G.K. GUJAR MEMORIAL CHARITABLE TRUST'S

DR. ASHOK GUJAR TECHNICAL INSTITUTE'S
DR. DAULATRAO AHER COLLEGE OF ENGINEERING, KARAD

UNIT TEST EXAMINATION - I / II / III

Name of Student: Shubham Shivaji Thotat.

Class: B.Tech (Mech) Division: FST2.0 09

Roll No. Date: 14/02/2024

Subject: Signature of Supervisor:

Que. No.	1	2	3	4	5	6	7	8	9	10	Total Marks
Marks Obtained											57/100

Class Assesement - 3

(Sign. of Subject Teacher)

Section A

Q.1. What is the difference a list and a tupple in python? Provide an example of when you would use each?

→ List: 1) Lists are mutable: You can modify, add or remove element after creation.

2) List defined using square brackets e.g. list=[1,2,3]

Tuple: 1) Tuple are immutable: Once created, their element cannot be changed.

2) Tuple use parentheses

e.g. Tuple = (1,2,3)

Example use cases:

1) Use list when mutability matter,

e.g. Managing a dynamic collection of user input.

User_data = ['John, 25]

User_data = 26 ≠ ~~Immutable data~~ modify age.

2) Use typele for Immutable data,

Co-ordinates = (10,20).

Choose list when you need a codlection that can be modified and tuple when you wsant a Fixed unchangeble set of value.

Q.2 Write a python function to calculate the factorial of given number.

→
```
def Factorial(n):
    If n==0 or n==1:
        return 1
    else:
        return * factorial (n-1)
```

Example usage:

```
number 5
result = Factorial (number)
Print (f"The Factorial of {number} is : {result}")
```

Q.3 Explain the concept of list comprehension in python. Provide an example of how it can be used to create list.

→ List comprehension is a concise way to create lists a python. It provides lists by sperifying the element to include and the condition to meet. The basic structure is,

[expression for item in iterable if condition]

e.g.

Here an example that generates a list of squares for even number between 0 & 9 using list compression,

```
Squares = [x**2 for x in range(10) if x%2==0]
print Squares
```

The result a list [0, 4, 16, 36, 64]

Q.4. Briefly explain the purpose of the following python libraries: Numpy, Pandas & Matplotlib.

→ 1) Numpy:
- Numpy is a powerful library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices.
- Along with a collection of mathematical functions to operate on these arrays efficiency.

2) Pandas:
- Pandas is a data manipulation and analysis library that provided easy-to-use data structures, such as dataframes, for working with structured data.
- It simplifies task like cleaning filtering & aggregating data
- Widly used in data analysis & preposition, handling missing data, and performing operations on structured dataset.

3) Matplotlib:
- Matplotlib is a 2D plotting library for creating static animated & interactive visualization in python. It allows user to create a coide variety of charts and graphs to visualize data.
- Ideal for data visualization in scientific computing, statistical analysis.

Section B

Q 1. Define Supervised learning & unsupervised learning provide an example of each.

→ Supervised learning:

- Supervised learning is a type of machine learning where the algorithm is trained on a labled dataset. In this approach the model learns to map input data to corresponding output labels.

Example:

- Task - Image classification [distiguishing bet cats & dogs]
- Dataset - Collection of images with label is
- Training - The algorithm learns the patterns and features associated with each class from image.
- Testing - Once trained the model can predict the labels of new images.

Unsupervised learning:

- Unsupervised learning involves training a machine algorithm on an unlabeled dataset.

Example:

- Task: Clustering customer performances in an e-commerce platform.
- Dataset: Purchasing history data without label categories.
- Algorithm: The model identifies natural grouping or clusters based on similarities in purchasing behaviour.
- Output: Clusters or segments of customers with similar performances emerge from analysis.

Q 3. Describe the steps involved in the machine learning pipeline provide a brief explanation of each step.

→ Steps:

1) Data Collection:

Gather relevent data for the problem at hand the quality and quantity of data significantly impact model performance.

2) Data cleaning & procesing:

Clean and preprocess the data to handle missing values outliers, and in consistancies convert data into a suitable format for training the model.

3) Feature Engineering:

Create new features or transform existing ones to enhance the models ability to capture patterns in the data this step improve model performance.

4) Data Spliting:

Divide the dataset into training & testing sets.

5) Model selection:

Choose a machinelearning algorithm & model. The selection cleped upon nature & characteristics of data.

6) Model training:

Train the selected model using the training data.

7) Hyperparameter Tuning:

Fine-Tune the hyperparameters of the model to optimize its performance.

8) Model evaluation:

Assess the model performance

9) Model deployement:

Integrate model into production environment.

10) Monitoring & maintenance:

Q.4. What is cross-validation and why is it important in machine learning? Provide an example of a cross validation.

→ Cross-Validation:-

- Cross-validation is a resampling technique used in a machine learning to access the performance of a model and to reduce the risk of over fitting or underfitting.
- It involves partitioning the dataset into subsets training the model on some of these subsets & evaluating it on the remaing subsets.

Importance of cross validation:

1) Better performance estimation: Cross-validation provides a more reliable estimate of how well a model will generalize to unseen data compared to a single train-test split.
2) Reduced overfitting or underfitting risk: By evaluating the model on multiple subsets, cross validation helps in identifying if model is overfitting & underfitting.

example:

```
from sklearn.model_selection import kfold
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import RandomForestClassifier.
```

≠ Example using Kfold cross-validation with random forest classifier.

```
model = RandomForestClassifier( )
KF = kfold (n_splits = 5, shuffle = True,
            random_state = 42)
```

≠ Performance cross-validation and calculate accuracy

```
accuracy_scores = cross_val_score(model, x, y, cv = kf,
                                  scoring = 'accuracy')
```

≠ Display average accuracy

```
print("Average accuracy", accuracy_scores.mean())
```

Q.5 Differentiate between regression and classification in the context of machine learning provide an example for each.

→

Regression	Classification
1) The output variable has to be real value or continues in nature.	1) The data output variable in discrete in the classification SML problem.
2) Regression algorithm helps to map the input value and the continous output variable.	2) Classification algorithm helps in mapping the input value with the output variable which is discrete in nature.
3) Regression algorithms are only used for data that is continous.	3) Classification algorithm are only used for data is deveate.
4) Linear regression, decision trees & natural network are common used for tasks. examples: Predicting house prices based on features like square footage, number of bedrooms & location.	4) Logistic regression, decision trees, support vector machine & natural network common task. example: Identifying whether an email is spam or not based on its content and characteristics.

Q.6. Briefly explain the K-nearest neighbors (KNN) algorithm. How does it work, and what are its main parameters?

→ The K-nearest neighbors (KNN) algorithm a simple and versatile supervised machine learning algorithm used for both classification & regression tasks. It make prediction based on the majority class or average of the K-nearest data points in the future space.

How KNN works:-

1) Training:

The algorithm stores the entire training dataset in memory.

2) Prediction (Classification):

For a new data point, the algorithm identified the K-nearest neighbors from the training set based on a distance metric.

- The majority class among these neighbors is assigned to the data point.

3) Prediction (Regression):

For regression tasks, the algorithm calculates the average of the target values of the K-nearest neighbors and assigns it to the new data point.

Main parameter:

1) Number of Neighbors (K):

2) Distance metric:

Measure similarity or dissimilarity betn data point

3) Weighting of neighbors:

Defines the contribution of each neighbor to the prediction.

Page No.
Date

Section c:

Q.1 Define the terms mean, median and mode explain when each measure of central tendency is most appropriate.

→ Mean :-

The mean is arithmetic average of a set of values. It is calculated by summing all values and dividing by the total number of values.

Median:

The median is the middle value in a sorted list of numbers. If there's an even number of values the median is the average of the two middle values.

Mode:

The mode is the ~~middle~~ value ~~in a sorted list of numbers. If there's an even number of values the median is the average of the two middle values~~ that appears most frequently in a dataset.

Appropriatness:

1) Use mean when:
- Dealing with a symmetric distribution.
- Outliers are not present or are minimal.
- The goal is to capture the average or typical value.

2) Use median when:
- Dealing with skewed distribution.
- Outliners are present and may significantly impact the mean.
- Wanting a measure less affected by extreme values.

3) Use mode when:
- Dealing with categorical data.
- Identifying the most common category is important.

- There may be multiple modes in a dataset.

Q.2. A dataset has a standard deviation of 10. If a data point is 2 standard deviations above the mean. What percentage of the data is below this point in a normal distribution?

→ In a normal distribution, approximately 68% of the data falls within one standard deviation of the mean, 95% falls within two standard deviation and about 99.7% falls within three standard deviation.

If a data point is 2 standard deviations above the mean, we are looking at the top 25% of the distribution [because 95% is within two std deviation. Leaving 5% in the tails. and we're interested in one tail.

So, the percentage of data below this point is
100% - 2.5% = 97.5%

Therfore, approximetly 97.5% of the data is below a data point that is 2 standard deviations above the mean in a normal distribution.

Q.4. Describe the difference betn correlation & causation. Provide an example to illustrate your explanation.

→ Correlation:

Defination :- Correlation measures the statistical relationship between two variables. If two variables are correlated it means that a change in one variable is associated with a change in the other but it doesn't imply a cause & effect relationship.

Example : Ice creame sales and the number of drowings are positively correlated. However, it does

mean that buying more icecream causes more drownings or vice versa.

2) Causation:

Definition: Causation implies a cause and effect relationship bet^n two variable change in one variable directly causes a change in order.

Example: The statement "smoking causes lung cancer" asserts a casual relationship bet^n smoking and the development of lung cancer.

Q.5 A random variable x follows a normal distribution with a mean of 50 and a standard deviation of 8. Calculate the z-score for a value of $X = 58$.

→ The z-score (or standard score) for a particular value in a normal distribution is a measure of how many standard deviation that value is from the mean,

$$z = \frac{x - \mu}{\sigma}$$

In this case,

$X = 58, \mu = 50, \sigma = 8$

Now plug these values into formula,

$$z = \frac{58 - 50}{8}$$

$$= \frac{8}{8}$$

$$z = 1$$

So, the z-score for $X = 58$ in this normal distribution is 1.

Page No.
Date

Section D

Q1. Explain the concept of overfitting in machine learning. How can it be mitigated?

→ Overfitting occurs when a machine learning model learns the training data too well, capturing noise and fluctuations in the data rather than the underlying patterns. This can lead ~~on~~ to poor generalization performance on new unseen data. Overfitting is characterized by a model that performs well on the training set by poorly on test & validation data.

Mitigation strategies for overfitting:

1) Cross validation: Use techniques like cross-validation to assess the model's performance on different subsets of the data.

2) Train with more data:
Increasing the size of the training dataset can help the model learn more robust patterns & reduce the chance of memorizing noise.

3) Simplify the model:
choose a simpler model with fewer parameters this helps prevent the model from being too flexible & capturing noise.

4) Regularization:
Regularization adds a penalty term to the loss function, discouraging extreme parameter values

5) Feature selection:
Carefully choose relevant features and eliminate unnecessary ones.

6) Ensemble methods:

Random forests or gardient boosting these methods combine predictions from multiple models often reducing overfitting.

7) Dropout.

8) Early stopping:

Monitor the models performance on a validation set during training.

Q.2. Briefly describe the support vector machine (SVM) algorithm. What is the role of the kernel in SVM.

→ Support vector machine (SVM)

Defination: SVM is a superviised machine learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates different classes in the feature space.

Role of the kernal SVM:

- Linear SVM:
  - In this smimplest form, SVM uses a linear kernel and the decision boundary is a straight line.
  - The linear kernel is effective when the data is already separable in the feature space.