Quiz-Course 9. IBM Machine Learning with Python (Coursera)

Week 1- Practice Quiz – Intro to Machine Learning

1.Supervised learning deals with unlabeled data, while unsupervised learning deals with labelled data.

- **True**
- False

2.The "Regression" technique in Machine Learning is a group of algorithms that are used for:

- **Predicting a continuous value; for example predicting the price of a house based on its characteristics.**
- Finding items/events that often co-occur; for example grocery items that are usually bought together by a customer.
- Prediction of class/category of a case; for example, a cell is benign or malignant, or a customer will churn or not.
  - → Regression techniques are used for continuous variable prediction, whereas classification techniques handle dependent variables with discrete classes.

3. When comparing Supervised with Unsupervised learning, is this sentence True or False?
In contrast to Supervised learning, Unsupervised learning has more models and more evaluation methods that can be used in order to ensure the outcome of the model is accurate.

- True
- **False**
  - → Unsupervised learning has fewer models and evaluation methods than Supervised learning.
  - → Unsupervised learning Has more difficult algorithms than supervised learning since we know little to no information about the data

Week 1- Quiz: Graded Quiz: Intro to Machine Learning

1.In a dataset, what do the columns represent?

- Independent Variables
- Observations
- **Features**
- Variable Type

2.What is a major benefit of **unsupervised learning** over supervised learning?

- **Discover previously unknown information about the dataset**.
- Explore the relationship between features and the target.
- Being able to produce a prediction based on unlabelled data.
- Better evaluates the performance of a built model.

3.What's the correct order for using a model?

- Split the data into training and test sets, fit the model on the train set, evaluate model accuracy.
- Clean the data, fit the model on the entire dataset, split the data into training and test sets, evaluate model accuracy.
- **Clean the data, split the data into training and test sets, fit the model on the train set, evaluate model accuracy.**
- Split the data into the training and test sets, fit the model on the train set, clean the data, evaluate model accuracy.

4. Which of the following is suitable for an **unsupervised learning**?

- Examine the relationship between academic performance and level of in-class participation using observations that include a feature recording each student's grade.
- Classifying benign and malignant tumors using historical data on tumor shape, color, etc.
- Predict house price based on location, house size, and number of rooms.
- **Segment customers into groups for discovering similar characteristics between them.**

5.The main purpose of the NumPy library is to:

- Visualize results in 2D and 3D plots.
- Achieve scientific computations.
- **Perform computations on arrays efficiently.**
- Construct machine learning models.

Week 2- Practice Quiz: Regression

1.Which of the following is the meaning of "Out of Sample Accuracy" in the context of evaluation of models?

- **"Out of Sample Accuracy" is the percentage of correct predictions that the model makes on data that the model has NOT been trained on**.
- "Out of Sample Accuracy" is the percentage of correct predictions that the model makes using the test dataset.
- "Out of Sample Accuracy" is the accuracy of an overly trained model (which may capture noise and produced a non-generalized model)
- "Out of Sample Accuracy" is the accuracy of a model on all the data available.
  - → Out-of-sample accuracy represents how well the model is able to perform on unknown data.

Quiz-Course 9. IBM Machine Learning with Python (Coursera)

2.When should we use Multiple Linear Regression? (Select two)

- When there are multiple dependent variables
- When we would like to examine the relationship between multiple variables.
- **When we would like to identify the strength of the effect that the independent variables have on a dependent variable**.
- **When we would like to predict impacts of changes in independent variables on a dependent variable.**
  - → Multiple linear regression is used for regression tasks involving more than one independent variable.
  - → We hope to understand how the dependent variable change when we change the independent variables.

3. Which sentence is TRUE about linear regression?

- Simple linear regression requires a linear relationship between the predictor and the response, but multiple linear regression does not.
- **A linear relationship is necessary between the independent variables and the dependent variable.**
- Multiple linear regression requires a linear relationship between the predictors and the response, but simple linear regression does not.
- A linear relationship is necessary between the independent and dependent variables as well as in between independent variables.
  - → If the relationship is non-linear, then we must use non-linear regression. One way to understand the relationship is doing a **scatter plot**

Week 2- Graded Quiz: Regression

1.What are the requirements for independent and dependent variables in regression?

- Independent and dependent variables can be either categorical or continuous.
- Independent variables must be continuous. Dependent variables can be either categorical or continuous.
- **Independent variables can be either categorical or continuous. Dependent variables must be continuous.**
- Independent and dependent variables must be continuous.

2.The key difference between simple and multiple regression is:

- Simple regression assumes a linear relationship between variables, whereas this assumption is not necessary for multiple regression.
- **To estimate a single dependent variable, simple regression uses one independent variable whereas multiple regression uses multiple.**
- Multiple linear regression introduces polynomial features.
- Simple linear regression compresses multidimensional space into one dimension.

3.Recall that we tried to predict CO2 emission with car information. Say that now we can describe the relationship as:
CO2_emission = 130 - 2.4*cylinders + 8.3*fuel_consumption What is TRUE of this relationship?

- **When "cylinders" decreases by 1 while fuel_consumption remains constant, CO2_emission increases by 2.4 units.**
- When "cylinders" increases by 1 while fuel_consumption remains constant, CO2_emission increases by 2.4 units.
- Since the coefficient for "fuel_consumption" is greater than that for "cylinders", "fuel_consumption" has higher impact on CO2_emission.
- When both "cylinders" and "fuel_consumption" increase by 1 unit, CO2_emission decreases.

4.What could be the cause of a model yielding **high training accuracy** and **low out-of-sample accuracy**?

- The model is training on the entire dataset, so it is underfitting.
- The model is training on a small training set, so it is underfitting.
- **The model is training on the entire dataset, so it is overfitting.**
- The model is training on a small training set, so it is overfitting.

5.**Multiple Linear Regression** is appropriate for:

- Predicting whether a drug is effective for a patient based on her characteristics.
- Predicting the sales amount based on month.
- **Predicting tomorrow's rainfall amount based on the wind speed and temperature.**

Week 3- Practice Quiz: Classification (KNN and Decision Tree)

1. Which one is TRUE about the **kNN algorithm**?

- **kNN algorithm can be used to estimate values for a continuous target**.
- kNN calculates similarity by measuring how close the two data points' response values are.
- The most similar point in kNN is the one with the smallest distance averaged across all normalized features.
- kNN is a classification algorithm that takes a bunch of unlabelled points and uses them to learn how to label other points.
  - → kNN is a supervised algorithm, so it requires a dataset with predefined labels to predict new cases.
  - → kNN can be used for both classification and regression prediction tasks. In the case of a continuous target, the prediction is taken as the average or median of the nearest neighbours.
  - → **(Response value= continuous**): To predict a value for a new case, **average or median** target value of the nearest neighbors is used
  - → **(Response value= categorical):** To predict a value for a new case with k=5, take the **majority vote among 5 closest points** (majority vote among 5 points whose features are closest to the new observation.)
  - → **(Response value= categorical):** kNN calculates similarity by determining the smallest distance based on two data points' features. When there are multiple features, kNN sums up the distance between two data points

2.If the information gain of the tree by using attribute A is 0.3, what can we infer?

- **The entropy of a tree before split minus weighted entropy after split by attribute A is 0.3**.
- Entropy in the decision tree increases by 0.3 if we make this split.
- By making this split, we increase the randomness in each child node by 0.3.
- Compared to attribute B with 0.65 information gain, attribute A should be selected first for splitting.

→ The entropy of a tree before split minus weighted entropy after split by attribute A is 0.3. This describes how information gain is calculated, measuring how much certainty has increased by making a split.
→ **There is an inverse relationship between information gain and entropy**. Entropy in the decision tree decreases by 0.3 if we make this split.
→ The randomness in each node is entropy, so with a positive information gain, we decrease randomness.
→ The tree should select the attribute yielding higher information gain (attribute B with 0.65 information gain) compared to attribute A with 0.3 information gain.

3.When we have a value of K for **KNN** that's too small, what will the model most likely look like?

- The model will be overly simple and does not capture enough noise.
- **The model will be highly complex and captures too much noise.**
- The model will have high accuracy on the test set.
- The model will have high out-of-sample accuracy.

→ By looking at too few neighbours, we can capture an anomaly in the data, which means that prediction isn't generalized enough.

Week 3- Graded Quiz: Classification

1.What can we infer about our kNN model when the value of **K is too big**?

- **The model is overly generalized and underfitted to the data.**
- The model will capture a lot of noise as a result of overfitting.
- The model will be too complex and not interpretable.
- The training accuracy will be high, while the out-of-sample accuracy will be low.

The value of k is too small (e.g. k=1)	**The value of k is too big** (e.g. k=20)
• **The model will be highly complex and captures too much noise.** • **The model is overfitted** • **The prediction is NOT generalized enough**	• **The model is overly generalized and underfitted to the data.**

→ If the K is **too small** (e.g. k=1) -> 1-we can capture the noise or anomaly in the data 2- also causes a highly complex model -> overfitting of the model (cannot be trusted for the out-of-sample cases (unknown samples) (To what extent can we trust our judgment which is based on the first nearest neighbor? It might be a poor judgment especially if the first nearest neighbor is a very specific case or an outlier.)
→ If the K is **too high** (e.g. k=20) -> model becomes overly generalized and underfitted

2.When splitting data into branches for a decision tree, what kind of feature is favored and chosen first?

- The feature with the greatest number of categories.
- **The feature that increases purity in the tree nodes**.
- The feature that increases entropy in the tree nodes.
- The feature that splits the data equally into groups.

3.What is the relationship between entropy and information gain?

- High entropy and low information gain is desired.
- When information gain decreases, entropy decreases.
- High entropy and high information gain is desired.
- **When information gain increases, entropy decreases**.

4.Predicting whether a customer responds to a particular advertising campaign or not is an example of what?

- **Classification problem**
- Machine learning
- Regression
- None of the above

5.For a new observation, how do we predict its response value (categorical) using a KNN model with k=5?

- **Take majority vote among 5 points whose features are closest to the new observation**.
- Take the majority vote among 5 points who are the most similar to each other.
- Take the average among 5 points whose features are closest to the new observation.
- Form 5 clusters and assign the new observation to the most similar cluster, taking the mean value as prediction.

Quiz-Course 9. IBM Machine Learning with Python (Coursera)

Week 4- Practice Quiz: Linear Classification (Logistic Regression and SVM)

1.Which of the following examples is/are a sample application of **Logistic Regression**? (select three)

- **Likelihood of a homeowner defaulting on a mortgage.**
- **The probability that a person has a heart attack within a specified time period using person's age and sex.**
- **Customer's propensity to purchase a product or halt a subscription in marketing applications.**
- Estimating the blood pressure of a patient based on her symptoms and biographical data.

→ First answer: we try to predict the possibility of defaulting versus not defaulting, which is a categorical response.
→ Second answer: The outcome is binary and uses other variables as predictors.
→ Third answer: The outcome is a probability of a categorical variable.

2.Which of the following statements comparing linear and logistic regressions is TRUE?

- Both linear and logistic regression can be used to predict categorical responses and attain a point's likelihood of belonging to each class.
- Independent variables in linear regression can be continuous or categorical, but can only be categorical in logistic regression.
- In this course, linear regression minimizes the mean absolute error, while logistic regression minimizes the mean squared error.
- **Linear regression is used for a continuous target whereas logistic regression is more suitable for a categorical target.**

→ Linear regression is not suitable for a categorical target because it tries to fit a line through the data, but the prediction is a step function that doesn't reflect class probability well.

3.How are gradient descent and learning rate used in **logistic regression**?

- Gradient descent will minimize learning rate to minimize the cost in fewer iterations.
- Gradient descent takes increasingly bigger steps towards the minimum with each iteration.
- We want to minimize the cost by maximizing the learning rate value.
- **Gradient descent specifies the steps to take in the current slope direction, learning rate is the step length.**

→ **Gradient descent** takes steps toward the minimum of the cost function and
→ **Learning rate** gives us control over how fast we move.

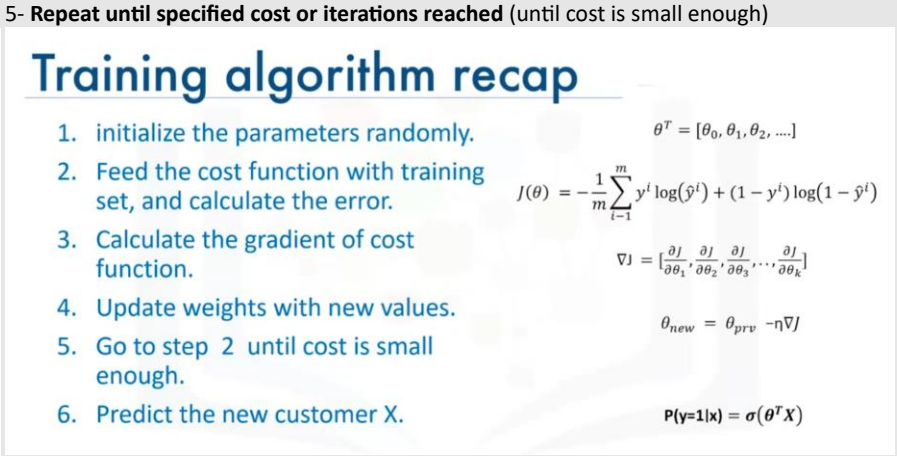
Week 4- Graded Quiz: Linear Classification

1.Which option lists the steps of training a **logistic regression model** in the correct order?

1. Use the cost function on the training set.
2. Update weights with new parameter values.
3. Calculate cost function gradient.
4. Initialize the parameters.
5. Repeat until specified cost or iterations reached.

- 1, 4, 3, 2, 5
- 3, 2, 5, 4, 1
- 4, 3, 2, 5, 1
- **4, 1, 3, 2, 5**

➔ 1- **Initialize the parameters** (pick some parameters randomly)
➔ 2- **Use the cost function on the training set** (first calculate the cost (error) for one case 1, then for the rest, i.e. add up the errors: J(θ)
--Note: the calculated cost (error) is high, as the initialized parameters was picked randomly) -> we will need new θs at some point
--*The coefficients (θ) refer to "weights", "weights factor" or "confidences of the equation"*
➔ 3- **Calculate cost function gradient**
➔ 4- **Update weights with new parameter values** (As the initial parameters led to high cost (error), the parameters need to be updated)
➔ 5- **Repeat until specified cost or iterations reached** (until cost is small enough)



2.What is the objective of **Support Vector Machine (SVM)** in terms of hyperplanes?

- Choose the hyperplane that's closest to one of the two classes.
- Minimize the distance between hyperplane and the support vectors.
- Find the hyperplane of the lowest dimension.
- **Choose the hyperplane that represents the largest margin between the two classes.**
  - → To find the hyperplane, we are looking for **largest margins from support vectors**
  - → The closest to the hyperplane are support vectors. Therefore, finding the hyperplane in such a way that it has the maximum distance to support vectors, will achieve the objective.

3.**Logistic regression** is used to predict the probability of a:

- Numerical dependent variable
- **Categorical dependent variable**
- Categorical independent variable
- Numerical independent variable

4.In which cases would we want to consider using **Support Vector Machine (SVM)**?

- When we desire efficiency when using large datasets.
- When we desire probability estimates for each class.
- When we want multiple decision boundaries with varying weights.
- **When mapping the data to a higher dimensional feature space can better separate classes.**
  - → Some applications of SVM include **text mining tasks** (detecting spam, text category assignment and sentiment analysis) and **gene expression data classification.** This is especially due its power in high-dimensional data classification.

5.What is a disadvantage of **one-vs-all classification**?

- **There's an ambiguous region where multiple classes are valid outputs.**
- It cannot output probability estimates of classes.
- It does not handle two-class classification well.
- It requires more models to be created compared to one-vs-one.
  - → In Multi-class classification, we classify data into multiple class labels (vs. binary classification)
  - → Multi-class classification techniques that can convert most two-class classifiers to a multi-class classifier: **One vs. All** and **One vs One**
  - → Read more at https://www.coursera.org/learn/machine-learning-with-python/supplement/sPzzF/multiclass-prediction

Week 5- Practice Quiz: Clustering

1.Which of the following is an application of **clustering**?

- **Customer segmentation**
- Price estimation
- Customer churn prediction
- Sales prediction

2.Which approach can be used to calculate dissimilarity of objects in **clustering**?

- Cosine similarity
- Minkowski distance
- Euclidian distance
- **All of the above**

3.How is a center point (centroid) picked for each cluster in k-means upon initialization? (select two)

- **We can create some random points as centroids of the clusters.**
- **We can randomly choose some observations out of the data set and use these observations as the initial means.**
- We can select it through correlation analysis.
- We select the k points closest to the mean/median of the entire dataset.
  - → Answer 1: We can randomly place k centroids, one for each cluster. Each data point is then assigned to its closest centroid.
  - → Answer 2: These centroids will be updated based on the clusters formed at each iteration.

Week 5- Graded Quiz: Clustering

1.The objective of k-means clustering is:

- Minimize the cost function via gradient descent
- Maximize the number of correctly classified data points
- **Separate dissimilar samples and group similar ones**
- Yield the highest out of sample accuracy

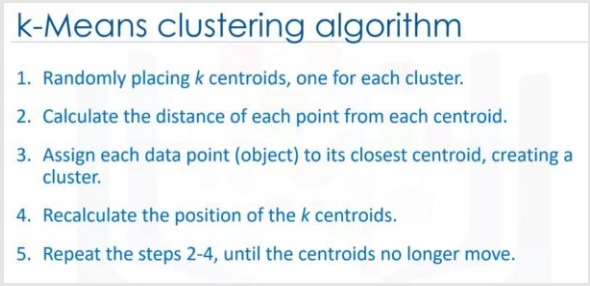
Quiz-Course 9. IBM Machine Learning with Python (Coursera)

2.Which option correctly orders the steps of k-means clustering?

1. Re-cluster the data points
2. Choose k random observations to calculate each cluster's mean
3. Update centroid to take cluster mean
4. Repeat until centroids are constant
5. Calculate data point distance to centroids

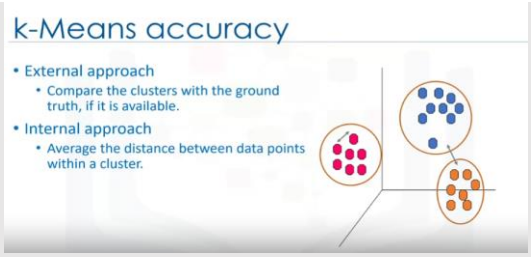
- **2, 5, 3, 1, 4**
- 2, 1, 4, 5, 3
- 2, 3, 4, 5, 1
- 3, 5, 1, 4, 2

→ 1. Choose k random observations to calculate each cluster's mean (First randomly placing k centroids, one for each cluster)
→ 2. Calculate data point distance to centroids (for instance using Euclidean distance)
→ 3. Update centroid to take cluster mean (assign each data point or object to its closest centroid creating a group)
→ 4. Re-cluster the data points (recalculate the position of the k centroids: The new centroid position determined by the mean of all points in the group)
→ 5. Repeat until centroids are constant (Repeat the steps 2-4, until the centroids no longer move)



3.How can we gauge the performance of a k-means clustering model when ground truth is not available?

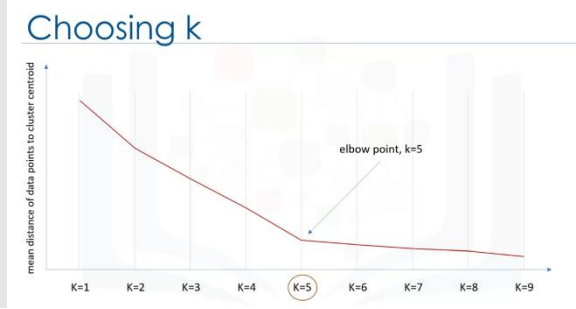
- Determine the prediction accuracy on the test set.
- Calculate the number of incorrectly classified observations in the training set.
- Calculate the R-squared value to measure model fit.
- **Take the average of the distance between data points and their cluster centroids.**



4.When the parameter K for k-means clustering increases, what happens to the error?

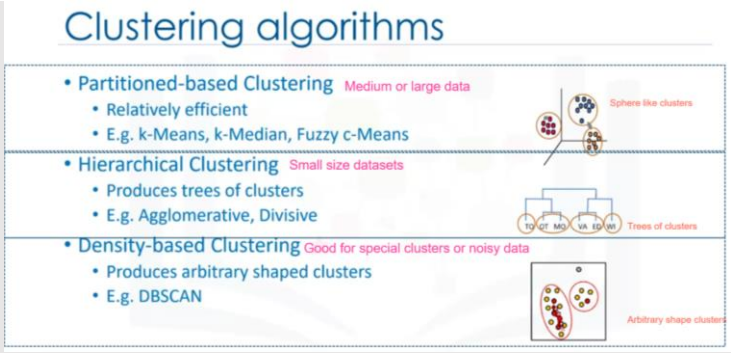
- It will increase because incorrectly classified points are further from the correct centroid.
- It will decrease because the data points are less possible to be in the wrong cluster.
- **It will decrease because distance between data points and centroid will decrease**.
- It might increase or decrease depending on if data points are closer to the centroid.

→ The problem is that with increasing the number of clusters (k), the distance of centroids to data points will always reduce. This means increasing K will **always** decrease the error-> that's why do elbow method (The mean distance of centroids to data points is plotted and the elbow point is determined where the rate of decrease sharply shifts. It is the right K for clustering. This method is called the elbow method.)



5.Which of the following is true for partition-based clustering but not hierarchical nor density-based clustering algorithms?

- **Partition-based clustering produces sphere-like clusters.**
- Partition-based clustering is a type of unsupervised learning algorithm.
- Partition-based clustering can handle spatial clusters and noisy data.
- Partition-based clustering produces arbitrary shaped clusters.



Week 6- Final Exam

1.Which of the following is **not true** about Machine Learning?

- Machine Learning was inspired by the learning process of human beings.
- Machine Learning models help us in tasks such as object recognition, summarization, and recommendation.
- **Machine learning gives computers the ability to make decision by writing down rules and methods and being explicitly programmed.**
- Machine Learning models iteratively learn from data and allow computers to find hidden insights.

→ Machine learning can learn without explicitly being programmed to do so.

2.Which of the following is a Machine Learning technique?

- Clustering
- Classification
- Regression/Estimation
- **All of the above**

→ All of the above are considered machine learning techniques along with association, anomaly detection, sequence mining, and recommendation systems.

3.In which of the following would you use **Multiple Linear Regression**?

- **Predicting the production of apples in an orchard based on temperature and rainfall.**
- Recommend products to customers based on their demographic characteristics.
- Predict whether a customer is likely to repay a loan based on age and income.
- Predict CO_2 emission of a car based on engine size.

→ We use multiple linear regression when there is more than one independent variable for predicting a continuous variable.

4.Which of the below is an example of a classification problem?

- To predict the category to which a customer belongs to.
- To predict whether a customer switches to another provider/brand.
- To predict whether a customer responds to a particular advertising campaign or not.
- **All of the above.**

→ All of these are examples of logistic regression as they try to predict the probability of a binary response.

5.Which of the following is an example of Logistic Regression?

- The odds of a particular individual having a heart attack based on how much they exercise and how much they weigh.
- The probability of a borrower defaulting on their mortgage based upon their credit score and age.
- The probability of a person purchasing life insurance based on age and income.
- **All of the above**.

6.What type of clustering divides the data into non-overlapping subsets without any cluster-internal structure?

- **k-mean clustering**
- Hierarchical clustering
- DBSCAN
- None of the above

→ Other algorithms divide data into clusters of varying shapes.

Quiz-Course 9. IBM Machine Learning with Python (Coursera)

7.Which one best describes the clustering process for k-means clustering?

- k-means clustering creates a tree of clusters.
- k-means divides the data into clusters with minimal overlap such that there are low chances of dissimilar samples in the same cluster.
- **The objective of k-means is to form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters.**
- k-means creates clusters by grouping data points with similar labels.

→ K-Means seeks to create non-overlapping clusters.

8.What is a statistical model that uses Logistic function to model the conditional probability?

- Ridge regression
- Linear regression
- Stepwise regression
- **Logistic regression**

→ Logistic regression uses the **logistic cost function** to return the probability of each class.

→ Logistic regression can not only be used for classification, but also determine the probability. The probability is determined by **Logistic function = Sigmoid function**. The output is probability (output is always between 0 and 1)

9.Precision and recall are suitable for measuring the performance of which tasks?

- **Classification**
- Clustering
- Regression
- All of the above

→ Precision measures the accuracy of class label prediction, whereas Recall is the true positive rate.

Measure of Accuracy		
Linear Regression	**Classification (Logistic R, Decision Tree, KNN, SVM)**	**Clustering (K Means Clustering)**
• **MAE: Mean Absolute Error** • **MSE: Mean Squared Error** • **RMSE: Root Mean Squared Error** • **RAE: Relative Absolute Error (**also known as **residual sum of square)** • **RSE: Relative Squared Error** • **R^2 = 1- RSE** (0-1: best:1) (Is not an error per say but is a popular metric for the accuracy of the model. It shows how close the data values are to the fitted regression line. The higher the R^2, the better the model fits data)	• **Jaccard index** (0-1: best:1) • **F1-Score** (0-1: best:1) • **Precision** (0-1: best:1) • **Recall** (0-1: best:1) • **Logarithmic loss = Log loss** (0-1: best:0): (It measures the performance of a classifier where the predicted output is a probability value between 0 and 1 (like logistic regression) ->) It only applies to the Logistic Regression	• **Average of the distances of data points from their cluster centroids** can be used as a metric of error for the clustering algorithm. (But the problem is that with increasing the number of clusters, the distance of centroids to data points will always reduce. This means increasing K will always decrease the error. To address the problem: Use elbow method for selecting K.

FP (Type I error)	The predicted value is Positive, but actually is False Type I higher-> Type II lower
FN (Type II error)	The predicted value is Negative, but actually is Positive Precision higher -> Recall lower
Accuracy	The ratio of correct predictions (+ or -) to the total number of predictions. It measures the **overall correctness** of the predictions. (diagonal of matrix/total) Max Accuracy score =1.0
Precision	Measures the ability of the model to correctly identify positive instances out of the total instances it classified as positive. It focuses on the correctness of the positive predictions. Max Precision= 1.0
Recall (Sensitivity)	Measures the ability of the model to correctly identify positive instances out of total actual positive instances. Max Recall= 1.0; *If No False Negative -> Recall=1.0*
F1-Score	Provides a single metric that balances both precision and recall. Sometimes both precision and recall is very important-> we use F1-score for model selection. Max F1-score=1.0

10.Which of the following is more suitable to solve with a decision tree?

- To segment customers into groups with similar characteristics.
- To predict the probability of raining based on current temperature and humidity.
- To predict the salary of a baseball player based on the number of home runs and years in the league.
- **To predict if the person will like a certain movie based on age, favorite actors and genre.**

→ Decision trees can split the data based on age, favorite actors, and genre to output a discrete prediction for whether the person likes/dislikes a movie.