ASSIGNMENT_NO_7

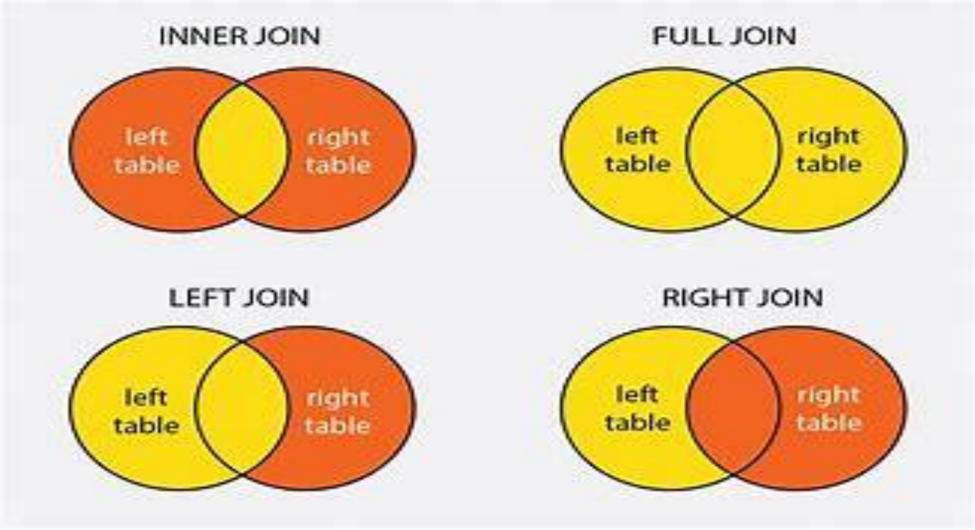
WORKSHEET_NO_7

Question No	Answers
1	B. Candidate keys
2	B. Primary keys cannot contain NULL values… C. A table can have only one primary key with single or multiple fields…
3	C. Insert
4	C. ORDERBY
5	C. SELECT
6	C. 3NF
7	C. All of the above can be done by SQL
8	B. DML
9	B. Table
10	A. 1 NF

Q12 Answer: -

The type of join statement you use depends on your use case. There are four different types of join operations:

- (INNER) JOIN: Returns dataset that have matching values in both tables
- LEFT (OUTER) JOIN: Returns all records from the left table and matched records from the right
- RIGHT (OUTER) JOIN: Returns all records from the right table and the matched records from the left
- FULL (OUTER) JOIN: Returns all records when there is a match in either the left table or right table.



Q11 Answer: -

SQL join statements allow us to access information from two or more tables at once. They also keep our database normalized. Normalization allows us to keep data redundancy low so that we can decrease the amount of data anomalies in our application when we delete or update a record.

Q13 Answer: -

SQL Server is a relational database management system (RDBMS) developed by Microsoft. It is primarily designed and developed to compete with MySQL and Oracle database. SQL Server supports ANSI SQL, which is the standard SQL (Structured Query Language) language. However, SQL Server comes with its own implementation of the SQL language, T-SQL (Transact-SQL). T-SQL is a Microsoft propriety Language known as Transact-SQL. It provides further capabilities of declaring variable, exception handling, stored procedure, etc. SQL Server Management Studio (SSMS) is the main interface tool for SQL Server, and it supports both 32-bit and 64-bit environments.

Q14 Answer: -

A primary key is a field in a table which uniquely identifies each row/record in a database table. Primary keys must contain unique values. A primary key column cannot have NULL values. A table can have only one primary key, which may consist of single or multiple fields. When multiple fields are used as a primary key, they are called a composite key. If a table has a primary key defined on any field(s), then you cannot have two records having the same value of that field(s).

Q15 Answer: -

ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load. It is tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse.

ASSIGNMENT_NO_7

MACHINE LEARNING

Question No	Answers
1	D) All of the above
2	A) Random forest
3	A) The regularization will increase
4	C) both A & B
5	A) It's an ensemble of weak learners
6	C) Both of them
7	B) Bias will decrease, Variance increase
8	C) model is performing good

Q9 Answer: - Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index and entropy of the dataset

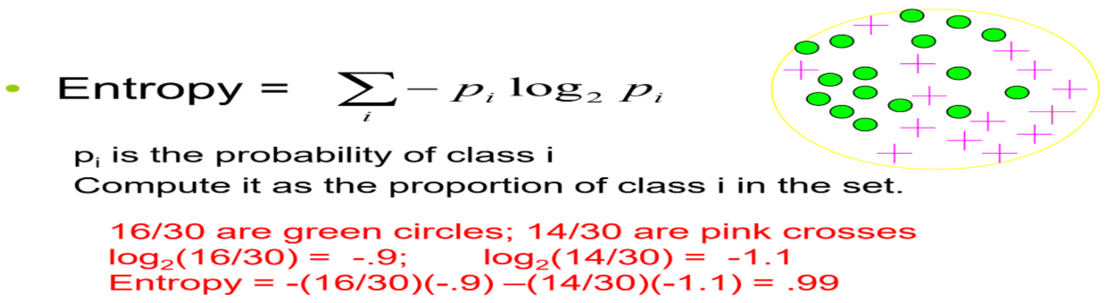
The Gini Index is simply a tree-splitting criterion. When your decision tree has to make a “split” in your data, it makes that split at that particular root node that minimizes the Gini index.

Below, we can see the Gini Index Formula:

$$Gini = 1 - \sum_{j} p_j^2$$

By putting all values, we get, Gini Index = 1 – (0.4^2 + $0.6^{2)}$ = 1-0.52 = 0.48

Entropy is simply “how much variance the dataset has.



- Entropy = $\sum_i - p_i \log_2 p_i$

p_i is the probability of class i
Compute it as the proportion of class i in the set.

16/30 are green circles; 14/30 are pink crosses
$\log_2(16/30)$ = -.9; $\log_2(14/30)$ = -1.1
Entropy = -(16/30)(-.9) –(14/30)(-1.1) = .99

Based on above we can calculate entropy as,

Log2 (40/100) = 0.1204

Log2 (60/100) = 0.1806

Therefore entropy = -(40/100)*(0.1204) – (60/100)*(0.1806) = 0.04816+0.10836 = 0.15625

Q10 Answer: -

Reduced risk of overfitting: Decision trees run the risk of overfitting as they tend to tightly fit all the samples within training data. However, when there's a robust number of decision trees in a random forest, the classifier won't overfit the model since the averaging of uncorrelated trees lowers the overall variance and prediction error.

Provides flexibility: Since random forest can handle both regression and classification tasks with a high degree of accuracy, it is a popular method among data scientists. Feature bagging also makes the random forest classifier an effective tool for estimating missing values as it maintains accuracy when a portion of the data is missing.

Easy to determine feature importance: Random Forest makes it easy to evaluate variable importance, or contribution, to the model. There are a few ways to evaluate feature importance. Gini importance and mean decrease in impurity (MDI) are usually used to measure how much the model's accuracy decreases when a given variable is excluded. However, permutation importance, also known as mean decrease accuracy (MDA), is another importance measure. MDA identifies the average decrease in accuracy by randomly permutating the feature values in job samples.

Q11 Answer: -

It refers to putting the values in the same range or same scale so that no variable is dominated by the other. Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidean distance between two data points in their computations, this is a problem.

If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Min-Max scaling: - min-max scaling is also a type of normalization, we transform the data such that the features are within a specific range e.g. [0, 1]. where x' is the normalized value. It can be easily seen that when x=min, then y=0, and When x=max, then y=1. This means, the minimum value in X is mapped to 0 and the maximum value in X is mapped to 1. So, the entire range of values of X from min to max are mapped to the range 0 to 1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization: - Standardization (also called z-score normalization) transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1. It is the definition that we read in the last paragraph. Where σ is the variance and x̄ is the mean.

$$x' = \frac{x - \bar{x}}{\sigma}$$

Q12 Answer: -

Gradient Descent is an iterative optimization algorithm, used to find the minimum value for a function. The general idea is to initialize the parameters to random values, and then take small steps in the direction of the “slope” at each iteration. Gradient descent is highly used in supervised learning to minimize the error function and find the optimal values for the parameters. Various extensions have been designed for the gradient descent algorithms.

- More stable convergence and error gradient than Stochastic Gradient descent.
- Embraces the benefits of vectorization.
- A more direct path is taken towards the minimum.
- Computationally efficient since updates are required after the run of an epoch.

Q13 Answer: -

Classification accuracy is the most-used metric for evaluating classification models. The reason for its wide use is because it is easy to calculate, easy to interpret, and is a single number to summarize the model’s capability. As such, it is natural to use it on imbalanced classification problems, where the distribution of examples in the training dataset across the classes is not equal. This is the most common mistake made by beginners to imbalanced classification.

When the class distribution is slightly skewed, accuracy can still be a useful metric. When the skew in the class distributions are severe, accuracy can become an unreliable measure of model performance. The reason for this unreliability is centred around the average machine learning practitioner and the intuitions for classification accuracy. Typically, classification predictive modelling is practiced with small datasets where the class distribution is equal or very close to equal. Therefore, most practitioners develop an intuition that large accuracy score (or conversely small error rate scores) are good, and values above 90 percent are great.

Achieving 90 percent classification accuracy, or even 99 percent classification accuracy, may be trivial on an imbalanced classification problem. This means that intuitions for classification accuracy developed on balanced class distributions will be applied and will be wrong, misleading the practitioner into thinking that a model has good or even excellent performance when it, in fact, does not.

Q15 Answer: -

We do this on the training set of data.

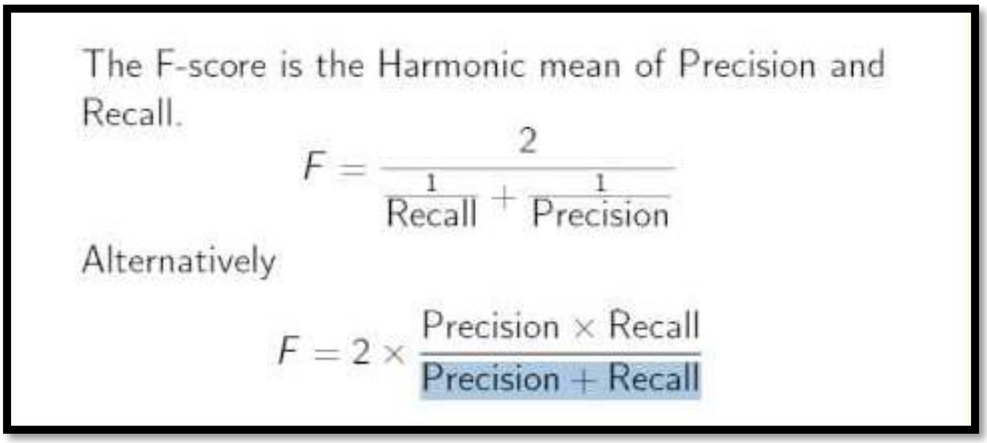
1.Fit (): Method calculates the parameters μ and σ and saves them as internal objects.

2.Transform (): Method using these calculated parameters apply the transformation to a particular dataset.

3.Fit_transform (): joins the fit() and transform() method for transformation of dataset.

Q14 Answer: -

F1-Score or F-measure is an evaluation metric for a classification defined as the harmonic mean of precision and recall. It is a statistical measure of the accuracy of a test or model. Mathematically, it is expressed as follows, here, the value of F-measure(F1-score) reaches the best value at 1 and the worst value at 0. F1-score 1 represents the perfect accuracy and recall of the model.



The F-score is the Harmonic mean of Precision and Recall.

$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

Alternatively

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ASSIGNMENT_NO_7

STATISTICS WORKSHEET_7

Question No	Answers
1	b) 0.135
2	d) 0.53
3	c) 0.745
4	b) 0.577
5	c) 0.6
6	a) 0.33
7	c) 0.33
8	b) 0.22
9	a) 0.66
10	a) 0.33
11	c) 0.5
12	a) 0.166
13	a) 0.375
14	d) 0.06

15	c) 1/2