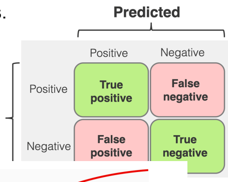
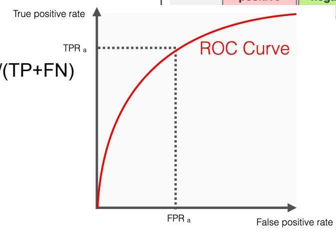
Loss Function **Definition**: A mathematical function that quantifies the difference between predicted and true values.

- **Types**: **Regression**: Mean Squared Error (MSE), Mean Absolute Error (MAE). **Classification**: Cross-Entropy Loss, Hinge Loss.
- **Purpose**: Guides model optimization during training by minimizing the error.
- **Gradient Descent**: Common optimization technique using the derivative of the loss function.



Confusion Matrix

- **Definition**: A table summarizing the performance of a classification model.
- **Components**: **True/False +ves/-ves [(T** or **F)(P** or **N])** - Correctly/Incorrectly predicted positive/negative instances.
- **Derived Metrics**: **Accuracy**: (TP+TN)/(TP+TN+FP+FN) **Precision**: TP/(TP+FP) – How many positive predictions were correct.
  - **Recall (Sensitivity)**: TP/(TP+FN) – How many actual positives were correctly predicted.
  - **F1 Score**: Harmonic mean of precision and recall: 2·(Precision·Recall)/(Precision+Recall).



ROC and AUC

- **ROC Curve**: Graphical plot of **True Positive Rate (Recall)** vs. **False Positive Rate**. **True/False Positive Rate**: (TP or FP)/(TP+FN)
- **AUC (Area Under Curve):**
  - Measures the ability of the model to distinguish between classes.
  - Ranges from 0.5 (random guessing) to 1.0 (perfect model).
  - Represents the probability that a randomly chosen positive instance ranks higher than a randomly chosen negative one.

Logistic Regression

- **Definition**: A statistical model for binary classification using a sigmoid function to output probabilities. **Equation**: σ(z)=1/(1 + e^{-z}), where z=β0+β1x1+⋯+βnxn
- **Loss Function**: Cross-Entropy Loss. **Regularization**: L1 (Lasso): Promotes sparsity in features. L2 (Ridge): Penalizes large coefficients to prevent overfitting.

Principal Component Analysis (PCA) **Definition**: Dimensionality reduction technique that transforms data into principal components, capturing maximum variance.

- **Steps**: 1. Standardize data. 2. Compute covariance matrix. 3. Perform eigendecomposition to get eigenvectors (directions) and eigenvalues (variance explained). 4. Select top components based on explained variance.
- **Use Case**: Reducing noise and computational complexity in high-dimensional data.

Naive Bayes (NB) **Definition**: Probabilistic classifier based on Bayes' theorem, assuming independence among predictors.

- **Bayes' Theorem**: P(A|B)=P(B|A)·P(A) / P(B). **Strengths**: Fast, simple, effective for text classification.
- **Variants**: Gaussian NB: For continuous data. Multinomial NB: For discrete counts. Bernoulli NB: For binary data.

Decision Tree (DT) **Definition**: A tree-structured algorithm for classification or regression.

- **Key Concepts**: Splits data recursively based on feature thresholds. Criteria: Gini Impurity, Entropy.
- **Advantages**: Interpretability, non-parametric. **Limitations**: Prone to overfitting (solved using pruning or ensembles).

k-Nearest Neighbors (kNN) **Definition**: Instance-based algorithm that assigns a class based on the majority vote of k nearest neighbors.

- **Key Steps**: 1.Compute distance (e.g., Euclidean, Manhattan). 2. Identify k nearest neighbors. 3. Predict based on majority class (classification) or average (regression).
- **Challenges**: Sensitive to outliers, computationally expensive for large datasets.

Support Vector Machine (SVM)

- **Definition**: A model that finds the hyperplane best separating data points of different classes.
- **Key Concepts**: **Margin**: Distance between the hyperplane and nearest data points. **Kernel Trick**: Transforms data into higher dimensions to find a linear separator. **Types:** Linear, Polynomial, RBF (Gaussian).
- **Loss Function**: Hinge Loss. **Strengths**: Effective in high-dimensional spaces, robust to outliers.

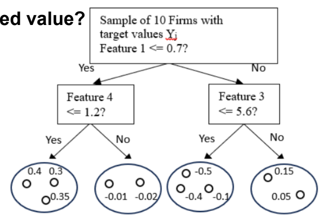
Neural Network (NN) **Definition**: Computational models inspired by biological neural networks, capable of approximating complex functions.

- **Components**: **Input Layer**: Features. **Hidden Layers**: Nonlinear transformations. **Output Layer**: Predictions. **Activation Functions**: Sigmoid, ReLU, Tanh.
- **Training**: Forward Propagation: Predicts output. Backward Propagation: Optimizes weights using gradients of the loss function.
- **Use Case**: Handles non-linear, high-dimensional data well.

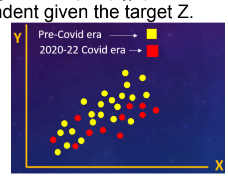
SHAP (SHapley Additive exPlanations) **Definition**: A framework for interpreting predictions by attributing feature contributions based on Shapley values.

- **Key Concepts**: Shapley Value: From game theory, measures the contribution of a feature by considering all permutations.
  - Additive Explanation: Breaks predictions into individual contributions.
- **Advantages**: Model-agnostic; Provides insights into feature importance and interaction. **Use Case**: Enhances trust and understanding in black-box models.

What is a key difference between classical regressions (CR) and ML methods? **Ans:** CR typically tests if features are statistically significant while ML does not typically do so.
Reason: Classical regressions focus on feature significance (hypothesis testing), whereas ML methods prioritize prediction accuracy.
In a ML prediction, the recall is 2/3, precision is 3/4, number of TP's is 1/5 of number of TN's. What is accuracy?
1. **Recall** = TP/(TP+FN)=2/3 2. **Precision** = TP/(TP+FP)=3/4 From the question, **TP** = 1/5 TN → **accuracy** comes out to be **75.67%**
Which of the following is an ensemble effect? **Ans:** Building cost estimation based on estimates submitted by a dozen established building corporations.
Reason: Ensemble methods combine multiple predictions (like averaging) from several models, similar to combining estimates here.
A simple DT predicting a target numerical value with conditions (Feature 3 > 5.6, Feature 1 > 0.7, Feature 4 <= 1.2). What is the predicted value?
Following the decision tree: **Feature 3 > 5.6** → **Yes** branch; **Feature 1 > 0.7** → **No** branch. Terminal node values are: 0.15, 0.05.
The **predicted value** is the average of these two values: (0.15+0.05)/2=0.10.



Minimizing objective function using L1 regularization term $\sum_{k=1}^{n}\left(Y_k - \sum_{j=0}^{p} b_j X_{jk}\right)^2 + \alpha \sum_{j=0}^{p}|b_j|$ **is Lasso**
Reason: The L1 norm (sum of absolute values) is characteristic of Lasso regression, which encourages sparsity in feature coefficients.
In an SVM classification, what is the perpendicular distance between parallel planes x+2y+z=5 and x+2y+z=7?
The formula for the perpendicular distance between two parallel planes ax+by+cz=d1 and ax+by+cz=d2 is: dist = |d2−d1|/sqrt(a^2+b^2+c^2)
Here: a=1, b=2, c=1, d1=5, d2=7. Substituting into the formula yields 2/sqrt(6).
In running the Neural Network (NN), how can you improve accuracy while reducing loss?
Increasing epochs improves training (but risks overfitting); Decreasing batch size allows more granular updates; Increasing the number of layers can enhance model complexity.
In logistic regression, prediction of class with label Y=1 is done through? **Ans:** Fitting Expected Value of Y given the features.
Reason: Logistic regression models the probability that Y=1 given the features using a sigmoid function.
PCA transforms original features into features with what property? **Ans:** Orthogonal transformed features.
Reason: PCA transforms features into uncorrelated (orthogonal) components.
Which ML algorithm is most sensitive to outliers? **Ans:** kNN (sensitive to outliers because distance-based calculations can be skewed by extreme values)
Effect of shrinking estimation: Shrinking estimation (e.g., through regularization) helps reduce overfitting, regularizes coefficients, and can indirectly perform dimension reduction.
Suppose Fj's are features, and Z is the target. If $P(F_1, F_2, \ldots, F_n|Z) = \prod_{j=1}^{n} P(F_j|Z)$, **what is the assumption?** **Ans:** P(Fj|Z,F1,F2,…,Fj−1,Fj+1,…,Fn)=P(Fj|Z)
Reason: This assumption reflects **conditional independence**, which is fundamental in Naive Bayes classifiers. Each feature Fj is conditionally independent given the target Z.
The graph shows data partitioned randomly without considering timing. What could be a critical problem? **Ans: Imbalanced data**
Reason: The graph shows two distinct distributions for **Pre-Covid** (yellow) and **Covid-era** (red) points.
Random partitioning may result in an **imbalanced dataset** where one class (e.g., Covid-era) dominates, affecting the model's ability to generalize well.



Which propagation in NN uses a learning rate? **Ans:** Gradient Descent
Reason: Gradient descent adjusts the weights during **backpropagation** using the learning rate to minimize the loss function.
In ML prediction, given a target of 1.70 and SHAP values +0.85, −0.65, +0.30, what is the average predicted target?
To compute the average predicted target: Base Value=1.70 − (0.85 − 0.65 + 0.30) = 1.20