Name: Ayush Jain
SAP ID: 60004200132
Branch: CS

Module 1:

1. What are the important objectives of machine learning?

"Learning is any process by which a system improves from experience" - Herbert Simon

Definition by Tom Mitchell (1998):

Machine Learning is the study of algorithms that

- improve their performance *P*
- at some task *T*
- with experience *E*.

A well-defined learning task is given by <*P*, *T*, *E*>.

2. What do you mean by a well –posed learning problem? Explain the important features that are required to well –define a learning problem.

A Machine learning system learns from historical data, builds the prediction models, ad whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data. As the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it we just need to feed the data to generic algorithms, and with the help of the algorithms machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem.

Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

3. **What are the basic design issues and approaches to machine learning?**

Common issues in Machine Learning

1. Inadequate Training Data
2. Poor quality of data
3. Non-representative training data
4. Overfitting and Underfitting
5. Monitoring and maintenance
6. Getting bad recommendations
7. Lack of skilled resources
8. Customer Segmentation
9. Process Complexity of Machine Learning
10. Data Bias
11. Lack of Explain ability
12. Slow implementations and results
13. Irrelevant features

4. Design a learning system with all stages involved in the process.
   - Choose the training experience
   - Choose exactly what is to be learned

– i.e. the target function
   - Choose how to represent the target function
   - Choose a learning algorithm to infer the target function from the experience.

5. Differentiate categorical and continuous valued feature with the help of following examples?

a. Ethnicity of a person- Categorical
b. Area (in sq. centimeter) of your laptop screen.- continuous
c.The color of the curtains in your room.- categorical

6.

Comparison Table

Criteria	Supervised ML	Unsupervised ML	Reinforcement ML
Definition	Learns by using labelled data	Trained using unlabelled data without any guidance.	Works on interacting with the environment
Type of data	Labelled data	Unlabelled data	No – predefined data
Type of problems	Regression and classification	Association and Clustering	Exploitation or Exploration
Supervision	Extra supervision	No supervision	No supervision
Algorithms	Linear Regression, Logistic Regression, SVM, KNN etc.	K – Means, C – Means, Apriori	Q – Learning, SARSA
Aim	Calculate outcomes	Discover underlying patterns	Learn a series of action
Application	Risk Evaluation, Forecast Sales	Recommendation System, Anomaly Detection	Self Driving Cars, Gaming, Healthcare

7.

(b) Grouping related documents from an unannotated corpus is an unsupervised task.

In supervised learning, the algorithm is trained on labelled data to predict the label of new data. However, in unsupervised learning, the algorithm is not given any labelled data, and it has to find patterns or structures in the data on its own.

Option (a) involves predicting a label (sweet or spicy) for new edible items based on the information of ingredients, their quantities, and labels for many other similar dishes. This is a supervised learning task as the algorithm is trained on labelled data (ingredients, quantities, and labels) to predict the label for new data.

Option (b) involves grouping related documents from an unannotated corpus. The algorithm is not given any labels or categories to group the documents. Instead, it has to find patterns or structures in the documents and group them based on their similarity. This is an unsupervised learning task.

Option (c) involves grouping hand-written digits based on their image. This is also an unsupervised learning task as the algorithm has to find patterns or structures in the images and group them based on their similarity, without any labelled information.

Option (d) involves predicting the degree of a scholar based on historical data such as qualifications, department, institute, research area, and time taken by other scholars to earn the degree. This is not a machine learning task and can be done using simple statistical methods, as it involves analyzing historical data rather than finding patterns or structures in unannotated data.

8.

Explain issues occuring in ML.

(i) Inadequate training data - Data quality can be affected by factors such as noisy data, incorrect data and generalizing of output data.

(ii) Poor quality of data - It is inaccurate data, such as missing/outdated information.

(iii) Non-representative data - Data points not represented/categorised by features.

(iv) Overfitting and underfitting - Whenever there is a huge amount of training data, the prediction is negatively affected due to the massive amount of biased data. This is called overfitting. When a ML model is provided with fewer amounts of data, it provides incomplete and inaccurate data, called underfitting.

(v) Monitoring and maintenance - Generalized output data is mandatory for any ML model, hence regular monitoring and maintenance become compulsory for the same.

(vi) Getting bad recommendations - It occurs when new data is introduced or interpretation of data changes.

(vii) Lack of skilled resources - Absence of skilled resources in the form of manpower is an issue.

(viii) Customer segmentation - An algorithm is necessary to recognize customer behaviour and trigger a relevant recommendation for user-based past experience.

(ix) Process complexity of ML - The ML process is very complex, which is another issue.

(x) Data bias - These errors exist when certain elements of dataset are heavily weighted or need more importance than others.

(xi) Lack of explainability - Outputs cannot be easily understood.

(xii) Slow implementations and results - Very common issue

(xiii) Irrelevant features - If we feed garbage data as input, we will get garbage result.

9.

Inductive bias is a concept in machine learning that refers to the set of assumptions or beliefs that a learning algorithm uses to make predictions or decisions. It is the set of assumptions that the algorithm makes about the

relationship between the input data and the output labels, and it guides the algorithm's search for the best hypothesis or model that fits the data.

Inductive bias helps the algorithm to generalize from the training data to the test data by making certain assumptions about the data that are not explicitly stated in the training data. The inductive bias is based on prior knowledge, assumptions, or heuristics that are built into the learning algorithm or provided by the domain expert.

For example, a decision tree algorithm has an inductive bias that prefers smaller, simpler trees that are easier to interpret and less prone to overfitting. This bias is based on the assumption that simpler models are more likely to generalize well to new data. Similarly, a neural network algorithm may have an inductive bias that prefers smooth functions that are easier to optimize.

Inductive bias is important because it helps the learning algorithm to make predictions or decisions in situations where there is uncertainty or ambiguity in the data. However, the choice of inductive bias can also have a significant impact on the performance of the learning algorithm. A biased algorithm may overlook important features or patterns in the data, while an unbiased algorithm may over fit the training data and perform poorly on the test data.

In summary, the concept of inductive bias refers to the assumptions or beliefs that a learning algorithm uses to make predictions or decisions. It is an important concept in machine learning because it helps the algorithm to generalize from the training data to the test data, but it can also have a significant impact on the performance of the algorithm.

12.

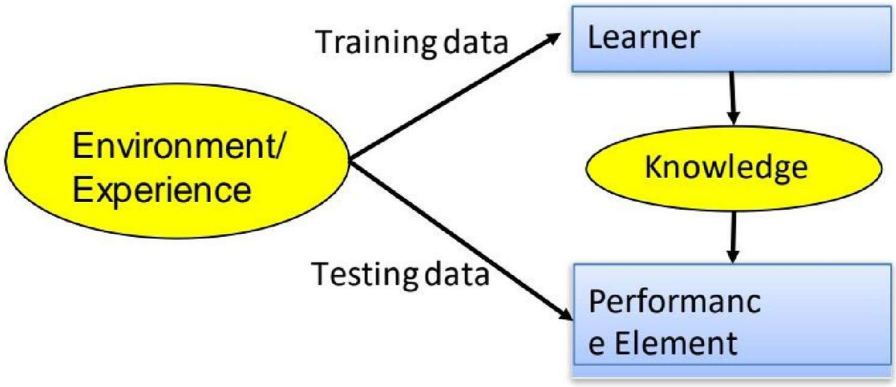
1) Designing a learning system

• choose the training experience

• choose exactly what is to be learner– i.e. the target function

• choose how to represent the target function

• choose a learning algorithm to infer the target function from the experience. learner



2) Training vs. Test Distribution

• We generally assume that the training and

• test examples are independently drawn from

• the same overall distribution of data – We call this “i.i.d” which stands for “independent and identically distributed”

• If examples are not independent, requires collective classification

• If test distribution is different, requires transfer learning

3) ML in a Nutshell

• Tens of thousands of machine learning

• algorithms

– Hundreds new every year

• Every ML algorithm has three components:

– Representation

– Optimization

– Evaluation

13.

Overfitting: A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise.

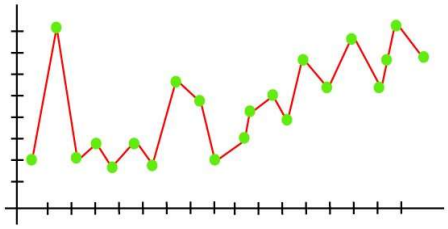
Reasons for Overfitting:

1. High Variance, Low bias

2. Model is too complex 3. Size of Training Data

Example: We have a training data set such as 1000 mangoes, 1000 apples, 1000 bananas, and 5000 papayas. Then there is a considerable probability if identification of an apple as papaya because we have a massive amount of biased data in training data set; hence prediction got negatively affected.

OR

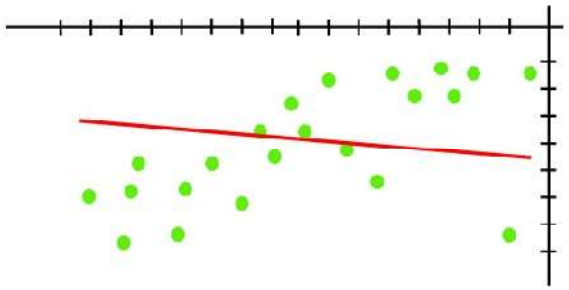


Underfitting:

A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data. Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data. In such cases, the rules of the machine learning model are too easy and flexible to be applied to such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.
Reasons for Underfitting:

1. High Bias and low variance
2. The size of training data set is not enough

Example: The model is unable to capture the data points present in the plot.

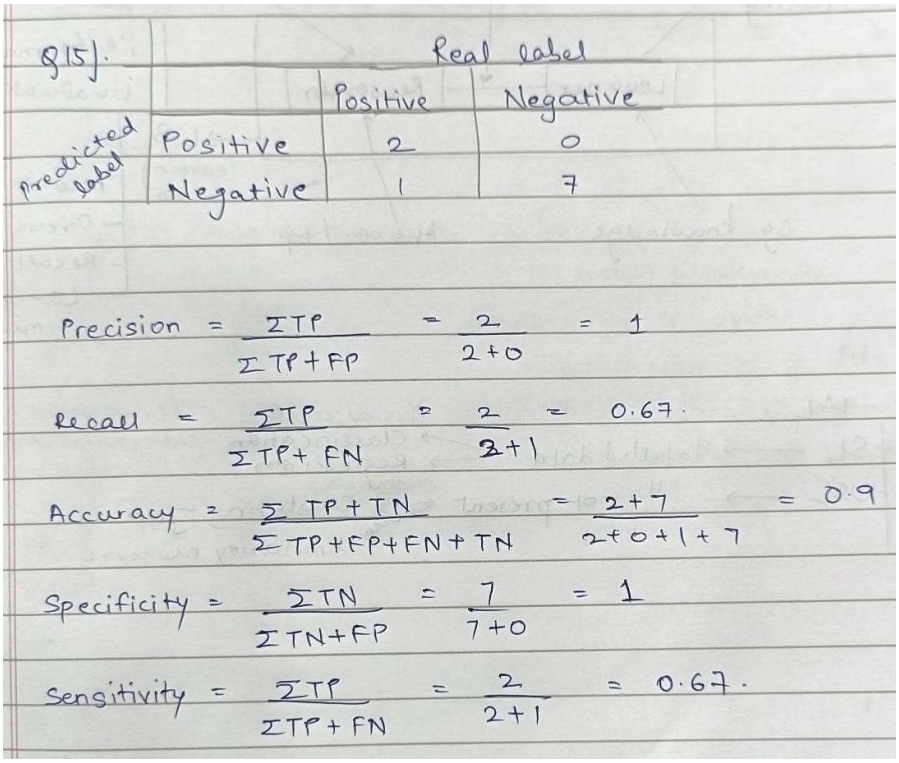


14.

a) Precision (tp / (tp + fp)) measures the ability of a classifier to identify only the correct instances for each class

b) Recall (tp / (tp + fn) is the ability of a classifier to find all correct instances per class.

15.



16.

Om Chhabra 60001200078

Machine Learning Question Bank:

• Module 1:

(Q16)

We have,

X_actual = [1 1 0 1 0 0 1 0 0 1 0]

Y_pred = [1 0 1 1 1 0 1 1 0 0]

→ Confusion Matrix:

	Predicted 0	Predicted 1	
Actual 0	$\Sigma TN = 3$	$\Sigma FP = 3$	= 6
Actual 1	$\Sigma FN = 1$	$\Sigma TP = 3$	= 4
	= 4	= 6	

→ Performance Parameters:

(i) Precision = $\frac{\Sigma TP}{\Sigma TP + \Sigma FP} = \frac{3}{3+3} = \frac{3}{6} = 0.5$

(ii) Recall (Sensitivity) = $\frac{\Sigma TP}{\Sigma TP + \Sigma FN} = \frac{3}{3+1} = \frac{3}{4} = 0.75$

(iii) Accuracy = $\frac{\Sigma TP + TN}{\Sigma TP + FP + TN + FN} = \frac{3+3}{3+3+3+1} = \frac{6}{10} = 0.6$

(iv) Specificity = $\frac{\Sigma TN}{\Sigma TN + FP} = \frac{3}{3+3} = \frac{3}{6} = 0.5$

17.

(Q17) We have

$x_{actual} = [5, -1, 2, 10]$

$y_{pred} = [3.5, -0.9, 2, 9.9]$

Mean Absolute Error = (MAE)

$= \frac{1}{N} \sum_{i=1}^{N} |y_i - \bar{y}|$

$N = 4, \quad \bar{y} = \frac{5 - 1 + 2 + 10}{4} = 4$

$$MAE = (|5-4|+|-1-4|+|2-4|+|10-4|) \times \frac{1}{4}$$

$$= (1+5+2+6) \times \frac{1}{4}$$

$$= \frac{14}{4} = 3.5$$

$$\text{Mean Square Error} = \frac{1}{N}\sum(y_i-\bar{y})^2$$

$$MSE = \frac{1}{4} \times [(1)^2+(5)^2+2^2+6^2)$$

$$= \frac{66}{4} = 16.5$$

$$\text{Variance} = R^2 = 1 - \frac{\sum(y_i-\hat{y})^2}{\sum(y_i-\bar{y})^2}$$

$$= 1 - \frac{\sum(y_i-\hat{y})^2}{66}$$

$$\text{Now } \sum(y_i-\hat{y})^2 = [(5-3.5)^2+(-1+0.9)^2+(2-2)^2 + (10-9.9)^2]$$

$$= 1.5^2+(-0.1)^2+0+(0.1)^2$$

$$= 2.27$$

$$R^2 = 1 - \frac{2.27}{66} = 0.9656$$

18.

Q18. In a good review 10 among 100 times it is bad and in a bad review, 70 among 100 times it is bad.

Hence, Probability $= \dfrac{10(1269-1272)+70(1272)}{\text{Total no. of times}}$

$$= \frac{10(1124)+70(172)}{1296 \times 100}$$

$$= \frac{11240+12040}{129600}$$

$$= \frac{23280}{129600}$$

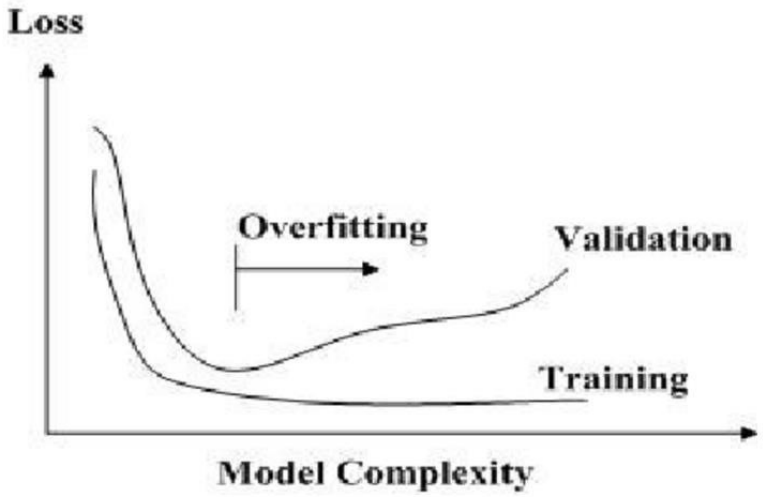
$$= 0.17962.$$

19.

More complex models will, in general, perform better on training data than simpler ones. Complex models have more parameters that can be adjusted during training to get a good fit to the desired result. Therefore, their error-rate when measured on the data-set used for training will usually be lower.

However, a model that is too complex can end up over-fitting to random effects that are present only in the specific data-set used for training. If these random effects are not present in new data when the model is asked to make predictions, then the model will produce incorrect results when it tries to use them. To judge whether this is happening, measure the model's error-rate on a validation data-set that was not used during training.

A complex model exhibiting high variance may improve in performance if trained on more data samples.



20.

Bias: Assumptions made by a model to make a function easier to learn. It is actually the error rate of the training data. When the error rate has a high value, we call it High Bias and when the error rate has a low value, we call it low Bias.

Variance: The difference between the error rate of training data and testing data is called variance. If the difference is high then it’s called high variance and when the difference of errors is low then it’s called low variance. Usually, we want to make a low variance for generalized our model.

Underfitting: A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data. Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data. In such cases, the rules of the machine learning model are too easy and flexible to be applied to such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.

Overfitting: A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the nonparametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

Module 2:

Q1) What are the four assumptions of linear regression?

Linear regression is a popular method for modelling the relationship between a dependent variable and one or more independent variables. There are four key assumptions that underlie linear regression:

1. Linearity: The relationship between the independent and dependent variables is linear, which means that the change in the dependent variable is proportional to the change in the independent variable(s). This assumption can be checked by plotting the data and looking for a linear relationship between the variables.
2. Independence: The observations in the dataset are independent of each other, which means that the value of the dependent variable for one observation is not affected by the value of the dependent variable for another observation. This assumption can be violated when the data is collected through time series or other types of correlated observations.
3. Homoscedasticity: The variance of the residuals (the difference between the predicted values and the actual values) is constant across all levels of the independent variable. This means that the errors are equally distributed across the range of the independent variable(s). This assumption can be checked by plotting the residuals against the predicted values and looking for a constant spread of the points.
4. Normality: The residuals follow a normal distribution, which means that the errors are randomly distributed around zero and have a constant variance. This assumption can be checked by plotting a histogram of the residuals and looking for a normal distribution.

These assumptions are important to check when using linear regression to model data, as violations of these assumptions can lead to biased or inaccurate results. If any of the assumptions are violated, it may be necessary to use a different modelling technique or to transform the data to meet the assumptions of linear regression.

Q2) What is meant by dependent and independent variables explain with an example.

In statistical analysis and machine learning, dependent and independent variables refer to the variables that are being measured or observed and the variables that are being used to predict or explain the observed outcomes.

The dependent variable is the variable that is being measured or observed and is expected to change in response to changes in the independent variable. It is sometimes referred to as the response variable or the outcome variable. In other words, it is the variable that we are interested in predicting or explaining. For example, in a study investigating the effect of a new drug on blood pressure, the dependent variable would be the blood pressure of the participants.

The independent variable is the variable that is being used to predict or explain changes in the dependent variable. It is sometimes referred to as the predictor variable or the explanatory variable. In other words, it is the variable that we believe is causing or influencing changes in the dependent variable. For example, in the same study investigating the effect of a new drug on blood pressure, the independent variable would be the dosage of the drug that the participants receive.

To summarize, the dependent variable is the variable that we want to understand or predict, and the independent variable is the variable that we believe is influencing or causing changes in the dependent variable.

In statistical analysis and machine learning, we use different methods to explore the relationship between the dependent and independent variables, such as regression analysis, correlation analysis,

and hypothesis testing. These methods help us to understand how changes in the independent variable(s) affect the dependent variable and to make predictions about the outcome variable based on the values of the predictor variables.

Q3) What is difference between simple linear and multiple linear regressions?

Simple linear regression and multiple linear regression are two techniques used in statistical analysis and machine learning to model the relationship between a dependent variable and one or more independent variables. The main difference between the two is the number of independent variables involved in the analysis.

In simple linear regression, there is only one independent variable that is used to predict the dependent variable. The relationship between the two variables is assumed to be linear, which means that the change in the dependent variable is proportional to the change in the independent variable. The equation for simple linear regression is $y = mx + b$, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept. For example, a simple linear regression could be used to model the relationship between a person's weight and their height.

In multiple linear regression, there are two or more independent variables that are used to predict the dependent variable. The relationship between the variables is also assumed to be linear, but the equation is more complex than the simple linear regression equation. The equation for multiple linear regression is $y = b0 + b1x1 + b2x2 + ... + bnxn$, where y is the dependent variable, x1, x2, xn are the independent variables, and b0, b1, b2, ..., bn are the regression coefficients. For example, a multiple linear regression could be used to model the relationship between a person's income, education level, and years of experience on their job performance.

In summary, the main difference between simple linear regression and multiple linear regression is the number of independent variables used to predict the dependent variable. Simple linear regression uses only one independent variable, while multiple linear regression uses two or more independent variables. Multiple linear regression is more complex than simple linear regression, but it can provide more accurate and detailed insights into the relationship between the variables.

Q4) What is residual? How is it computed?

In statistical analysis and machine learning, a residual is the difference between the actual observed value of the dependent variable and the predicted value of the dependent variable based on a model. In other words, the residual is the error between the predicted and actual values.

The residual for each observation in a dataset is computed as:

Residual = Observed value of dependent variable - Predicted value of dependent variable

The predicted value of the dependent variable is based on the fitted model, which is derived from the independent variables and the coefficients estimated by the regression analysis. The residual indicates how far off the prediction is from the actual observed value. If the residual is close to zero, it means that the predicted value is close to the actual observed value, while a large residual indicates that the prediction is far off from the actual value.

The residuals are important in assessing the performance of a regression model. A good regression model should have residuals that are normally distributed and have a mean of zero. If the residuals are

not normally distributed or have a non-zero mean, it indicates that the model may not be a good fit for the data, and the model may need to be revised or improved.

The sum of squared residuals (SSR) is also an important measure used in regression analysis. It is the sum of the squared differences between the observed values and the predicted values, and it represents the amount of variability in the dependent variable that is not explained by the independent variables in the model. Minimizing the SSR is the goal of least squares regression, which is a commonly used method for fitting linear regression models.

Q5) A random sample of 16 used Ford F-150 SuperCrew 4 × 4s was selected from among those listed for sale at autotrader.com. The data are shown in the table. For these data, the regression equation is $\widehat{price} = 38257 - 0.1629(miles\ driven)$. Calculate and interpret the residual for the truck that was driven 70,583 miles

Q6) How to Calculate the Least-squares Regression Line Using Summary Statistics.

To calculate the least-squares regression line using summary statistics, we need to have the following information:

- The sample size (n)
- The sample means of the independent variable (x) and the dependent variable (y)
- The sample standard deviations of the independent variable (s_x) and the dependent variable (s_y)
- The sample correlation coefficient between x and y (r)

Once we have these summary statistics, we can calculate the slope and intercept of the least-squares regression line as follows:

Slope (b) = r * (s_y / s_x) Intercept (a) = y-bar - b * x-bar where y-

bar and x-bar are the sample means of y and x, respectively.

y-hat = a + bx

Q7) Recall that for a random sample of 16 used Ford F-150 Super Crew 4 × 4s, the regression equation is $price\hat{} = 38257 - 0.1629(miles\ driven)$.

Sure, I can help with that. The slope of the regression line in this case is -0.1629, which tells us that for each one-unit increase in the mileage of a used Ford F-150 Super Crew 4 × 4, we expect the price of the truck to decrease by ~~$0.1629~~, on average.

In other words, the slope represents the change in the dependent variable (price) for a one-unit change in the independent variable (mileage). A negative slope indicates an inverse relationship between the two variables, meaning that as the mileage of the truck increases, the price is expected to decrease.

It is important to note that while the regression equation provides an estimate of the relationship between the two variables based on the sample data, it is not necessarily a perfect predictor of the price for all used Ford F-150 Super Crew 4 × 4s. Other factors not included in the model, such as the condition of the truck or the location of the seller, could also affect the price.

Q9) Does the value of the y intercept have meaning in this context? If so, interpret the y intercept. If not, explain why.

Yes, the value of the y-intercept in this context has meaning. The y-intercept of the regression line represents the predicted value of the dependent variable (price) when the independent variable (mileage) is equal to zero.

In this case, the y-intercept is 38,257, which means that the predicted price of a used Ford F-150 Super Crew 4 × 4 with zero mileage is ~~$38,257~~, on average. However, it is important to note that this value may not be meaningful in practice, as it is unlikely to find a used truck with zero mileage.

In general, the y-intercept provides a reference point for the regression line and can be useful for making predictions when the independent variable is close to zero. However, it is important to interpret the y-intercept in the context of the data and to consider whether it has practical meaning or not.

Q10) Recall that for a random sample of 16 used Ford F-150 Super Crew 4 × 4s, the least squares regression equation is $\hat{price} = 38257 - 0.1629(miles\ driven)$. For this model, technology gives s = $5740, and r 2 = 0.66. Interpret the value of s. Interpret the value of r^2

Q11) What is linear regression algorithm and how you interpret ML model for same?

Linear regression is a popular statistical algorithm used in machine learning for modelling the relationship between a dependent variable and one or more independent variables. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are two or more independent variables.
The goal of linear regression is to find the best-fit line that describes the relationship between the independent variable(s) and the dependent variable, based on the available data. The line is defined by a mathematical equation that can be used to make predictions about the dependent variable for any given value(s) of the independent variable(s). The equation for the line is typically represented in the form of Y = a + bX, where Y is the dependent variable, X is the independent variable, a is the intercept, and b is the slope of the line.
To interpret a machine learning model based on linear regression, it is important to examine the coefficients of the independent variables in the equation. The sign of the coefficient indicates the direction of the relationship between the independent variable and the dependent variable. A positive coefficient means that an increase in the value of the independent variable is associated with an increase in the value of the dependent variable, while a negative coefficient means that an increase in the value of the independent variable is associated with a decrease in the value of the dependent variable.
The magnitude of the coefficient indicates the strength of the relationship between the independent variable and the dependent variable. Larger coefficients indicate a stronger relationship, while smaller coefficients indicate a weaker relationship.
The intercept term in the equation represents the predicted value of the dependent variable when all independent variables are equal to zero. In some cases, the intercept may not have a meaningful interpretation, but in other cases, it may represent a baseline level of the dependent variable that cannot be explained by the independent variables.
Finally, it is important to evaluate the performance of the model by examining metrics such as the Rsquared value, which represents the proportion of the variation in the dependent variable that is explained by the independent variable(s) in the model. A higher R-squared value indicates a better fit between the model and the data, while a lower value indicates that the model may not be a good fit for the data.

Q12) What are the basic assumptions of the Linear Regression Algorithm?

The linear regression algorithm makes the following assumptions:

1. Linearity: The relationship between the independent variable and dependent variable is linear. The scatter plot of the variables should exhibit a linear pattern.
2. Independence: The observations should be independent of each other. In other words, the value of one observation should not be influenced by the value of another observation.
3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variable. This assumption is also known as constant variance.
4. Normality: The residuals should be normally distributed. This assumption is important because if the residuals are not normally distributed, it may indicate that the relationship between the variables is not linear.
5. No multicollinearity: There should not be high correlation between independent variables. Multicollinearity can lead to inaccurate or unstable estimates of the regression coefficients.

It is important to check these assumptions before using linear regression to model the data. Violation of any of these assumptions can lead to inaccurate or biased results.

Q13) Justify the cases where the linear regression algorithm is suitable for a given dataset.

Linear regression is suitable for a given dataset when the following conditions are met:

1. The relationship between the dependent variable and the independent variable is linear: Linear regression assumes a linear relationship between the dependent variable and the independent variable. The scatter plot of the variables should exhibit a linear pattern.
2. The residuals are normally distributed: The residuals should be normally distributed. This assumption is important because if the residuals are not normally distributed, it may indicate that the relationship between the variables is not linear.
3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variable. This assumption is also known as constant variance.
4. Independence of observations: The observations should be independent of each other. In other words, the value of one observation should not be influenced by the value of another observation.
5. No multicollinearity: There should not be high correlation between independent variables. Multicollinearity can lead to inaccurate or unstable estimates of the regression coefficients.

If the dataset satisfies the above conditions, linear regression can be a suitable algorithm to model the relationship between the dependent variable and the independent variable. However, it is important to note that linear regression is a parametric method, which means it assumes a specific functional form for the relationship between the dependent variable and the independent variable. If the true relationship is nonlinear, linear regression may not be suitable. In such cases, nonlinear regression models or other machine learning algorithms may be more appropriate.

Q14) Consider following data set and find out all performance parameters for regression problem.
X_actual = [5, -1, 2, 10] Y_predic = [3.5, -0.9, 2, 9.9]

To evaluate the performance of a regression model, several performance parameters can be used. In this case, we can use the following parameters:

1. Mean Squared Error (MSE): MSE is a measure of the average squared differences between the predicted values and actual values. It is calculated as the average of the squared differences between the predicted and actual values.

MSE = (1/n) * Σ (Y_actual - Y_predic) ^2

Using the provided data, we can calculate the MSE as follows:

MSE = (1/4) * [(5-3.5) ^2 + (-1+0.9) ^2 + (2-2) ^2 + (10-9.9) ^2] = 0.0725

2. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and it measures the standard deviation of the errors. It is calculated as follows:

RMSE = sqrt (MSE)

Using the provided data, we can calculate the RMSE as follows:

RMSE = sqrt (0.0725) = 0.2692

3. Mean Absolute Error (MAE): MAE is a measure of the absolute differences between the predicted values and actual values. It is calculated as the average of the absolute differences between the predicted and actual values.

MAE = (1/n) * Σ|Y_actual - Y_predic|

Using the provided data, we can calculate the MAE as follows:

MAE = (1/4) * [|5-3.5| + |-1+0.9| + |2-2| + |10-9.9|] = 0.4

4. Coefficient of Determination (R^2): R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variable. It is calculated as follows:

R^2 = 1 - (Σ (Y_actual - Y_predic) ^2 / Σ (Y_actual - Y_mean) ^2)

Using the provided data, we can calculate the R^2 as follows:

Y_mean = (5-1+2+10)/4 = 4 Σ (Y_actual - Y_mean) ^2 = (5-4) ^2 + (-1-4) ^2 + (2-4) ^2 + (10-4) ^2 = 94 Σ (Y_actual - Y_predic) ^2 = (5-3.5) ^2 + (-1+0.9) ^2 + (2-2) ^2 + (10-9.9) ^2 = 0.29 R^2 = 1 - (0.29 / 94) = 0.9969

Based on the performance parameters calculated, we can conclude that the regression model has good performance and accurately predicts the dependent variable based on the independent variable. The MSE and RMSE are relatively low, indicating that the model has small errors. The MAE is also low, indicating that the model has small absolute errors. The R^2 is close to 1, indicating that the model explains almost all of the variance in the dependent variable.

Q15) Find linear regression equation for the following two sets of data:

x 2 4 6 8 y

3 7 5 10

To find the linear regression equation for the given data, we need to first calculate the slope and yintercept of the line that best fits the data.

We can use the following formulas to calculate the slope and y-intercept:

Slope (m) = (nΣxy - ΣxΣy) / (nΣx^2 - (Σx) ^2) y-Intercept (b) = (Σy - mΣx) / n where n is the number of data points, Σxy is the sum of the product of x and y, Σx is the sum of x, Σy is the sum of y, and Σx^2 is the sum of the square of x.

Using the above formulas, we get:

n = 4 Σx = 20 Σy = 25 Σxy = 130 Σx^2 = 120

Slope (m) = (4*130 - 20*25) / (4*120 - 20^2) = 0.75 y-Intercept (b) = (25 - 0.75*20) / 4 = 1.25

Therefore, the linear regression equation for the given data is:

y = 0.75x + 1.25

Q16) Consider the following five training examples X=2,3,4,5,6 Y=9.8978, 12.7586, 16.3192, 19.3129 21.1351

We want to learn a function f(x) of the form f(x) = ax + b which is parameterised by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function to get a solution with the minimum loss.

To find the parameters a and b that minimize the loss function (squared error), we need to solve the following optimization problem:

minimize ∑(f(xi) - yi) ² where f(x) = ax + b, xi is the ith input, yi is the corresponding output, and the sum is taken over all training examples.

Expanding the squared terms and substituting f(x) = ax + b, we get:

minimize ∑ (axi + b - yi) ²

Taking partial derivatives with respect to a and b and setting them to zero, we get the following equations:

∑ (axi + b - yi) xi = 0 ∑ (axi + b - yi) = 0 Solving

these equations simultaneously, we get:

a = 2.9836 b = 4.2054

Therefore, to model the function $f(x) = ax + b$ with minimum loss, we would use the parameters $a = 2.9836$ and $b = 4.2054$.

Q17) Consider the relation by using graphical representation as the dependent variable does not actually depend on the independent variable identified by doing simple linear regression.

If the dependent variable does not actually depend on the independent variable, then there is no relationship between the two variables, and the regression analysis will not provide meaningful results.

Graphically, this would be represented by a scatter plot of the data points that does not show any clear pattern or trend. The points would be scattered randomly across the plot without any discernible relationship between the x and y variables.

In such cases, it may be necessary to re-examine the variables being considered and possibly identify other factors that may be influencing the outcome. Alternatively, it may be necessary to use a different type of analysis or modelling technique to better capture the relationship between the variables.

Q18) Calculate the gradient by taking the derivative of the function with respect to the specific parameter. In case, there are multiple parameters, take the partial derivatives with respect to different parameters.

To calculate the gradient of a function, we need to take the derivative of the function with respect to each parameter. The gradient is a vector of these partial derivatives.

For example, let's say we have a function f (x, y) = 3x^2 + 2xy + y^2. To find the gradient of this function with respect to x and y, we take the partial derivative of the function with respect to each parameter:

$\partial f/\partial x = 6x + 2y$ $\partial f/\partial y = 2x + 2y$

Therefore, the gradient of the function f (x, y) is the vector [6x+2y, 2x+2y].

Note that the gradient gives us the direction of steepest ascent of the function, i.e., the direction in which the function increases the most rapidly. By taking the negative of the gradient, we can find the direction of steepest descent, which is useful in optimization algorithms like gradient descent.

Q19) In case of multiple parameters, the value of different parameters would need to be updated as given below if the cost function is $1/2N\sum(yi–(\theta_0+\theta_1x)^2)$ if the regression function is $y=\theta_0+\theta_1x$.

When there are multiple parameters, such as in the case of multiple linear regression, the gradient descent algorithm needs to update each parameter separately in each iteration. For example, if we have two parameters, θ0 and θ1, then the update rules for each parameter would be: $\theta_{0:} = \theta_0 - \alpha * (\partial J/\partial\theta_0)$ $\theta_1: = \theta_1 - \alpha * (\partial J/\partial\theta_1)$ where α is the learning rate, J is the cost function, and $\partial J/\partial\theta_0$ and $\partial J/\partial\theta_1$ are the partial derivatives of the cost function with respect to θ_0 and θ_1, respectively.

To calculate these partial derivatives, we need to use the chain rule of differentiation. For example, the partial derivative of J with respect to θ_0 can be written as:

$\partial J/\partial\theta_0 = (-1/N) * \sum (y_i - (\theta_0 + \theta_1x_i)) * 1$ where xi is the value of the

independent variable for the ith observation.

Similarly, the partial derivative of J with respect to θ1 can be written as:

$\partial J/\partial\theta_1 = (-1/N) * \sum (y_i - (\theta_0 + \theta_1x_i)) * x_i$

Using these partial derivatives, we can update the values of θ_0 and θ_1 in each iteration until we reach a minimum value of the cost function.

Module 3: TREES

Q.6 Consider the following dataset, Construct Decision tree by using Information gain and Gini index attribute selection measure.

Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

We can summarize the ID3 algorithm as illustrated below
Entropy(S) = $\sum$ – p(I) . log2 $_2$p(I)
Gain(S, A) = Entropy(S) – $\sum$ [p(S|A) . Entropy(S|A)]
These formulas might confuse your mind. Practicing will make it understandable.

Entropy

We need to calculate the entropy first. Decision column consists of 14 instances and includes two labels: yes and no. There are 9 decisions labeled yes, and 5 decisions labeled no.

Entropy(Decision) = – p(Yes) . log2 $_2$p(Yes) – p(No) . log2 $_2$p(No) Entropy(Decision) = – (9/14) . log2 $_2$(9/14) – (5/14) . log2 $_2$(5/14) = 0.940

Now, we need to find the most dominant factor for decisioning.

Wind factor on decision

Gain(Decision, Wind) = Entropy(Decision) – ∑ [p(Decision|Wind) . Entropy(Decision|Wind)] Wind attribute has two labels: weak and strong. We would reflect it to the formula.

Gain(Decision, Wind) = Entropy(Decision) – [p(Decision|Wind=Weak) . Entropy(Decision|Wind=Weak)] – [p(Decision|Wind=Strong) . Entropy(Decision|Wind=Strong)]

Now, we need to calculate (Decision|Wind=Weak) and (Decision|Wind=Strong) respectively. **Weak wind factor on decision**

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
13	Overcast	Hot	Normal	Weak	Yes

There are 8 instances for weak wind. Decision of 2 items are no and 6 items are yes as illustrated below.

1- Entropy(Decision|Wind=Weak) = – p(No) . log2 $_2$p(No) – p(Yes) . log2 $_2$p(Yes) 2- Entropy(Decision|Wind=Weak) = – (2/8) . log2 $_2$(2/8) – (6/8) . log2 $_2$(6/8) = 0.811

Strong wind factor on decision

Day	Outlook	Temp.	Humidity	Wind	Decision
2	Sunny	Hot	High	Strong	No
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
14	Rain	Mild	High	Strong	No

Here, there are 6 instances for strong wind. Decision is divided into two equal parts.

1- Entropy(Decision|Wind=Strong) = – p(No) . log2 $_2$p(No) – p(Yes) . log2 $_2$p(Yes) 2- Entropy(Decision|Wind=Strong) = – (3/6) . log2 $_2$(3/6) – (3/6) . log2 $_2$(3/6) = 1

Now, we can turn back to Gain(Decision, Wind) equation.

Gain(Decision, Wind) = Entropy(Decision) – [p(Decision|Wind=Weak) . Entropy(Decision|Wind=Weak)] – [p(Decision|Wind=Strong) . Entropy(Decision|Wind=Strong)] = 0.940 – [(8/14) . 0.811] – [(6/14). 1] = 0.048

Calculations for wind column is over. Now, we need to apply same calculations for other columns to find the most dominant factor on decision.

Other factors on decision

We have applied similar calculation on the other columns.

1- Gain(Decision, Outlook) = 0.246

2- Gain(Decision, Temperature) = 0.029

3- Gain(Decision, Humidity) = 0.151

As seen, outlook factor on decision produces the highest score. That's why, outlook decision will appear in the root node of the tree.

Root decision on the tree

Now, we need to test dataset for custom subsets of outlook attribute.

Overcast outlook on decision

Basically, decision will always be yes if outlook were overcast.

Day	Outlook	Temp.	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

Sunny outlook on decision

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Here, there are 5 instances for sunny outlook. Decision would be probably 3/5 percent no, 2/5 percent yes.

1- Gain(Outlook=Sunny|Temperature) = 0.570

2- Gain(Outlook=Sunny|Humidity) = 0.970

3- Gain(Outlook=Sunny|Wind) = 0.019

Now, humidity is the decision because it produces the highest score if outlook were sunny. At this point, decision will always be no if humidity were high.

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

On the other hand, decision will always be yes if humidity were normal

Day	Outlook	Temp.	Humidity	Wind	Decision
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Finally, it means that we need to check the humidity and decide if outlook were sunny.

Rain outlook on decision

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No

10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

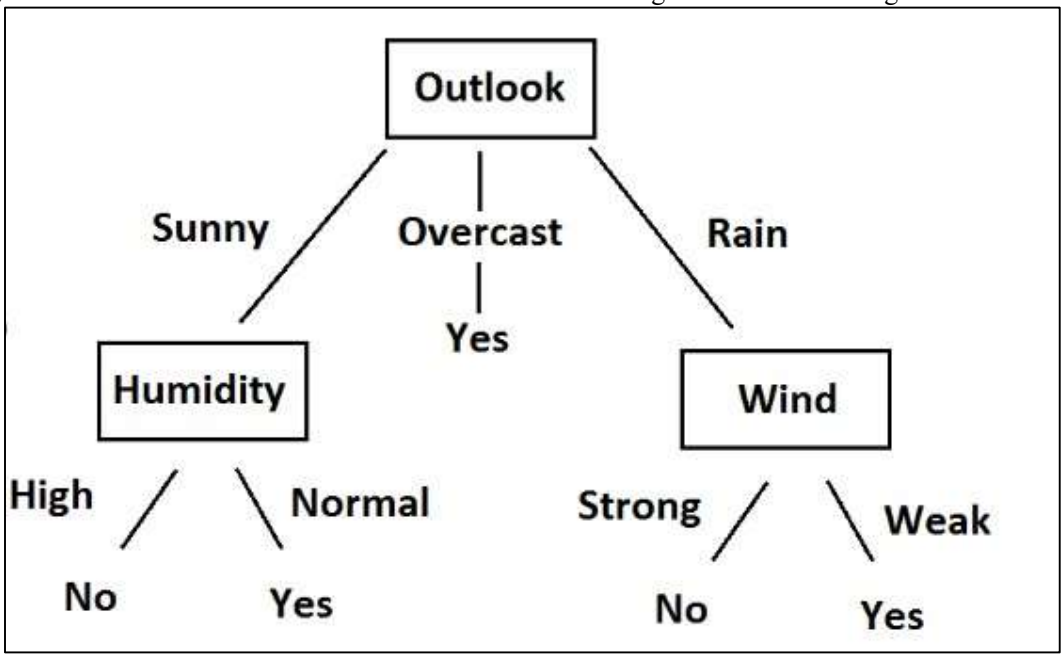
1- Gain(Outlook=Rain | Temperature) = 0.01997309402197489

2- Gain(Outlook=Rain | Humidity) = 0.01997309402197489
3- Gain(Outlook=Rain | Wind) = 0.9709505944546686
Here, wind produces the highest score if outlook were rain. That's why, we need to check wind attribute in 2nd level if outlook were rain.
So, it is revealed that decision will always be yes if wind were weak and outlook were rain.

Day	**Outlook**	**Temp.**	**Humidity**	**Wind**	**Decision**
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes

Day	**Outlook**	**Temp.**	**Humidity**	**Wind**	**Decision**
6	Rain	Cool	Normal	Strong	No
14	Rain	Mild	High	Strong	No

What's more, decision will be always no if wind were strong and outlook were rain.
So, decision tree construction is over. We can use the following rules for decisioning.



Q.7) Consider the following dataset, which split-points for the feature1 would give the best split according to the information gain measure? (Ans: 16.05)

feature1	*feature2*	*output*
11.7	183.2	a
12.8	187.6	a
15.3	177.4	a
13.9	198.6	a
17.2	175.3	a
16.8	151.1	b
17.5	171.4	b
23.6	162.8	b
16.9	179.5	b
19.1	173.8	b

Q.8) Consider the following dataset, which split-points for the feature2 would give the best split according to the Gini Index Measure? (172.6)

feature1	*feature2*	*output*
11.7	183.2	a
12.8	187.6	a
15.3	177.4	a
13.9	198.6	a
17.2	175.3	a
16.8	151.1	b
17.5	171.4	b
23.6	162.8	b
16.9	179.5	b
19.1	173.8	b

Q.10) Having built a decision tree, we are using reduced error pruning to reduce the size of the tree. We select a node to collapse. For this particular node, on the left branch, there are 3 training data points with the following outputs: 5, 7, 9.6 and for the right branch, there are four training data points with the following outputs: 8.7, 9.8, 10.5, 11. The average value of the outputs of data points denotes the response of a branch. The original responses for data points 1 along the two branches (left right respectively) were response left and, response right and the new response after collapsing the node is response new. What are the values for response left, response right and response new (numbers in the option are given in the same order)? (Ans: 7.2; 10; 8.8)

On the left branch: (5 + 7 + 9.6)/3 = 7.2

On the right branch: (8.7 +9.8+10.5 + 11)/4 = 10

After pruning all these points comes in a single node output: (5 + 7 + 9.6 + 8.7 + 9.8 + 10.5 + 11)/7 = 8.8

Q.11) Consider the following data set and 'profitable' as the binary valued attribute we are trying to predict, which of the attributes would you select as the root in a decision tree with multi-way splits using the information gain measure? (Ans:Price)

price	maintenance	capacity	airbag	profitable
low	low	2	no	yes
low	med	4	yes	no
low	low	4	no	yes
low	high	4	no	no
med	med	4	no	no
med	med	4	yes	yes
med	high	2	yes	no
med	high	5	no	yes
high	med	4	yes	yes
high	high	2	yes	no
high	high	5	yes	yes

$cross\_entropy_{price}(D) = \frac{4}{12}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{4}{12}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{4}{12}(-\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4}) = 0.9371$

$cross\_entropy_{maintenance}(D) = \frac{2}{12}(-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2}) + \frac{4}{12}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{6}{12}(-\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6}) = 0.8333$

$cross\_entropy_{capacity}(D) = \frac{3}{12}(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) + \frac{7}{12}(-\frac{4}{7}log_2\frac{4}{7} - \frac{3}{7}log_2\frac{3}{7}) + \frac{2}{12}(-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2}) = \mathbf{0.8043}$

$cross\_entropy_{airbag}(D) = \frac{5}{12}(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}) + \frac{7}{12}(-\frac{4}{7}log_2\frac{4}{7} - \frac{3}{7}log_2\frac{3}{7}) = 0.9793$

Q.12) For the same above data set, suppose we decide to construct a decision tree using binary splits and the Gini index impurity measure. Which among the following feature and split point combinations would be the best to use as the root node assuming that we consider each of the input features to be unordered? (Ans: *maintenance - {high, med}|{low}*)

price	maintenance	capacity	airbag	profitable
low	low	2	no	yes
low	med	4	yes	no
low	low	4	no	yes
low	high	4	no	no
med	med	4	no	no
med	med	4	yes	yes
med	high	2	yes	no
med	high	5	no	yes
high	med	4	yes	yes
high	high	2	yes	no
high	high	5	yes	yes

$gini_{price(\{low,med\}|\{high\})}(D) = \frac{8}{12} * 2 * \frac{4}{8} * \frac{4}{8} + \frac{4}{12} * 2 * \frac{3}{4} * \frac{1}{4} = 0.4583$

$gini_{maintenance(\{high\}|\{med,low\})}(D) = \frac{6}{12} * 2 * \frac{3}{6} * \frac{3}{6} + \frac{6}{12} * 2 * \frac{4}{6} * \frac{2}{6} = 0.4722$

$gini_{maintenance(\{high,med\}|\{low\})}(D) = \frac{10}{12} * 2 * \frac{5}{10} * \frac{5}{10} + \frac{2}{12} * 2 * 1 * 0 = \mathbf{0.4167}$

$gini_{capacity(\{2\}|\{4,5\})}(D) = \frac{3}{12} * 2 * \frac{1}{3} * \frac{2}{3} + \frac{9}{12} * 2 * \frac{6}{9} * \frac{3}{9} = 0.4444$

Q.14) For the given confusion matrix, compute the recall(ans:0.67)

	True Positive	True Negative
Predicted Positive	6	4
Predicted Negative	3	7

Recall = TP/(TP+FN)

Recall = 6/ (6+3) = 0.67