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: Aavej Vahid Patel

Class: Batch No - D9 Division:

Roll No. Date: 14/2/24

Subject: Assesment 3 Signature of Supervisor:

Que. No.	1	2	3	4	5	6	7	8	9	10	Total Marks
Marks Obtained											53/100

(Sign. of Subject Teacher)

Sec. A.

Q.1]

→ List :- ① lists are mutable; you can modify add or remove element after creation.

② List defined using square brackets

eg. list = [1,2,3]

Tuple :- ① Tuple are immutable; once created, their element cannot be changed

② Tuple use parentheses

eg. Tuple = (1,2,3)

* Example use cases.

1) Use List when mutability matter

eg- managing a dynamic collection of user input

ud = ['John', 25]

ud = 26 ≠ modify age

2] Use- tuple for Immutable data

coordinates = (10,20)

Choose list when you need a collection that can be modified & tuple when you want a fixed unchangeable set of value.

Q.2. → ~~write a pgm~~

```
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: {res
```

Q.3]

→ List comprehension is a concise way to create lists: a python. It provides a more readab & compact syntax for generating lists by specify ng the element to include & the condition to meet. The basic structure is

[expression for item in trable if condition

eg :-

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

```
Squares = [x** 2 for x in range (10) if x%, L=
print squares
```

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

3.4]

→ 1] Numpy :-

- Numpy is a powerful library for numerical computing in python. It provides support for large, multi-densional arrays & matrices, along with a collection of mathematical functions to operate on these array efficiency

2] Pandas :

- Pandas is a data manipulation & analysis library that provides 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 & preparation, handling missing data, & performing operations on structured.

3] Matplotlib :-

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

Sec 'B'

Q.1]

→ Supervised learning :-

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

Example :-

- Task - Image classification [distiguishing Be cats & dogs]
- Dataset - Collection of images with labels
- Training - The algorithm leans the patterr & features associated with each class
- Testing - Once trained the model can predict the labels of new images.

3/ Unsupervised learning :

- Unsupervised learning involves training a mach algorithm on an unlabeled dataset

Example :

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

Q.3]

→ Steps:

① Data collection:

Gather relevant data for the problem at hand the quality & quantity of data significantly impact model performance.

② Data cleaning & processing:-

Clean & preprocess the data to handle missing values outliers & consistencies. Convert data into a suitable format for training the model

③ feature Engineering:-

Create new features or transform existing ones to enhance the model's ability to capture patterns inthe data. this step helps improve model performance

④ Data spliting:-

Divide the dataset into training & testing sets.

⑤ Model selection:

Choose a machine learning & model The selection depend upon nature & characteristic of dat.

⑥ Model training-

Train the selected model using the training data.

⑦ Model evaluation:

Assess the model performance.

⑧ Model deployment:

Integrate model into production enviroment.

⑨ Monitoring & maintenance.

Q.4]

→ Cross - validation :-

- cross - validation is resampling technique us in a machine learing to assess the performan of a model & to reduce the risk of overfitting or underfitting
- It involves partitioning the dataset into subsets training the model on these subsets & evaluating it on remaining subsets

Importance of cross validation

1] Better performance Estimation :- Cross - Valid provides a more reliable estimate of how well a model will generlize to unseen data compared to a single train - test split

2] Reduced overfitting or underfitting Risk :- By evaluating the model on multiple subse cross validation helps in indentifying if model is overfitting & underfitting

example :

```
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_sco
from sklearn.ensemble import RandomForestClassifie
```

Example using KFold cross - validation with Random forest classifier

```
model = RandomForestClassifier()
kf = KFold(n_splits = 5, shuffle = True,
           random_state = 42)
```

Performance cross-validation & calculate accuracy

```
accurancy_scores = cross_val_score(model, x, y,
                   cv = kf, scoring = 'accurancy')
```

Display average accurancy

```
print("Average Accurancy", accurancy_scores
                   mean())
```

Q.5]

→

Regression	Classification
1] The output variable has to be real value or continuous in nature	1] The data output variable is discreate in the best function SML problem
2] Regression algorithm helps to map the input value & the continuous 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 continuous	3] Classification algorithm are only used for data is decrete.
4] linear regression, decision trees, neutral networks are common used for tasks eg. predicting house prices based on features like square footage number of bedroom & location.	4] Logistic regression decision trees, support vector machine & neutral commontasks eg. Identifying whether an email is spam or not based on its content & characteristics.

Q.6] →

The k-nearest neighbors (KNN) algorithm a simple & 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] Trainning -
The algorithm stores the entire training dataset in memory.

2] Prediction (Classification) :-
- for a new data point, the algorithm identifies the K-nearest neighbors from the training set based on a distance metric.

3] Prediction :- (Regression)
for regression tasks the algorithm calculates the average of the target values of the K-nearest neighbors & assigns it to the new data point

Main parameter
① Number of Neighbors (K) :
② Distance Metric :-
Measure similarity or dissimilarity bet^n data points.
③ Weighting of Neighbors :-
Defines the contribution of each neighbor to the prediction

Date : / /
Page No.:

Section 'C'

Q.1 → 1] Mean :- It is the average of a set of values calculated by summing at values & dividing by total number of value

$$\text{Mean} : \quad \Sigma i^n = \frac{1}{n} x_i$$

2] Median :- It is a middle value of sorted list of values. It the Num if the middle value
- The median is most appropriate whe the data is skewed or contains outliers. It provides a better representation of the control tendancy in such cases.

3] Mode :- It is the value that appears most frequency in dataset.
The mode is most appropiate for categorial or nominal data, where calculating the mean & median may not be meaningful. It is also useful for identifying the most common value in dataset.

Q.2, 1) In a normal distribution, approximately 95% of data falls within 2 standardention of the mean. This is known as 95% rule or the empirical rule.

2) If a data point is 2 standard variation above the mean it falls within the top 2.5% of distribution (since approximately 95% of data is with in 2 standard deviation of the mean. 5% of the data outside

Date : / /
Page No.:

this range divided equally beth the upper & lower
iii) The percentage of the data below t
point in normal distribution is approxim
100% - 2.5% = 97.5%

Q.4]
→ The P value in hypothesis testing represents th
probability of observing a test static such
as extreme as or more than extreme, th
one observed in sample data, under the assumpti
that the null hypothesis is trap.

i) Formulate hypothesis :- In hypothesis testing we
start with a null hypothesis (Ho), which repres
the default assumption or the status ; &
alternative hypothesis (H1 or Ha) which represents
what we are trying to find evidance for
ii) Select a significance level :- Before conducting
the hypothesis for asignificance level (α)
is chosen, typically set at 0.05 or 0.01. This
represents threshold for rejecting null
hypothesis

Q.5.
→ $z = \frac{x - u}{\sigma}$ - x is value of random variable
- u is mean of the distribution &
- σ is the standard deviation of the distributio

$z = \frac{58 - 50}{8} = \frac{8}{8} = 1$

So the Z-score for the value of x = 58 is
z = 1

number do the predict

Date : / /
Page No.:

Section 'D'

Q1] → Overfitting is a common problem in machine learning where a model learns the learning data too well capturing noise & random function in the dat instead of underlying pattern. As a result the model performs well on the training data but fails to generalize to new unseen data.

Causes :- i) Model complexity ii) Insufficient data iii) Noise in the Data.

Effect :- i) Power generalization ii) High variance

mitigation Techniques for overfitting:-

i) Simplifying the model ii) cross validation iii) Regularization iv) feature selection v) Data Augmentation vi) Early stopping vii) Ensemble method

Q2] →

Support vector machine (SVM)

Definatn :- SVM is a supervised machine learning algorithm used for classification & regression tasks. It works by finding the hyperplane that best separate different classes in the feature space

Role of the kernel in SVM

Date : / /
Page No.:

• Linear SVM :-

- In its simplest form, SVM uses a linear kernal & the decision boundary is a straight line
- The linear kernal is effective when the data is already linearly separable in the feature space.

• Non-linear SVM :-

- For non-linearly separable data SVM can be non-linear kernals to map the input feature into higher dimensional space.
- This transformation allows for finding a hyperplane in the transformed space that corresponds to a non-linear decision boundary in original space