Machine Learning

1. C
2. B
3. C
4. C
5. B
6. A, D
7. B
8. A
9. A,B
10. Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 has a propensity to overestimate how well the linear regression fits. It always rises as more factors are incorporated into the model. In order to account for this overestimation, adjusted R2 is used. If a single effect does not help the model, adjusted R2 may fall.
11. A variation of linear regression known as lasso, penalises the model for the total of the absolute weight values. As a result, the absolute values of weight will generally be lower and likely to be zero. The coefficient used by Lasso's hyperparameter, alpha, to punish weights. Ridge goes one step far and penalises the model based on the weights' sum squared value. As a result, the weights not only have smaller absolute values but also have a tendency to penalise the weights' extremes, leading to a collection of weights that are more equally distributed.
12. One indicator of the degree of multicollinearity in regression analysis is the variance inflation factor (VIF). In a multivariate regression model, multicollinearity occurs when there is a correlation between several independent variables. The regression results may suffer as a result. The variance inflation factor can therefore be used to calculate the degree to which multicollinearity has inflated the variance of a regression coefficient.
    - VIF equal to 1 = variables are not correlated
    - VIF between 1 and 5 = variables are moderately correlated
    - VIF greater than 5 = variables are highly correlated

    The likelihood of multicollinearity increases with VIF, necessitating more investigation. There is severe multicollinearity that needs to be adjusted when VIF is higher than 10.
13. The discrepancy in feature ranges will result in different step sizes for each feature if an algorithm uses gradient descent. We scale the data before feeding it to the model to make sure that the gradient descent progresses smoothly towards the minima and that the steps for gradient descent are updated at the same pace for all the features. Similar-scale characteristics will accelerate the gradient descent's convergence to the minima.

14. The popular metrics for regression models are:
    - Mean Squared Error-MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.
    - Root Mean Squared Error-An addition to the mean squared error is the Root Mean Squared Error, or RMSE. The square root of the error is determined, which is significant since it means that the units of the RMSE and the goal value that is being forecasted are the same.
    - Mean Absolute Error-The average of the absolute error numbers is used to generate the MAE score. The mathematical function abs() only turns a negative value positive. As a result, while calculating the MAE, the difference between an expected and forecasted number may be positive or negative.
15. Sensitivity=TP/(TP+FN)=0.8
    Specificity= TN/(TN+FP)=0.96
    Precision=TP /(TP + FP)=0.95
    Recall=Sensitivity=TP/(TP+FN)=0.8
    Accuracy= TP + TN / TP + TN + FP + FN=0.88

SQL

1. A,C,D
2. A,C,D
3. B
4. C
5. B
6. B
7. A
8. D
9. D
10. A
11. With a database optimization approach known as denormalization, we add duplicated data to one or more tables. This can help relational databases avoid pricey joins. Denormalization does not imply "reversing normalisation" or "not to normalise," as some may believe. It is an optimization method used following normalisation. Denormalization, in its simplest form, is the process of taking a normalised schema and rendering it non-normalized. Designers use it to optimise the efficiency of systems to handle time-critical processes.
12. One way to conceptualise a database cursor is as a pointer to a specific row within a query result. From one row to the next, the pointer can be shifted. It is possible to move the cursor to the previous row depending on its type.
13. There are 5 types of widely used SQL queries:
    - Data Definition Language-We can define the database structure or schema with the aid of data definition language. Eg: CREATE, DROP, ALTER, TRUNCATE

- Data Manipulation Language-By adding, changing, and deleting data, the Data Manipulation Language enables you to alter the database instance. It is in charge of carrying out all different kinds of data alteration in a database. Eg: INSERT, UPDATE, DELETE.
- Data Control Language-To grant "rights & permissions," the DCL (Data Control Language) includes commands like GRANT and REVOKE.
- Transaction Control Language- The transactions within the database are handled via TCL commands, or transaction control language. Eg: ROLLBACK, SAVEPOINT.
- Data Query Language-The database's data is retrieved using Data Query Language (DQL). "SELECT" is the only command that is used.

14. Rules for the data in a table can be specified using SQL constraints. The kinds of data that can be entered into a table are restricted by constraints. This guarantees the reliability and accuracy of the data in the table. The action is stopped if there is a violation between the constraint and the data action. Column-level or table-level constraints are both possible. Table level restrictions apply to the entire table, while column level constraints just affect the specified column.
15. When a new record is entered into a table, auto-increment enables a unique number to be created automatically. This is frequently the primary key field that we want to be automatically produced each time a new record is inserted.

Statistics

1. D
2. A
3. A
4. C
5. C
6. A
7. C
8. B
9. B
10. The five most crucial descriptive values for a data collection are graphically represented in a box plot, also known as a box-and-whisker plot. The minimum value, the first quartile, the median, the third quartile, and the highest value are among these values. Just the horizontal axis on the graph of this five-number summary shows values.

    A histogram is a style of bar chart in which the frequencies of a collection of data are graphically represented. A histogram depicts the frequency, or raw count, on the Y-axis and the variable being measured on the X-axis, much like a bar chart does.

    Even while histograms and box plots are both considered chart aids, they are actually completely different kinds of charts. Both charts effectively depict various sets of data; but, in some cases, one chart may be more effective than the other in recognising data

variations. The kind of data gathered, a preliminary study of data trends, and the project objectives all influence the type of chart aid that is used.

11.
    - When a sample is being compared to the population based on a hypothesis, one-sample tests are applicable. The theory or calculations used to determine the population characteristics.
    - Two-sample tests can be used to compare two samples, usually experimental and control samples from a carefully planned experiment.
    - For comparing two samples where it is impossible to control crucial variables, paired tests should be used. Members are paired between samples rather than being compared to two sets, making the difference between the members the sample. The mean of the discrepancies is then often contrasted with zero.
    - With strict requirements for normality and a known standard deviation, Z-tests are appropriate for comparing means.
    - Under flexible conditions, a t-test is adequate for comparing means (less is assumed).
    - F-tests (analysis of variance, ANOVA) are commonly used when deciding whether groupings of data by category are meaningful. If the variance of test scores of the left-handed in a class is much smaller than the variance of the whole class, then it may be useful to study lefties as a group. The null hypothesis is that two variances are the same – so the proposed grouping is not meaningful.
    - To ascertain whether a normal population has a given variance, chi-squared tests are utilised. The assumption at zero is that it does.

12. The use of hypothesis testing is necessary to evaluate statistical significance. The alternative and null hypotheses would be presented first. Second, the p-value, or probability of receiving the test's observed results if the null hypothesis is true, will be computed. Lastly, the null hypothesis will be rejected using the threshold of significance (alpha). The result is statistically significant if the p-value is less than the alpha.
13. There is no Gaussian or log-normal distribution for exponential distributions. In actuality, categorical data of any kind won't have these distributions either. Time until the next earthquake, for instance, or the length of a phone call are examples of such data.
14. If we were to establish the average wage for everyone in the United States based solely on their salaries, the mean would be greatly skewed due to the huge concentration of high-earning individuals. On the other hand, the median provides the 50-percentile data that, in this instance, best explains the central trend. As a result, the median in this situation provides a better knowledge of an average person's wage than the mean, which tends to move closer to the salaries of the wealthier people and is often not a good predictor.

15. Simply said, the term "likelihood" refers to the process of modifying the dataset distribution's features in order to improve the likelihood that a specific circumstance will occur.