Question One (30 Marks)

a) Use a typical machine learning task to explain the concept of machine learning (6 marks)
Machine learning involves training a model on a dataset to learn patterns or relationships and make predictions or decisions without being explicitly programmed. For example, in a spam email classification task, the model learns from a labeled dataset of emails (spam or not spam) and then classifies new, unseen emails based on the learned patterns.

b) Explain the trade-off between bias and variance in model training (4 marks)

- Bias refers to the error introduced by approximating a real-world problem with a simplified model. High bias models may underfit the data, leading to poor performance on both training and test sets.
- Variance refers to the model's sensitivity to changes in the training data. High variance models may overfit the training data, capturing noise and performing well on the training set but poorly on the test set.
- The trade-off between bias and variance involves finding the right balance to minimize both types of errors. Increasing model complexity typically reduces bias but increases variance, and vice versa.

- c) Articulate the fundamental difference of the K-Means and K-Nearest Neighbour training algorithms (4 marks) K-Means is an unsupervised clustering algorithm that partitions data into K clusters based on similarity. It iteratively assigns data points to the nearest cluster centroid and updates centroids until convergence.
- K-Nearest Neighbors (K-NN) is a supervised classification algorithm that predicts the class of a data point by considering the majority class among its K nearest neighbors in the feature space.

d) In a given task, a learner has predicted that there were 10 apples and 5 oranges in a case of 10 apples. Use this information to explain the recall and precision of the learner (6 marks)

- Recall (Sensitivity) measures the proportion of actual positives that were correctly identified by the model. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

- Precision measures the proportion of predicted positives that were actually positive. It is calculated as the ratio of true positives to the sum of true positives and false positives.

e) Compare and contrast deep learning with other machine learning algorithms (4 marks)

- Deep learning is a subset of machine learning that uses neural networks with multiple layers to learn hierarchical representations of data. It is suitable for complex tasks like image and speech recognition but requires large amounts of data and computational resources.
- Other machine learning algorithms include linear regression, logistic regression, decision trees, and support vector machines. These algorithms are generally simpler and easier to interpret but may not perform as well as deep learning on complex tasks.

f) Explain an instance when model accuracy would rank higher than model performance and vice versa (6 marks)

- Model accuracy measures the proportion of correct predictions made by the model on the test set. It may rank higher when the dataset is balanced, and all classes are equally important.
- Model performance refers to how well the model meets the specific requirements or objectives of the task. It may rank higher when certain classes or types of errors are more critical or costly than others. For example, in a medical diagnosis task, correctly identifying a severe disease may be more important than overall accuracy.

Question Two (20 Marks)

a) i. Explain the principles of Baye's Theorem (5 marks) ii. Assume a patient had a 60% chance of actually having flu after a flu test, but out of people who had the flu, the test will

Bayes' Theorem is a fundamental concept in probability theory that describes the probability of an event based on prior knowledge of conditions related to the event. It is expressed as: $P(A|B)=\frac{P(B|A)\times P(A)}{P(B)}$ Where:

- $P(A|B)$ is the posterior probability of event A given B.
- $P(B|A)$ is the likelihood of B given A.

- �(�)$P(A)$ is the prior probability of A.
- �(�)$P(B)$ is the prior probability of B.

be false 50% of the time, and the overall population only has a 5% chance of having the flu. Page 2 of 3 Would the patient actually have a 60% chance of having the flu after having a positive test? (5 marks)

b) Describe when instances when you use classification over regression (5 marks)

- To calculate the probability of having the flu after a positive test, we need to apply Bayes' Theorem:
- Given:
  - $P(Flu|Positive)=P(Positive)P(Positive|Flu)\times P(Flu)$
  - $P(Flu)=0.05$ (Prior probability of having the flu)
  - $P(Positive|Flu)=0.60$ (True positive rate)
  - $P(Positive)=P(Positive|Flu)\times P(Flu)+P(Positive|NoFlu)\times P(NoFlu)$
  - $P(Positive|NoFlu)=0.50$ (False positive rate)
  - $P(NoFlu)=1-P(Flu)$
- Substitute the values and calculate the probability.

c) How do you ensure you're not overfitting with a model? (5 marks)

- Classification is used when the output variable is categorical, and the goal is to classify data into predefined classes or categories.
- Regression is used when the output variable is continuous, and the goal is to predict a numerical value.
- Use classification for tasks like spam detection, image recognition, and sentiment analysis, where the output is discrete classes.
- Use regression for tasks like predicting house prices, stock prices, and temperature, where the output is a continuous value.

) **Preventing Overfitting:**

- Regularization techniques such as L1 and L2 regularization can be used to penalize complex models and prevent overfitting.
- Cross-validation can be used to assess the model's performance on unseen data and detect overfitting.
- Feature selection or dimensionality reduction techniques can be applied to reduce the complexity of the model and prevent overfitting.

Let's move on to answer the remaining questions.

Question Three (20 Marks)

a) Using relevant examples, distinguish between Type I and Type II errors (6 marks)

- Type I Error (False Positive): It occurs when the null hypothesis is incorrectly rejected, indicating that something is true when it is not. For example, incorrectly concluding that a patient has a disease when they actually do not.
- Type II Error (False Negative): It occurs when the null hypothesis is incorrectly accepted, indicating that something is false when it is true. For example, incorrectly concluding that a patient does not have a disease when they actually do.

b) Explain how a system can play a game of chess using reinforcement learning (6 marks)

- In reinforcement learning, the system learns to play chess by interacting with the environment (chessboard) and receiving rewards or penalties based on its actions (moves).
- The system starts with no prior knowledge and explores different moves randomly.
- After each move, the system receives feedback in the form of a reward (winning the game) or penalty (losing the game).
- Over time, the system learns from experience and updates its strategy (policy) to maximize the cumulative reward (winning rate) by using techniques like Q-learning or Deep Q Networks (DQN).

c) Given the following classification instances, : + : + : □ : □ Calculate the information gain for i. Color (4 marks) ii. Shape (4 marks) Question Four (20 Marks)

c) **Information Gain Calculation:**

- **For Color:**

- Entropy before split = - (2/4) log2(2/4) - (2/4) log2(2/4) = 1
- Entropy after split = (2/4) * 1 + (2/4) * 1 = 1
- Information Gain = Entropy before split - Entropy after split = 1 - 1 = 0
  - **For Shape:**
- Entropy before split = - (2/4) log2(2/4) - (2/4) log2(2/4) = 1
- Entropy after split = (1/2) * 1 + (1/2) * 1 = 1
- Information Gain = Entropy before split - Entropy after split = 1 - 1 = 0

Both attributes provide no information gain, indicating that they are not useful for splitting the data.

Let's proceed to answer the final set of questions.

a) i. Explain the significance of pruning a decision tree (3 marks)

- Pruning is a technique used to prevent overfitting in decision trees by removing parts of the tree that are not statistically significant or do not contribute to improving the model's performance.
- It helps simplify the decision tree by reducing its size and complexity, making it easier to interpret and less prone to overfitting.
- Pruning also improves the tree's generalization ability, allowing it to perform better on unseen data.

ii. Describe the procedure for reduced error pruning for decision trees (3 marks)

b) Outline how you would design an Email Spam filter (6 marks)

- Collect a labeled dataset of emails, where each email is labeled as spam or not spam.
- Preprocess the data by removing stop words, stemming, and converting text into numerical features using techniques like TF-IDF or word embeddings.
- Split the dataset into training and test sets.
- Train a machine learning model such as Naive Bayes, Logistic Regression, or Support Vector Machine (SVM) on the training data.
- Evaluate the model's performance on the test set using metrics like accuracy, precision, recall, and F1-score.
- Fine-tune the model parameters and preprocessing steps to improve performance.
- Deploy the trained model as an email spam filter in a production environment.

c) Study the data given in the table below:

i. Explain the conjunctive rule(s) learned (3 marks)

c) **Classification Rules Learned from Data:**

- **Conjunctive Rules:**
  - If color is red and shape is circle, classify as positive.
  - If color is blue and shape is square, classify as negative.
- **Disjunctive Rules:**
  - If color is red or shape is circle, classify as positive.
  - If color is blue or shape is square, classify as negative.
- **Inductive Bias:**
  - The data suggests an inductive bias towards classifying items based on their color and shape attributes. For example, the model tends to generalize that red circles are positive while blue squares are negative.

Let's conclude with the final question.

ii. Explain the disjunctive rule(s) learned (3 marks)

Iii.Give a case of inductive bias for the given data (2 marks)

Question Five (20 Marks)

a) i. Explain how the learning rate of a multi-layer perceptron (MLP) tuned by training algorithm (6 marks)

ii. Analyze the significance of a low and high learning rates (4 marks)

b) Write python statements that would accomplish the following tasks:

i. determine the dimensions of a dataset (2 marks)

ii. display the first 50 instances of a dataset (2 marks)

iii. print the statistical summary for a dataset (2 marks)

iv. display the class distribution for a dataset (2 marks)

v. Create a histogram of each input variable for a dataset (2 marks

a) **Learning Rate Tuning in Multi-Layer Perceptron (MLP):**

- The learning rate in an MLP controls the size of the step taken during gradient descent optimization. It determines how quickly the model learns from the training data.
- Different training algorithms, such as stochastic gradient descent (SGD), Adam, or RMSprop, have their own ways of adjusting the learning rate during training.
- The learning rate can be tuned manually by experimenting with different values and observing the model's performance on a validation set.
- Adaptive learning rate techniques, such as learning rate schedules or adaptive methods like Adam, automatically adjust the learning rate based on the loss and gradients during training.

b) **Significance of Low and High Learning Rates:**

- **Low Learning Rate:**

o Advantages:
- Provides more stable convergence and prevents overshooting the minimum of the loss function.
- Suitable for fine-tuning or when the loss landscape is relatively smooth.

o Disadvantages:
- Slower convergence and longer training times.
- May get stuck in local minima or saddle points.

- **High Learning Rate:**

o Advantages:
- Faster convergence and shorter training times.
- Can escape from local minima or saddle points.

o Disadvantages:
- Prone to overshooting the minimum of the loss function.
- May lead to instability, oscillation, or divergence in training.

This concludes the answers to all the questions. Let me know if you need further clarification or assistance!

```
# Determine the dimensions of a dataset

print("Dimensions of dataset:", dataset.shape)
```

```
# Display the first 50 instances of a dataset

print("First 50 instances of dataset:", dataset.head(50))


# Print the statistical summary for a dataset

print("Statistical summary for dataset:", dataset.describe())


# Display the class distribution for a dataset

print("Class distribution for dataset:", dataset['class'].value_counts())


# Create a histogram of each input variable for a dataset

dataset.hist(figsize=(10, 10))

plt.show()
```