NAME-RAFIUR RAHMAN MASHRAFI

STUDENT ID: 2221971042

CSE445.4

Problem#1

1.Supervised Learning: when a computer algorithm trained on data that already labeled or categorized. After training phase it can apply what it is learned to classify or predict outcomes for fresh data.

2 .Unsupervised Learning: A type of machine learning uses algorithms to analyze unstructured data to find hidden patterns, structures, and relationships without human guidance.

3. Semi-Supervised Learning : A machine learning technique that uses a combination of both labeled and unlabeled data for training models.

4. Reinforcement Learning : Reinforcement learning is when a computer learns by trial and error. It tries different actions, gets rewards for good choices or penalties for bad ones, and figures out the best way to behave over time.

5. Feature / Attribute : A measurable property or column of the data used as input. A feature is like a main part of something, and then an attribute is more like a detail that describes that part.

6. Parameter:  a parameter is internal variable that the model learns from data to make predictions.

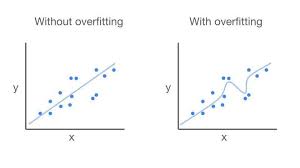
7. Training Data: Training data is the initial dataset used to teach an AI model what to do. It uses this information to learn the rules and patterns needed for its task.

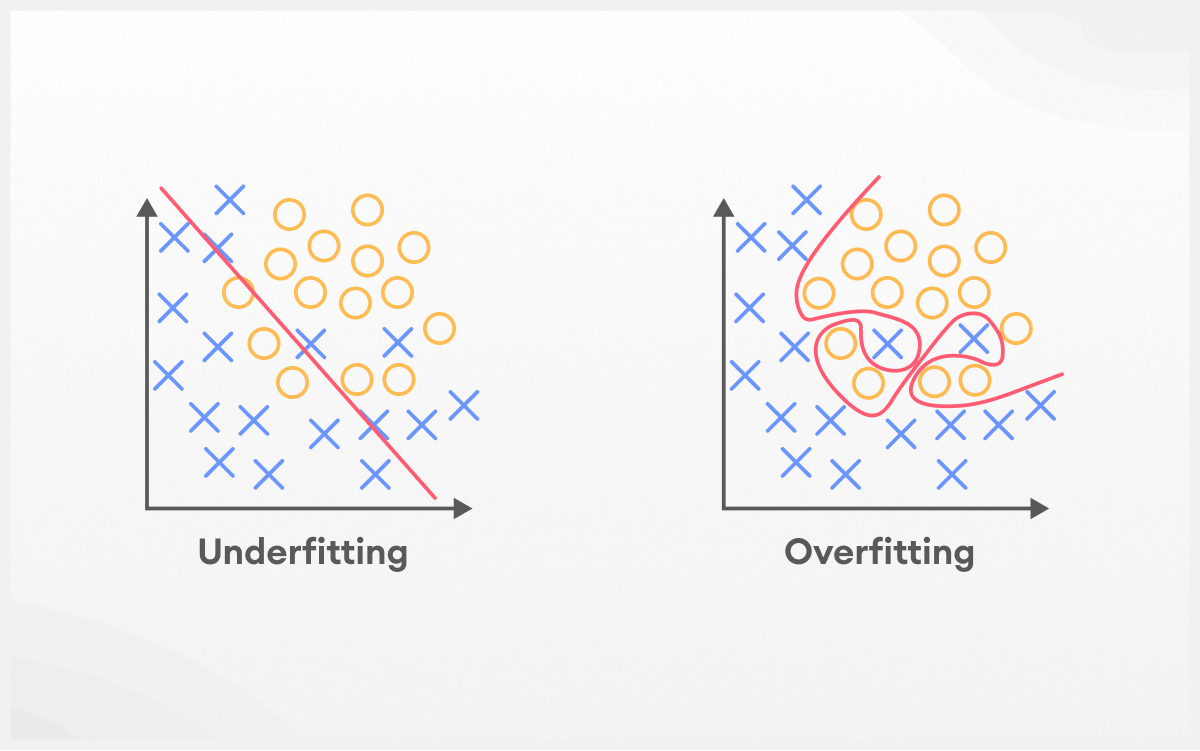
**8. Validation Set:** It’s a subset of data used when ml model training to tune hyperparameters and prevent overfitting. Its give a unbiased result of the model performance.

Train dataset---VALIDATION DATASET---Testing dataset

(hyperparameters and overfitting)

**9. Test Set:** The portion of dataset not use during the training or validation in a ML model.its main purpose is works efficiently in invisible dataset on the final evaluation of the models performance.

**10. Overfitting:**It is happened in ML when a model learns the training data erfectly,includes its noise,irrelevents details, that causes it to performed poorly unseen data.it happens for that the models is too complex ,specific to the training set.

**11. Underfitting : Its occurs in ML when model is so simple to gain the underlying patterns in the training data set.** Model is too simple and cannot capture underlying patterns; poor training and test performance.

**12. Bias–Variance Tradeoff:** Bias is the error that occurs when a model is too simple to capture the underlying patterns in the data. Variance is the error caused by a model's excessive sensitivity to small fluctuations or noise in the training data. It’s like finding good spot where it learns the real patterns without getting distracted by the noise. **Bias–Variance Tradeoff its like ML that ability to learn the training data and its ability to create to new data.**

**13. Loss Function:**This is result for a single training example. It tell the model how wrong its prediction was for just one piece of data.

**14. Cost Function:**This is the average result of the model training . It tells the model how wrong it is on average across all the model. This is the main score the model tries to minimize.

**15. Gradient Descent:**This is the model's way of finding the lowest point in a valley (the lowest error). It looks around, figures out which direction is downhill, and takes a small step that way. It repeats this until it finds the bottom.

**16. Learning Rate: This refers to the magnitude of the steps the model takes during optimization. If the step size is too large (high learning rate), you risk overshooting the minimum. Conversely, if the step size is too small (low learning rate), it will take a long time to converge.**

**17. Epoch: This term denotes a complete cycle through the entire training dataset. It’s akin to the model thoroughly reading a book from start to finish once. Typically, several epochs (repeated readings) are needed for effective learning.**

**18. Cross-validation: This is an advanced method of evaluation. Instead of just creating one train/test split, you divide your dataset into multiple segments. You train on some parts and validate on others, cycling through all sections to ensure consistent model performance rather than relying on chance.**

**19. Batch Size: This represents the quantity of training samples processed before the model is updated. A small batch size resembles studying using flashcards one at a time, while a large batch size is akin to reviewing an entire chapter prior to testing yourself.**

**20. Regularization: These are techniques employed to prevent the model from overfitting. It’s similar to encouraging the model to adopt a simpler rationale, which generally performs better on unseen data by imposing a penalty on complexity.**

**21. Linear Regression: This is a straightforward model that determines the best-fitting straight line through a dataset. It is used for predicting a continuous outcome, such as estimating house prices based on their size.**

**22. Logistic Regression: Logistic Regression is a supervised machine learning method mainly employed for classification problems, especially binary classification. Although it includes the term "regression" in its name, it does not function as a regression method for forecasting continuous values; instead, it is a classification technique that estimates the likelihood of a specific outcome.**

**23. Decision Tree: A diagrammatic framework utilized for decision-making or for a supervised machine learning technique that forecasts outcomes by learning basic decision rules derived from data.**

**24. Random Forest: This approach utilizes multiple decision trees. Instead of relying on a single tree, it consults a collection of trees and selects the outcome that receives the most votes. This enhances both robustness and accuracy compared to using a solitary tree alone.**

**25. Support Vector Machine (SVM):** A Support Vector Machine (SVM) is a form of supervised machine learning algorithm that identifies the optimal separation, known as a hyperplane, to differentiate data points belonging to various classes. While it is primarily used for classification purposes, it can also be utilized for regression and outlier detection. SVMs work by maximizing the distance, or margin, between the hyperplane and the closest data points from each class, which improves classification effectiveness.

**26. K-Nearest Neighbors (KNN)**  
K-Nearest Neighbors (KNN) is a supervised machine learning technique that does not assume any specific parameters and can be used for both classification and regression tasks, although its main application is in classification. This method works by categorizing a new data point into the class that is most prevalent among its 'k' nearest neighbors in the training set, with the neighbors determined through a chosen distance metric.

How KNN functions Choose 'k': Select a value, 'k', representing the number of neighbors to consider. Compute distances: Assess the distance from the new data point to each point in the training set. Identify nearest neighbors: Find the 'k' training data points that are nearest to the new data point based on the calculated distances.

Make predictions: For classification: Assign the new data point to the class that appears most frequently among its 'k' neighbors (a "majority vote").

For regression: Predict the value of the new data point by averaging the values of its 'k' neighbors.

**27. Naive Bayes**  
Naive bayes is part of supervised machine learning algorithms it is use for classification takes ,such as spam filter and sentiment analysis. Its work process : Training🡪prediction🡪Naïve assumption P(a|b)=(P(B|A)\*P(A))/P(B)

**28. Principal Component Analysis (PCA)**  
Principal Component Analysis (PCA) reduces a dataset's dimensionality by creating new, uncorrelated variables called principal components.These components are linear combinations of the original variables, ordered to capture the maximum variance in the data. The first component (PC1) captures the most variance, the second (PC2) captures the next most while being orthogonal to the first, and so on.By keeping only the first few components, you can simplify the data while retaining most of its information, which is useful for visualization and preprocessing.

**29. Clustering**  
The goal of unsupervised learning where you group similar data points together without using pre-defined labels. Finding natural "clumps" or "gangs" in data. The objective is to enhance similarity among items within the same cluster while reducing similarity across different clusters, which aids in uncovering concealed patterns and structures in unlabeled data.

**30. Dimensionality Reduction**  
The process of reducing the number of random variables (features) in model data. PCA is one way to do this. the method of decreasing the quantity of features (or variables) in a dataset while preserving as much significant information as feasible. This approach tackles the "curse of dimensionality," which occurs when an excessive number of features leads to slower training, overfitting, and diminished model performance. It streamlines intricate, high-dimensional data, making it more manageable to analyze, visualize, and process, thereby enhancing model accuracy and efficiency.

**31. Perceptron**  
The simplest possible type of neural network. A perceptron is a basic representation of an artificial neuron utilized for binary classification and serves as a key component in neural networks. It accepts several inputs, multiplies them by corresponding weights, aggregates the results, and then applies an activation function to generate a binary output (0 or 1). The learning process involves modifying these weights and a bias to enhance the classification of data.

**32. Activation Function**  
A function in a neural network that decides mathematical functions base on the input and inputs outputs determined deciding whether the neuron should be "activated" or not.

**33. Backpropagation**  
The "learning" algorithm for neural networks. It calculates how much each neuron contributed to the final error and then adjusts all the weights backwards through the network to do better next time.

Backpropagation is a method utilized to train neural networks by determining the error at the output and sending it back through the network to modify the weights and biases. It employs the chain rule from calculus to evaluate the gradient of the loss function, which shows how to adjust the parameters of the network to reduce errors and enhance accuracy. This process is repetitive, occurring over numerous training examples, until the predictions of the network reach an acceptable level of precision.

**34. Convolutional Neural Network (CNN)**  
A type of neural network super good at processing images. It uses "filters" to scan across an image and detect patterns like edges, shapes, and eventually complex objects.

IT is a deep learning model mainly employed for analyzing images and videos, identifying patterns through convolutional layers that handle input data.

**35. Recurrent Neural Network (RNN)**

**It’s a type of deep learning model make for processing sequential data, as text, speech or time-series data.it memory use information from previous inputs and influence the current input and output.** simple said neural network designed for sequential data (like time series or text). It has a "memory" of previous inputs, which helps it understand context.

**36. Long Short-Term Memory (LSTM)**  
Its a smarter and more complex type of RNN. LSTMs are capable of capturing long-term relationships within sequential data, making them suited for applications such as language translation, speech recognition, and time series prediction. In contrast to standard RNNs, which rely on a single hidden state carried through time, LSTMs incorporate a memory cell that retains information over longer durations, effectively tackling the issue of learning long-term dependencies.

**37. Transformer**  
A modern neural network architecture that relies entirely on an "Attention Mechanism" to process data. A Transformer is an architecture for deep learning that is particularly effective at managing sequential data, mainly utilized in Natural Language Processing (NLP) for activities such as translation and text generation. Its main breakthrough is the self-attention mechanism,

**38. Attention Mechanism**  
A technique that allows a model to focus on the most relevant parts of the input when making a prediction. Like when translating a sentence, it "pays attention" to the specific words in the original sentence that are most relevant to the word it's currently writing.

**39. Dropout**  
A regularization technique used during training where random neurons are temporarily ignored.

**40. Gradient Vanishing Problem**  
The vanishing gradient problem describes a difficulty encountered when training deep neural networks, where gradients become significantly smaller as they are propagated back to the earlier layers, which can effectively hinder or slow down learning in those layers. This issue arises because gradients are repeatedly multiplied by small values (typically the derivatives of activation functions) during the backpropagation process in deep networks, causing an exponential reduction in their magnitude. As a result, the weights in the initial layers receive minimal updates, making it challenging for the model to learn effectively from the data.

Causes:

Activation Functions: Functions like sigmoid and hyperbolic tangent (tanh) have gradients that fall between 0 and 1, leading to the diminishing of gradients during multiplication.

Network Depth: The increased number of multiplications in deeper networks amplifies the vanishing gradient effect in the initial layers.

Weight Initialization: Poor weight initialization can worsen this problem.

Consequences:

Slow or Halted Training: The model undergoes a significantly slow learning process or may entirely cease learning, especially in the initial layers. Poor Performance: The model might struggle to identify complex patterns and demonstrate inadequate performance.

**41. Transfer Learning**   
A machine learning technique where a model trained for one task is re-used as the starting point for a new, related task. This approach conserves time and resources by utilizing insights acquired from earlier training, proving particularly beneficial in scenarios where data for a new task is scarce. Rather than creating a model from the ground up, transfer learning modifies an existing pre-trained model for a different challenge, which may require only the final layers to be retrained.

**42. Fine-Tuning**  
The second step of transfer learning. After take a pre-trained model, further train , it on own smaller, specific dataset to make it an expert on your particular problem.

**43. Hyperparameter Optimization**  
The process of finding the best settings (hyperparameters) for your model before training even starts. Things like the learning rate or number of layers. It's like tuning a race car's engine for peak performance before the race.

**44. Specificity**  
A metric that measures how good your model is at correctly identifying the negative cases. It's the "True Negative Rate." For a disease test, it's the percentage of healthy people correctly identified as healthy.

**45. Confusion Matrix**

**This is a table that outlines the performance of your classification model. It categorizes predictions into four groups: True Positives, True Negatives, False Positives, and False Negatives. This table serves as the basis for calculating most other performance metrics.**

**46. Precision**

**This measures the proportion of instances predicted as positive that were actually positive. It evaluates the model's reliability when it predicts a positive outcome. High precision indicates that you can have confidence in its affirmative predictions.**

**47. F1 Score**

**This is a combined metric that incorporates both Precision and Recall into a single value. It represents the "harmonic mean" of the two and is valuable when seeking a balance between them, especially when dealing with imbalanced classes.**

**48. ROC Curve**

**This is a graph that illustrates the effectiveness of a classification model across various classification thresholds. It charts the True Positive Rate against the False Positive Rate.**

**49. AUC (Area Under the Curve)**

**This refers to the area beneath the ROC Curve. It provides a single value summarizing the model's capability to differentiate between classes. An AUC of 1 indicates a flawless model, while 0.5 suggests performance is equivalent to random guessing.**

**Problem#2**

* **EDA**: Helps understand data patterns and relationships
* **Train-Test Split**: Ensures unbiased model evaluation
* **Linear Regression**: Simple, interpretable model
* **Decision Tree**: Can capture non-linear relationships
* **Evaluation Metrics**: MAE, MSE, RMSE, R² help compare models
* **Visualization**: Helps interpret model performance

<https://colab.research.google.com/drive/1Oq7dgZco7_LBawB5ou9YbZsN9R77NmPN?usp=sharing>

**Problem#3**

1. **Title:** "Bangladeshi Traffic Sign Recognition and Classification using Convolutional Neural Networks"

This project will develop a deep learning system to recognize and classify Bangladeshi traffic signs, addressing the unique challenges of the local transportation environment. The project will utilize the publicly available "Bangladeshi Traffic Sign Dataset" containing images of various traffic signs commonly found in Bangladesh.

Data set: <https://www.kaggle.com/datasets/tusher7575/traffic-sign-in-bangladesh>

1. **Title:** Skin Cancer Classification using Deep Neural Networks

Skin Cancer Classification using Deep Neural Networks (DNNs) is a computer vision-based approach that automatically identifies different types of skin cancer from dermatoscopic images. The system is trained on large image datasets, such as the ISIC Skin Cancer Dataset, to learn patterns and features that distinguish malignant lesions (like melanoma) from benign ones. A DNN, often implemented using Convolutional Neural Networks (CNNs), extracts hierarchical image features — from edges and textures to complex skin patterns. After training, the model can classify new images with high accuracy, assisting dermatologists in early detection and diagnosis.

Data set : <https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000?fbclid=IwY2xjawNxLYpleHRuA2FlbQIxMABicmlkETE2MVd4WmFRbnZTZXU3NHB3AR5kAAoY4CaEohrlMd9sl-8NoHF6Zbz3oyAeRzQ5MGzfvM35wopNZMs0vAQr4g_aem_ctRsX1od96RIbRqMtZYCXw>

1. Title : Automated Blood Cell Detection & Counting on the BCCD Dataset using CNN-based Object Detection

This project focuses on developing an automated computer vision system for detecting and counting different types of blood cells in microscopic images using Convolutional Neural Networks (CNN). The system will identify and classify three main types of blood cells: Red Blood Cells (RBCs), White Blood Cells (WBCs), and Platelets from the BCCD (Blood Cell Count and Detection) dataset. This application has significant potential in medical diagnostics by automating the traditionally manual and time-consuming process of blood cell counting, reducing human error, and increasing efficiency in clinical laboratories.

Data set: <https://www.kaggle.com/datasets/paultimothymooney/blood-cells>