WORKSHEET ASSIGMENT-7 SQL

ANS1. Candidate key.

ANS2. Primary key cannot contain NULL values.

. A table can have only one primary key with single or multiples fields.

ANS3. Insert

ANS 4. Order by

ANS5. Select

ANS6. 3NF

ANS7. Create database structures only

ANS8. DML

ANS9. Table

ANS10. 2NF

ANS11. In SQL, a join is an operation that allows you to combine data from two or more tables based on a common column or columns. The join operation is used when data needs to be retrieved from multiple tables at the same time.

ANS 12.There are several types of joins in SQL, including:

Inner join: This is the most common type of join in SQL. It returns only the rows that have matching values in both tables.

Left outer join: This join returns all the rows from the left table and the matching rows from the right table. If there are no matching rows in the right table, NULL values are returned.

Right outer join: This join returns all the rows from the right table and the matching rows from the left table. If there are no matching rows in the left table, NULL values are returned.

Full outer join: This join returns all the rows from both tables. If there are no matching rows in either table, NULL values are returned.

Cross join: This join returns the Cartesian product of the two tables, which means that every row from the first table is combined with every row from the second table.

Joins are an essential tool for working with relational databases, as they allow you to combine data from multiple tables in a single query.

ANS 13. SQL server is a rational database management system(RDBMS) developed by Microsoft corporation . It is a software product used to store and mange data on computers and servers. SQL server provides an efficient way to manage and relative data, as well as robust security and backup system.

SQL server uses structure query language (SQL) to manipulate data. SQL is a standard language used to interact with databases and is used to create, modify and delete database, tables and records. SQL server supports various editions, from the free express edition to the more advance Enterprise edition, which can handle large-scale data management and analytics.

SQL server provides a variety of features such as transactions processing, business intelligence , data warehousing, and reporting. It can be used in various applications, from small scale web applications to large enterprise-level systems. SQL server is widely used by businesses and organizations around the world to manage and store their data efficiency and securely.

ANS 14. In SQL, a primary key is a column or a group of column in a tablet that uniquely identifies each row in that table. The primary key ensures that each rows in a table has unique identify and can be accessed and manipulated uniquely.

The primary key is used to enforce data integrity is used to enforce data integrity and to establish relationship between tables in a realtioanl database. It serves as a reference point for other tables that need to link to the data in that table,

In order for a column or a group of column to be a primary key, it must meet the following requirements:

1. Uniqueness: Each value in the column or group of columns must be unique.
2. Non – nullability : The column or group of column cannot contain null values.
3. Irreducibility: The primary key must be composed of the minimum number of columns required to uniquely identify each row in the table.

Some of the benefits of using a primary key includes improved data retrieval speed, data consistency and accurarcy, and the ability to enforce referential integrity constraints when creating relationship between tables in a databse.

ANS 15. ETL stands for Extract, Transform, and Load, which is a process used in database management to move data from one system or database to another. ETL is a critical part of data warehousing and business.

It involves three main steps:

1. Extract: In this step, data is extracted from one or more source systems or databases. The source systems or databases. The source data can be in various formats such as CSV, EXCEL, or a relational database.
2. Transformed: once the data is extracted, it may need to be transformed or cleaned to ensure it can be loaded into the target system. This may includes tasks like filtering, sorting, joining, aggregating, or restructuring data.
3. Load: Finally the transformed data is loaded into the target system, such as data warehouse or a reporting database. The target system may use SQL to store and manage the data.

The ETL process is typically automated using specialized software tools that can perform each step of the process efficiency and accurately. SQL is often used in the transform and load stages of the ETL process, where it can be used to manipulate data and load it into the target system.

MACHINE LEARNING ASSINGNMENT 7

ANS.1-(B)

ANS.2-(A)

ANS.3-(A)

ANS.4-(A)

ANS.8-(B)

ANS.9-

To calculate the Gini index and entropy of the dataset, we need to first calculate the probability of each class:

Probability of class A = 40% = 0.4

Probability of class B = 60% = 0.6

Gini Index:

Gini index is a measure of impurity or inequality of a dataset. It can be calculated as:

Gini index = 1 - (p\_A^2 + p\_B^2)

where p\_A and p\_B are the probabilities of class A and class B, respectively.

Substituting the values, we get:

Gini index = 1 - (0.4^2 + 0.6^2) = 0.48

Therefore, the Gini index of the dataset is 0.48.

Entropy:

Entropy is a measure of impurity or uncertainty of a dataset. It can be calculated as:

Entropy = -p\_A log2(p\_A) - p\_B log2(p\_B)

where p\_A and p\_B are the probabilities of class A and class B, respectively.

Substituting the values, we get:

Entropy = -0.4 log2(0.4) - 0.6 log2(0.6) = 0.971

Therefore, the entropy of the dataset is 0.971.

ANS.10- Decision Tree Random Forest

A decision tree is a tree-like model of decisions along with possible outcomes in a diagram.

There is always a scope for overfitting, caused due to the presence of variance.

The results are not accurate.

Decision trees require low computation, thus reducing time to implement and carrying low accuracy

It is easy to visualize. The only task is to fit the decision tree model.

Random forest:

A classification algorithm consisting of many decision trees combined to get a more accurate result as compared to a single tree.

Random forest algorithm avoids and prevents overfitting by using multiple trees.

This gives accurate and precise results.

This consumes more computation. The process of generation and analyzing its time consuming.

This has complex visualization as it determine the patterns behind the data.

ANS.11- Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.

Two type of scaling:

Comparative scales:

It involves the direct comparison of objects. Comparative scale data must be interpreted in corresponding terms and have either ordinal or rank order properties.

1. Paired comparison:

This technique is a widely used comparative scaling technique.

In this technique, the respondent is asked to pick one object among the two objects with the help of some criterion.

The respondent makes a series of judgments’ between objects.

The data obtained is ordinal in nature.

With n brands, [n(n-1)/2] paired comparisons are required. For example: A survey was conducted to find out consumer’s preference for dark chocolate or white chocolate. The outcome was as follows:

2.Rank order:

In this technique, the respondent judges one item against others.

Respondent are present with several objects and are asked to rank or order them according to some criterion.

Rank order scaling is also ordinal in nature.

Only (n-1) scaling decisions need to be made in this technique.

ANS.12-Gradient Descent is the most common optimization algorithm in machine learning and deep learning. It is a first-order optimization algorithm. This means it only takes into account the first derivative when performing the updates on the parameters. On each iteration, we update the parameters in the opposite direction of the gradient of the objective function J(w) w.r.t the parameters where the gradient gives the direction of the steepest ascent. The size of the step we take on each iteration to reach the local minimum is determined by the learning rate α. Therefore, we follow the direction of the slope downhill until we reach a local minimum.

ANS.13-The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.

The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall.

The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing.

It is possible to adjust the F-score to give more importance to precision over recall, or vice-versa. Common adjusted F-scores are the F0.5-score and the F2-score, as well as the standard F1-score.

ANS.14-The fit() method helps in fitting the data into a model, transform() method helps in transforming the data into a form that is more suitable for the model. Fit\_transform() method, on the other hand, combines the functionalities of both fit() and transform() methods in one step.

STATISTICS WORKSHEET-7

ANS.1-(B)

ANS.2-(D)

ANS.3-(C)

ANS.4-(B)

ANS.5-(C)

ANS.6-(A)

ANS.7-(A)

ANS.8-(A)

ANS.9-(A)

ANS.10-(A)

ANS.11-(A)

ANS.12-(A)

ANS.13-(D)

ANS.14-(D)

ANS.15-(D)

STATISTICS WORKSHEET-8

ANS.1-(B)

ANS.2-(B)

ANS.3-(D)

ANS.4-(B)

ANS.5-(B)

ANS.6-(D)

ANS.7-(B)

ANS.8-(A)

ANS.9-(A)

ANS.10-(C)

ANS.11-(A)

ANS.12-(A)

ANS.13-Analysis of variance (ANOVA) is used when comparing the mean scores of more than two groups. One-way analysis of variance involves one independent variable (referred to as factor) which has a number of different levels (groups or conditions). The dependent variable is a continuous variable.

Analysis of variance compares the variability in scores between the different groups and the variability within each group. An F ratio is calculated; a large F ratio indicates that there is more variability between the groups (cause by the independent variable) than there is within each group (error term). A significant F test indicates that the groups differ. However, it does not indicate which of the groups differ. For this, you will need to conduct post-hoc tests.

ANS.14-There are three primary assumptions in ANOVA:

The responses for each factor level have a normal population distribution.

These distributions have the same variance.

The data are independent.

A general rule of thumb for equal variances is to compare the smallest and largest sample standard deviations. This is much like the rule of thumb for equal variances for the test for independent means. If the ratio of these two sample standard deviations falls within 0.5 to 2, then it may be that the assumption is not violated.

ANS.15-The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two.

One-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.

Two-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master’s), and race finishing times in a marathon.

MACHINE LEARNING-8

ANS.1-(B)

ANS.3-(C)

ANS.5-(D)

ANS.6-(C)

ANS.7-(C)

ANS.8-(D)

ANS.9-(B)

ANS.10-(A)

ANS.11- One-hot coding is a widely used technique for converting categorical data into numerical data. However, there are some situations where one-hot encoding may not be the best choice. One such situation is when dealing with high carinality categorical variables, where the numbers of categories is very large.

In this case, one-hot encoding can lead to an explosion in the number of features, which can lead to overfitting, increased computation time, and reduced model performance. In such situations, alternative encoding, frequency or binary encoding may be more appropriate.

Another situation where one-hot encoding may not be suitable is when dealing with ordinal categorical variables where the categories have a natural order or hierarchy. In such cases, ordinal encoding, where the categories are assigned integer values based on their order, may be a better choice.

ANS 12- Data imabalance is a common problem in classification tasks where the number of instances in one class is significantly higher or lower than the number of instances in another class. This can be lead to a biased model that performs poorly on the minority class. There are several techniques that can be used to balance the dataset, some of which are:

1. Undersampling: this involves randomly instances from the majority class to match the number of instances in he moniority class. This can be lead to a more balanced dataset, but it may result in the loss of useful information from the majority calss.
2. Oversampling: This involves creating new instances for the miniority class to match the number of instaces in the majority class. This can be done by replacing exisiting instances or generating new instances using techniques such as SMOTE (synthetic miniority over-sampling technique).
3. Class weighting: This involves assigning higher weights to the miniority class and lower weights to the majority class during model traning. This can help the model to focus more on the minority class and improve its performance on that class.
4. Ensembles methods: This involves combining multiple models to improve the classification performance. On way to do this is by using techniques such as bagging or boosting, which can help to reduce the impact of class imbalance on the model’s performance.
5. Data augmentation: This involves creating new instances by applying transformations such as rotation ,scaling, or flipping to the existing instances. This can help to increase the number of instances in the minority class and improve the model’s performance on that class.

ANS.13- SMOTE(synthetic minority over sampling technique) and ADASYN (adaptive synthetic sampling are two popular techniques used for handling class imbalance in machine learning.

SMOTE generates new synthetic minority class samples by creating new data points along the line segments connecting minority class instance, whereas ADASYN generates synthetic samples by computing the density distribution of the minority class samples and generating new samples in regions of higher density.

The main difference between SMOTE and ADYSN lies in how they generate synthetic samples. While SMOTE generates synthetic samples in a fixed manner, ADASYN generates synthetic samples adaptively based on the local distribution of the minority class.

SMOTE can be less effective when the minority class is highly overlapped with the majority class, as it can generates noisy samples in those areas. ADASYN, on the other hand, can better handle such scenarios by generating more synthetic samples in areas where the class imbalance is more severe.

Overall, both techniques can be effective in handling class imbalance, but the choice between SMOTE AND ADASYN may depend on the specific characteristics of the data and the task at hand.

ANS 14- The purpose of using GridsearchCV is to tune the hyper parameters of a machine learning model in order to obtain the best performance on a given dataset.

GridsearchCv is a technique that exhaustively searches over a specified parameter grid to find the optimal combination of hyper parameters for a machine learning algorithm. It involves training and evaluating the model on each combination of hyper parameters in the grid and selecting the set of hyperparameters that yields the best performance.

GridsearchCV is particularly useful when dealing with models that have a large number of hyperparameters, as it can help to automate the process of finding the best combination of hyper parameters . it can also used to avoid overfitting or underfitting by selecting hyper parameters that generalized well to the test data.

In terms of using GridsearchCV with large database, it can become computationally expensive, as it involves training and evaluating a model on each combination of hyper parameter in the grid. However there are techniques that can be used to speed up the process, such as reducing the dimensionally of the features space, using parallel processing, or using randomized search instead of exhaustive search.

Overall, gridsearchCV is a powerful technique that can help to optimized machimne learning models, but its use should be carefully considered based on the specific charachetrstic of the dataset and the resources avaialabel.

in