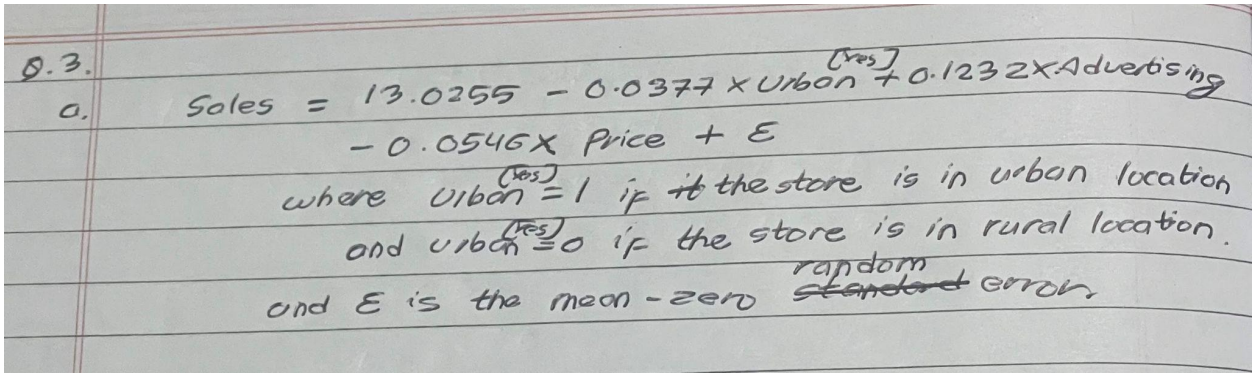
although in most of the usage this might be the case.

c. **False** because subsequent principal components are not just sometimes but always orthogonal to each other.

d. **True** as there are only a set of features and no outcome variables.

Q.no.3.

a.



b. Null Hypotheses and their analysis in terms of variables

Null hypothesis associated with the p-value of the f-statistic of the model: "The model doesn't have even a single predictor that has significant relationship with sales". This null hypothesis is rejected as the p value (2.85*10^-28) is very small and less than 0.05 and the alternative hypothesis is proposed, which is, the model has at least one variable that has a significant relationship with the sales and hence the model is valid.

Null hypothesis associated with the p value of the variable Urban: "The location of store (be it urban or rural) does not have significant relationship with the sales". Since the pvalue (0.886) is larger than 0.05, the null hypothesis cannot be rejected.

Null hypothesis associated with the p value of the variable Advertising: "The local advertising budget for the company does not have significant relationship with the sales". Since the pvalue (shown as 0.000) is less than 0.05, the null hypothesis is rejected and an alternative hypothesis is

proposed, which is, the local advertising budget for the company has a significant relationship with the sales.

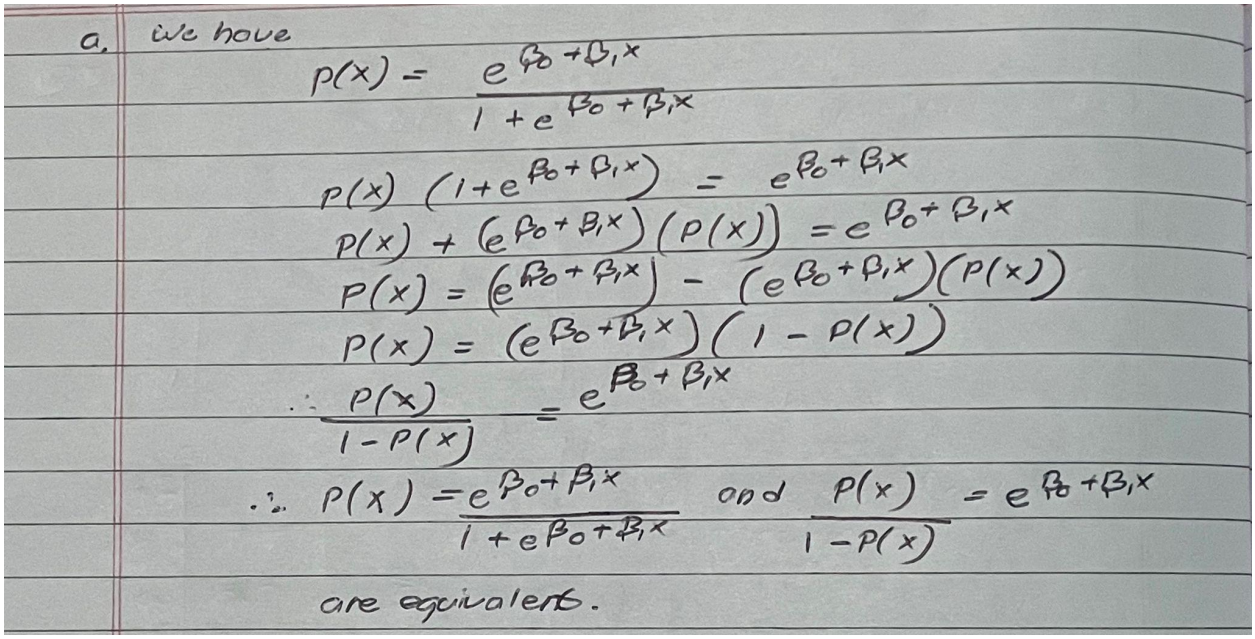
Null hypothesis associated with the p value of the variable Price: “The price company charges for car seats at each site does not have significant relationship with the sales”. Since the pvalue (shown as 0.000) is less than 0.05, the null hypothesis is rejected and an alternative hypothesis is proposed, which is, the price company charges for car seats at each site has a significant relationship with the sales.

Explanation of model in terms of variables

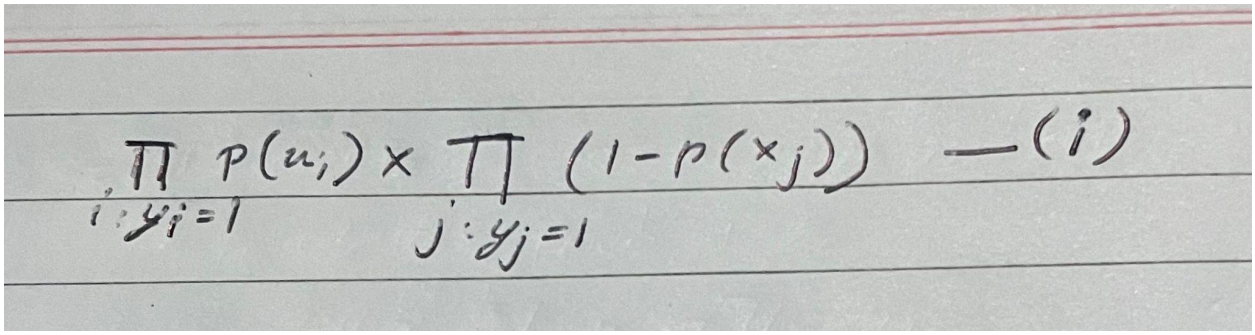
The Sales is 13025 units on average in a rural location when no advertising budget is allocated and no amount is charged for car seats. This is interpreted from the intercept which has a value of 13.0255. Urban location has on average 37.7 units less (because the coefficient is negative) sales than in rural location when all other predictors are fixed, which is interpreted from the value of the coefficient of Urban[Yes] (-0.0377). For every thousand dollars spent in advertising, the number of sales increases by 123.2 units on average when all other predictors are fixed, which is interpreted from the value of the coefficient of Advertising (-0.0546). Lastly, for every unit (exact unit is not given in description above) increase in charge for the car seats, the sales decreases by 54.6 units on average when all other predictors are fixed, which is interpreted from the coefficient of Price (-0.0546).

Q.no.4.

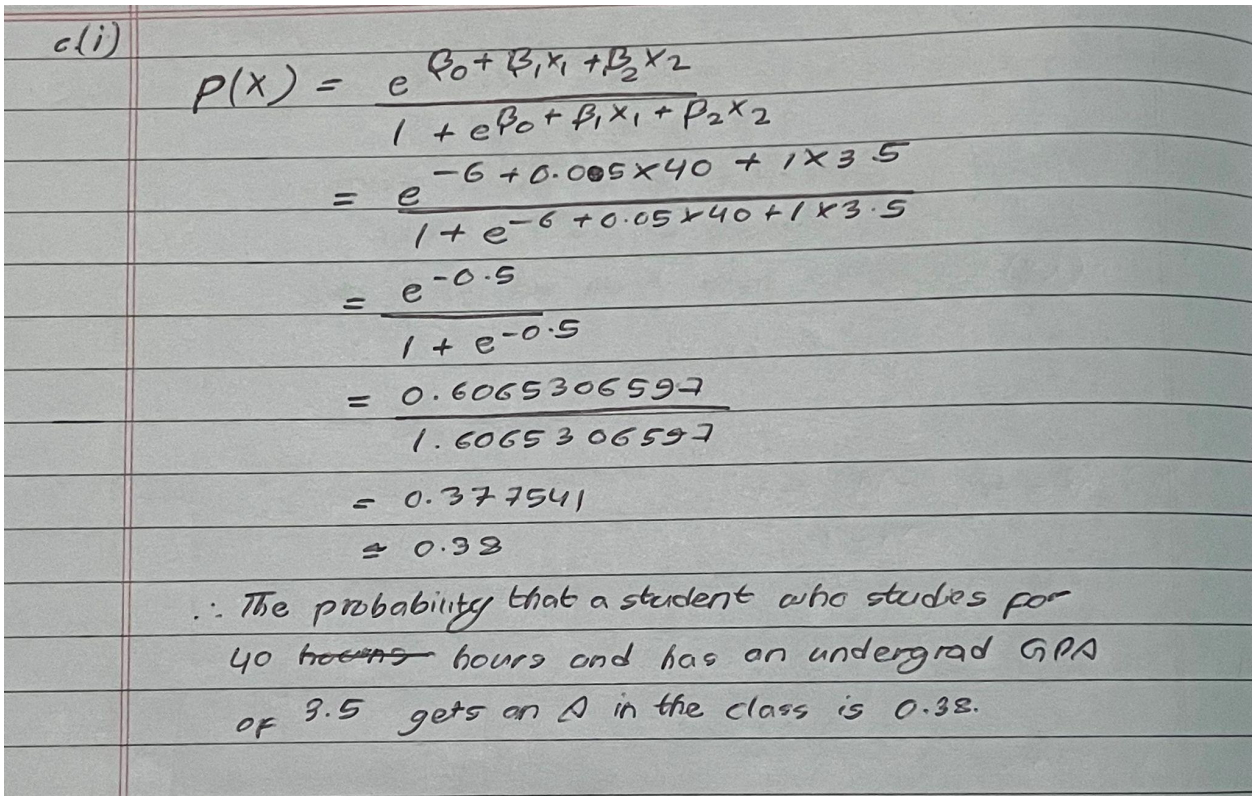
a.



b. The coefficients are estimated using the maximum likelihood method based on the available training data. The values of these coefficients are chosen such that the model that uses the logistic function (represented as equation 1 in Q.4) produces values with high probability, or close to 1, for observations in the class and values with low probability, or close to 0, for observations not in the class. In other words, we want both the classes to be predicted as falling under the correct classes and hence will choose the values of coefficients that will maximize the likelihood (represented mathematically by equation in (i) in image below) of this condition.



c. i.



c(i)

$$P(X) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

$$= \frac{e^{-6 + 0.05 \times 40 + 1 \times 3.5}}{1 + e^{-6 + 0.05 \times 40 + 1 \times 3.5}}$$

$$= \frac{e^{-0.5}}{1 + e^{-0.5}}$$

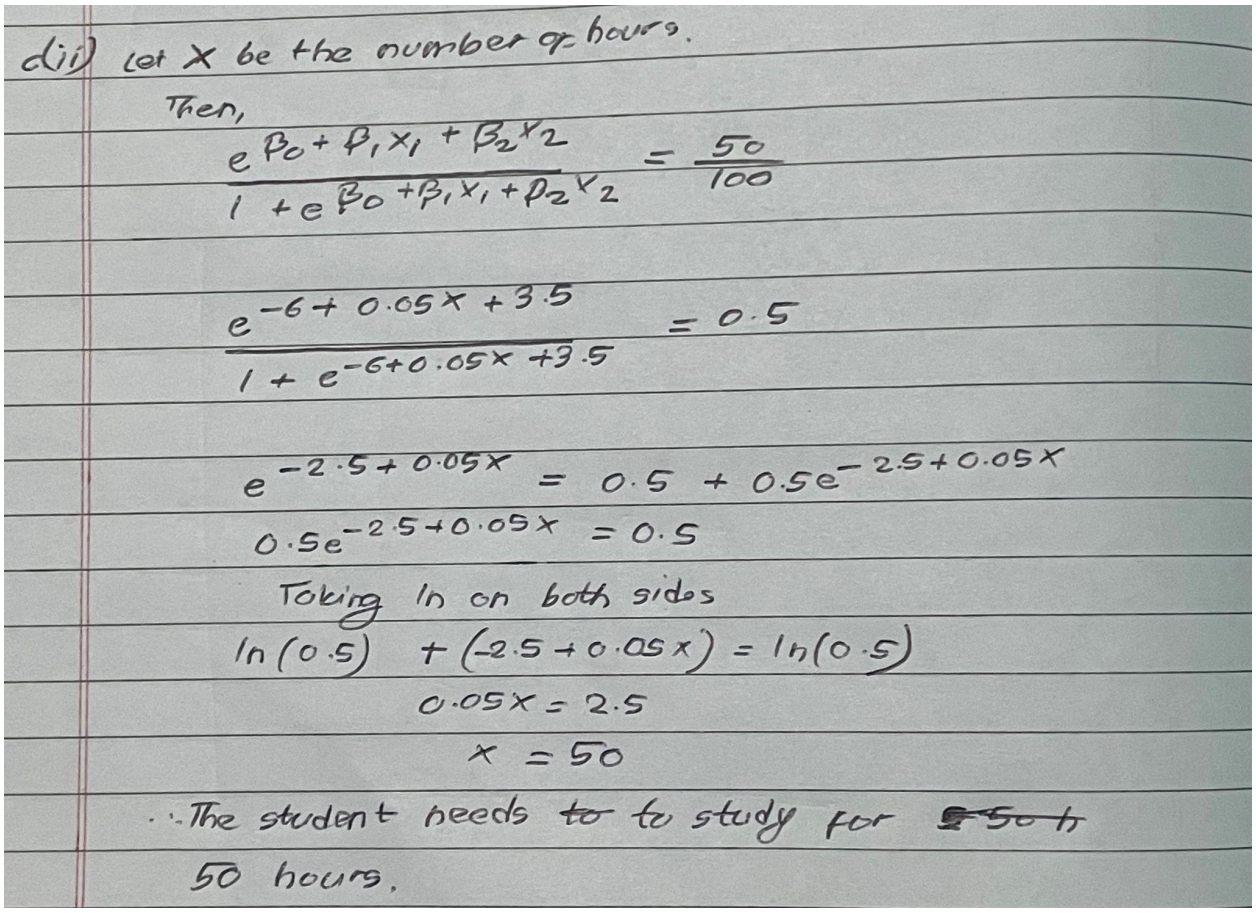
$$= \frac{0.6065306597}{1.6065306597}$$

$$= 0.377541$$

$$\approx 0.38$$

∴ The probability that a student who studies for 40 ~~hours~~ hours and has an undergrad GPA of 3.5 gets an A in the class is 0.38.

ii.



d(ii) Let X be the number of hours.

Then,

$$\frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}} = \frac{50}{100}$$

$$\frac{e^{-6 + 0.05X + 3.5}}{1 + e^{-6 + 0.05X + 3.5}} = 0.5$$

$$e^{-2.5 + 0.05X} = 0.5 + 0.5e^{-2.5 + 0.05X}$$

$$0.5e^{-2.5 + 0.05X} = 0.5$$

Taking ln on both sides

$$\ln(0.5) + (-2.5 + 0.05X) = \ln(0.5)$$

$$0.05X = 2.5$$

$$X = 50$$

∴ The student needs ~~to~~ to study for ~~50 h~~ 50 hours.

Q.no.5.

a.

DATE

Q.5. a. Complete linkage hierarchial procedure

Taking the maximum distance as the recalibrated distance in every single step

	1	2	3	4
1	0			
2	(0.3)	0		
3	0.4	0.5	0	
4	0.7	0.8	0.45	0

Link between 1 and 2 with distance 0.3.

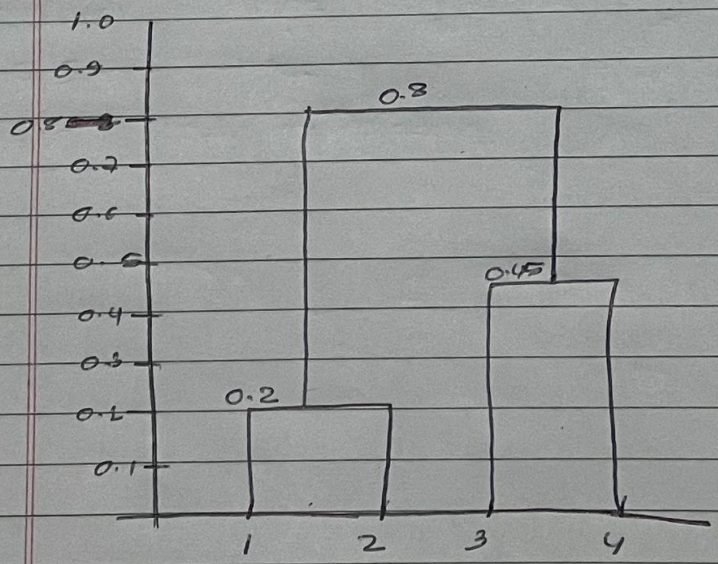
	1,2	3	4
1,2	0		
3	0.5	0	
4	0.8	(0.45)	0

Link between 3 and 4 with distance 0.45

	1,2	3,4
1,2	0	
3,4	(0.8)	0

Link between 1,2 and 3,4 with distance 0.8

Dendrogram (complete linkage)



b. We do not have enough information to conclude as depending on the dissimilarity value of these clusters the height the clusters fuse might vary. For instance, if the dissimilarity value between each of the components (1, 2, 3, 4, 5) are same (say, 5), then because complete linkage takes maximum distance and single linkage takes minimum distance as the recalibrated distance, all the clusters would fuse at the same height because both the minimum and maximum distance would be the same hence both the single and complete linkage will result in dendrogram with clusters {1,2,3} and {4,5} fusing at the same height. However, to give an example of another condition, there might be some other values for which these aforementioned clusters might fuse at lower height in single linkage (as this takes minimum distance as the recalibrated distance) than in the complete linkage. Hence with the information we are provided, we cannot conclude which fusion will occur higher on the tree or if they will fuse at the same height.

Q.no.6.

a. Variance is a term that indicates how much the estimate of the target function in the machine learning model will change if a different set of training data was used. In other words, variance is the variability in the prediction of a model (how much the Machine Learning model can adjust depending on the given dataset). If a model is complex or it is being trained on a dataset with a significant amount of noise, then the model will tend to overfit and subsequently have a higher variance. Contrary to a model with high bias, a model with high variance pays a lot of attention to training data and does not generalize on the data which it has not seen before. A model with high variance has low bias and a model with low variance has high bias.

   Bias is the amount that a prediction made by a machine learning model differs from target value when compared to the data on which it was trained. It is a systematic error that occurs in a machine learning model due to incorrect assumptions made by model. If a model has high bias, then it fails in capturing proper data trends and has high potential of underfitting because it is overly generalized (as it oversimplifies the

pattern) and subsequently has a high error rate. A model with high bias has low variance and a model with low bias has high variance.

Bias and variance are inversely connected and hence there needs to be a balance between these two when deploying a model (also known as bias-variance trade off). In other words we must not ignore the increase in one aspect when we are making changes to decrease another.

b. Restrictive machine learning techniques have an advantage over very flexible approaches in that they are easy to interpret and this might be a reason one would choose a more restrictive machine learning technique than a very flexible approach. For instance, linear regression or even better lasso regression (which has fewer variables in its final equation) can be used to see how each of the variables have effect in the outcome variables by looking into the equation that can be interpreted from the outcome of these models. However, it is really difficult to understand the effects of each of the variables in the outcome variable in flexible models like boosting and neural networks. These restrictive models might have higher accuracy or perform better but it is hard to make inference on how they actually work (what affects each of the predictor variables have in different steps). Hence, one would choose a restrictive model over flexible approach for better interpretability.

Q.no.7.

a. The models, all of which have achieved a precision score of 95%, can be combined into an ensemble for voting as voting methods, which combine more than one machine learning model, can obtain better results than individual classifiers alone. However, a voting ensemble might not lead to an increase in the precision if the models used are not independent because if they are similar (for instance decision trees and random forest), then the errors will not be independent leading to even poorer results than the individual models. Hence there is a high likelihood of increase in precision if the five models are different from one another. In addition to the method described above, one can also try training the ensemble of models on different sets by taking samples either with replacement (bagging) or

without replacement (pasting) to see if the precision scores will increase.

b. It is possible to speed up training of a random forest by distributing it across multiple servers as each of the predictions made by the model does not rely on the other predictors. In other words, the predictors are independent. Hence, their load can be distributed across multiple servers to speed up the training. However, it is not possible to speed up training of boosting ensemble by distributing across multiple servers as each of the predictor relies on the results of earlier predictor as boosting works by improving the prediction by evaluating the results of the earlier predictor (also known as sequential learning technique); in other words, distributing the training across servers will not decrease the training time as predictor on any given server except the first will still have to wait for earlier predictor to continue make predictions.

c. The main difference between hard and soft voting classifiers lies in the way in which they make their final predictions. Hard voting counts the number of votes given by each of the classifier and predicts the class that gets the highest number of votes while soft voting calculates the average probability by taking into account the probabilities predicted by each of the classifiers in the ensemble and makes prediction as the class that has the highest average class probability. Secondly, soft voting is generally more efficient than hard voting as it gives more importance to higher probabilities than just counting the number of votes for each class as hard voting does. Lastly, hard voting doesn't require its classifiers to predict the probabilities while in soft voting classifiers need to be able to predict the probability of class.