Statistics for Machine Learning

week N ~ week N+ 3

Statistics for ML

Source : https://pabloinsente.github.io/intro-linear-algebra

Learning Objectives

- [x] What is Statistics?
- [x] Data Sampling
- [x] Descriptive Statistics
- [x] Inferential Statistics
- [x] Regression Analysis
- [x] Model Evaluation

What is Statistics?

- Statistics is the body of techniques used to facilitate the collection, organization, presentation, analysis, and interpretation of data for the purpose of making better decisions.

Learning Objectives

- [x] What is Statistics?
- [x] Data Sampling
- [x] Descriptive Statistics
- [x] Inferential Statistics
- [x] Regression Analysis
- [x] Model Evaluation

Data sampling

- Random Sampling(무작위표본)- Every item or member in the population has an equal chance of being selected for the sample. It reduces bias and ensures that the sample is representative of the population.

- Stratified Sampling(층화표본)- The population is divided into subgroups or strata based on certain characteristics (e.g., age, gender, location). Then, random sampling is performed within each stratum to ensure representation of all groups.

- Systematic Sampling- The starting point is randomly chosen, and then every “kth” item is included in the sample. It’s simple and often more efficient than simple random sampling.

Learning Objectives

- [x] What is Statistics?
- [x] Data Sampling
- [x] Descriptive Statistics
- [x] Inferential Statistics
- [x] Regression Analysis
- [x] Model Evaluation

Descriptive statistics

- To deal with the presentation and summary of data.
  - Mean (Average)- Measure the average value in the distribution of numerical data.
  - Median- Provide the average information with more efficient way compared to Mean and it is not affected by outlier in data.
  - Variance- Measure the Spread in data.
  - Standard Deviation — The square root of the variance, providing a more interpretable measure of data variability.
  - Percentile- It is a measure that indicated the percentage of data points that are equal to or below a specific value in a dataset.
  - IQR (Interquartile range)- It is the measure of range between first quartile and third quartile which helps to identify middle of 50 % of data.
  - Histogram- It is the measure of frequency or count of data points falling into specific intervals (bins) along the horizontal axis.
  - PDF (Probability Density Function)-It is a statistical function that describes the likelihood of a continuous random variable taking on a specific value within a given range.
  - CDF (Cumulative Density Function)- It is a statistical function that gives the cumulative probability that a random variable is less than or equal to a specific value.
  - Skewness- It describes the asymmetry in the distribution of data.
  - Kurtosis- It measures the tailedness of the data distribution.

Descriptive statistics

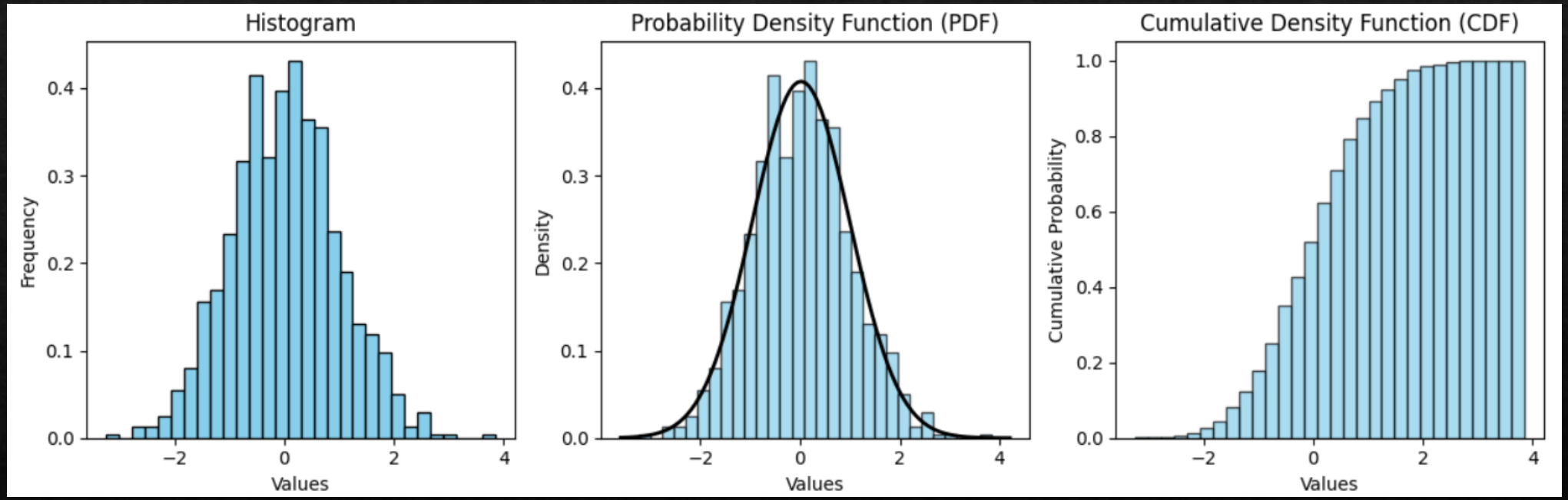
Chat GPT:

- Histogram, PDF (Probability Density Function), CDF (Cumulative Density Function)

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# Generate a random dataset
np.random.seed(42)
data = np.random.normal(size=1000)

# Plot Histogram
plt.figure(figsize=(12, 4))

plt.subplot(1, 3, 1)
plt.hist(data, bins=30, density=True, color='skyblue', edgecolor='black')
plt.title('Histogram')
plt.xlabel('Values')
plt.ylabel('Frequency')

# Plot Probability Density Function (PDF)
plt.subplot(1, 3, 2)
plt.hist(data, bins=30, density=True, color='skyblue', edgecolor='black', alpha=0.7)

# Fit a normal distribution to the data
mu, std = norm.fit(data)
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mu, std)
plt.plot(x, p, 'k', linewidth=2)

plt.title('Probability Density Function (PDF)')
plt.xlabel('Values')
plt.ylabel('Density')

# Plot Cumulative Density Function (CDF)
plt.subplot(1, 3, 3)
plt.hist(data, bins=30, density=True, color='skyblue', edgecolor='black', cumulative=True, alpha=0.7)

plt.title('Cumulative Density Function (CDF)')
plt.xlabel('Values')
plt.ylabel('Cumulative Probability')

plt.tight_layout()
plt.show()
```

Descriptive statistics

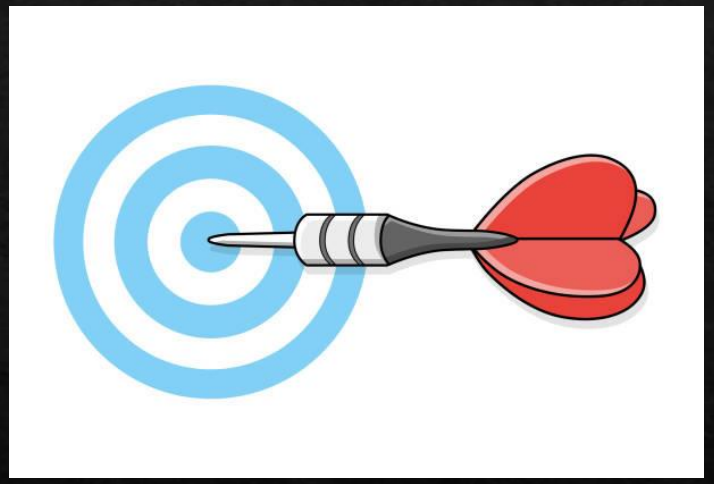
Chat GPT:



Learning Objectives

- [x] What is Statistics?
- [x] Data Sampling
- [x] Descriptive Statistics
- [x] Inferential Statistics
- [x] Regression Analysis
- [x] Model Evaluation

점/구간 추정(Point/Interval Estimation)



점/구간 추정(Point/Interval Estimation)

Chat GPT:

- **구간추정 (Interval Estimation)**
  - 모수가 존재할 가능성이 높은 구간을 제시하는 방식

- **모평균의 구간추정**
  - 신뢰수준(Confidence level) : 구간 $(\bar{X}-d, \bar{X}+d)$ 에 모수 μ가 포함될 확률
  - 보통 100(1 - α)%로 나타냄
  - 신뢰구간(Confidence interval) : 신뢰수준에 대응하여 도출되는 μ 의 추정구간
  - 신뢰수준이 커질수록 신뢰구간은 넓어짐.

$$P\{\bar{X}-d<\mu<\bar{X}+d\}=1-\alpha$$

"정부의 이번 정책에 대한 표본조사를 실시한 결과 찬성한다는 비율이 **74%**였다. … 이번 조사의 신뢰수준은 **95%**이고 오차한계는 **±4%**이다."

점/구간 추정(Point/Interval Estimation)

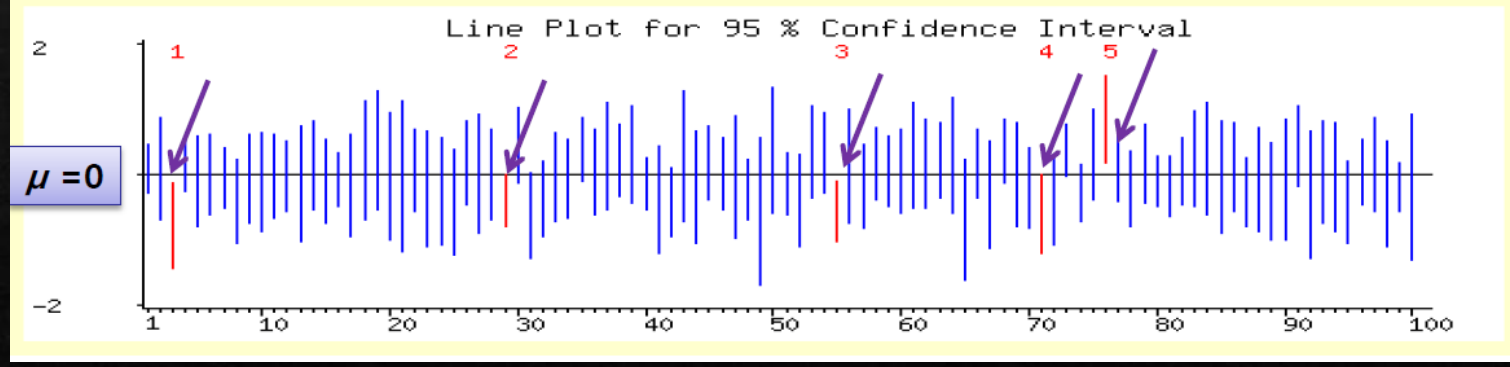
Chat GPT:

95% 신뢰구간이 의미하는 것은?

'100번의 반복샘플링을 통해 얻은 평균과 편차로 계산한 100개의 신뢰구간 중 5개는 실제모평균(μ)을 포함하고 있지 않는다' 혹은 '표본을 통해 얻은 95% 신뢰구간에 실제 모평균이 포함되지 않을 확률은 5%이다' 라고 해석할 수 있다.

(예 : 아래 그림에서 5개의 신뢰구간은 0을 지나가지 않는다)

신뢰수준이 높아지면, 신뢰구간은 넓어짐.



실습하기

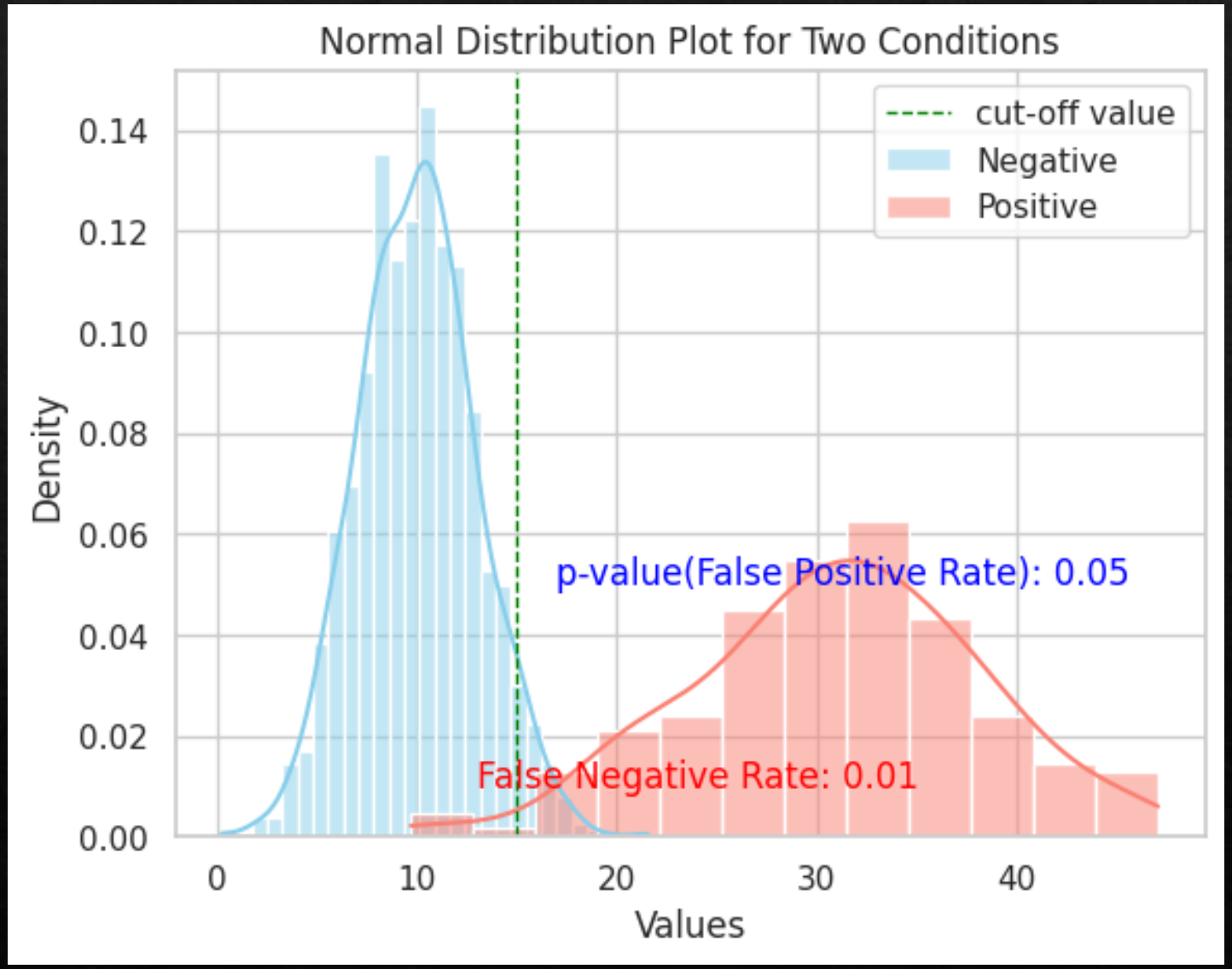
P-value

Chat GPT:

```
# Set the style for Seaborn
sns.set(style="whitegrid")

# Generate two sets of data (two conditions)
np.random.seed(42)
condition_1 = np.random.normal(loc=10, scale=3, size=1000)
condition_2 = np.random.normal(loc=30, scale=7, size=200)

cutoff = 15

# Calculate CDF values for Condition 1 and Condition 2 up to the cutoff value
cdf_condition1 = stats.norm.cdf(cutoff, loc=np.mean(condition_1), scale=np.std(condition_1))
cdf_condition2 = stats.norm.cdf(cutoff, loc=np.mean(condition_2), scale=np.std(condition_2))

# Plot for Condition 1/2
sns.histplot(condition_1, kde=True, color='skyblue', label='Negative', stat='density')
sns.histplot(condition_2, kde=True, color='salmon', label='Positive', stat='density')

# Display the p-value on condition_1
plt.text(cutoff+2, 0.05, f'p-value(False Positive Rate): {(1-cdf_condition1):.2f}', color='blue')
plt.axvline(x=cutoff, color='green', linestyle='dashed', linewidth=1, label='cut-off value')
# Display the p-value on condition_2
plt.text(cutoff-2, 0.01, f'False Negative Rate: {(cdf_condition2):.2f}', color='red')

plt.title('Normal Distribution Plot for Two Conditions')
plt.xlabel('Values'); plt.ylabel('Density')
plt.legend()
plt.show()
```

P-value

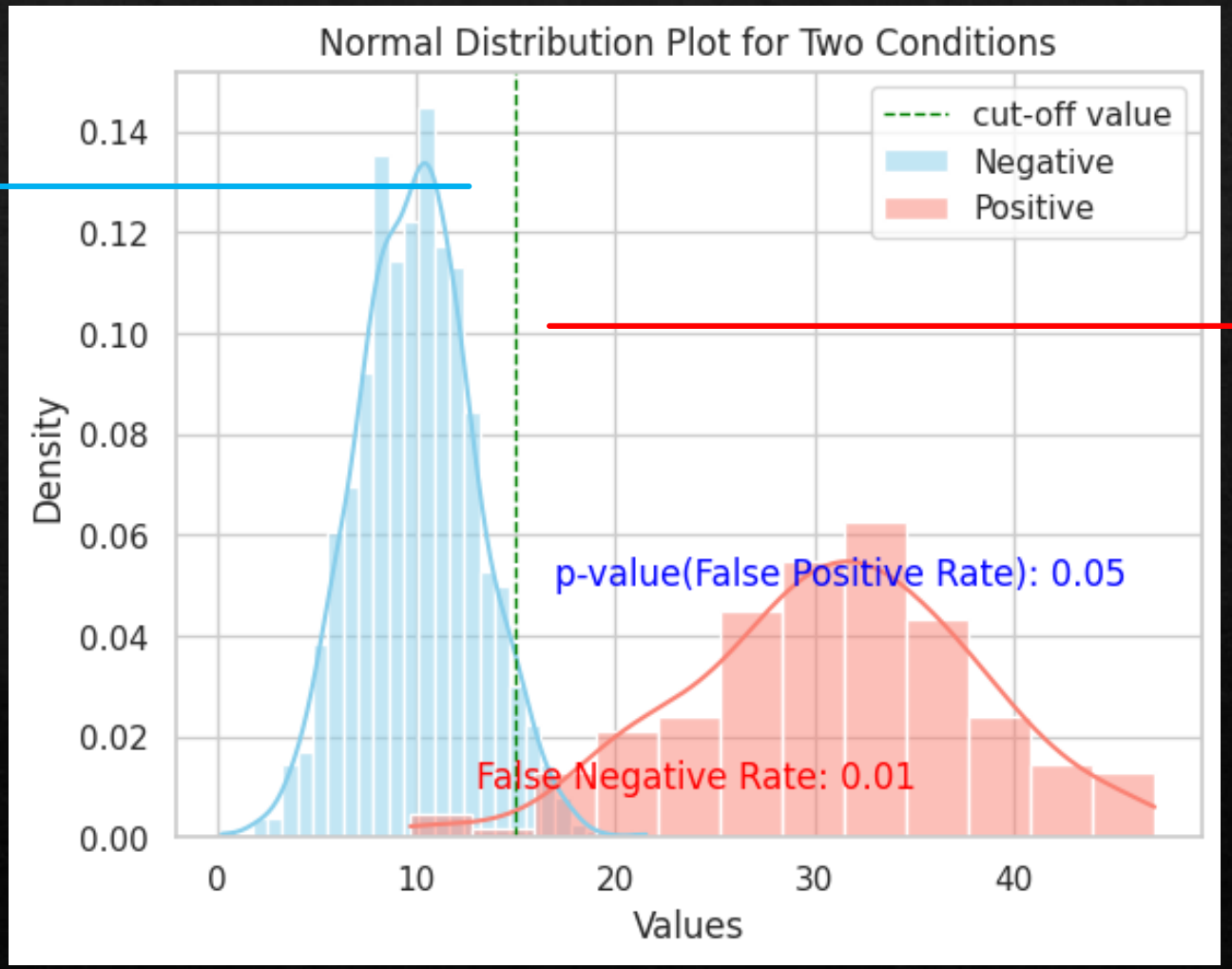
Chat GPT:



Confusion Matrix

Chat GPT:

Negative 판정
(Null Hypothesis)



Positive 판정
(Alternative Hypothssis)

Learning Objectives

- [x] What is Statistics?
- [x] Data Sampling
- [x] Descriptive Statistics
- [x] Inferential Statistics
- [x] Regression Analysis
- [x] Model Evaluation

Regression

- **Linear Regression-** It makes relationship between a dependent variable and one or more independent variables by fitting a linear equation to the data.

- **Multiple Regression-** It incorporate two or more independent variables to predict a single dependent variable.

- **Polynomial Regression-** It make relationship between variables appears to be nonlinear, this model fits a polynomial (e.g., quadratic or cubic) equation to the data.

- **Ridge Regression and Lasso Regression-** Variations of linear regression that incorporate regularization techniques to handle multicollinearity and prevent overfitting.

Learning Objectives

- [x] What is Statistics?
- [x] Data Sampling
- [x] Descriptive Statistics
- [x] Inferential Statistics
- [x] Regression Analysis
- [x] Model Evaluation

Model Evaluation

Chat GPT:

Regression Model

- Accuracy- Accuracy measures the proportion of correctly classified instances in a classification model.
- Mean Absolute Error (MAE)- MAE measures the average absolute difference between the predicted values and the actual values.
- Mean Squared Error (MSE)- MSE calculates the average of the squared differences between predicted and actual values.
- Root Mean Squared Error (RMSE)- RMSE is the square root of MSE, providing an interpretable metric in the same units as the target variable.
- R-squared (R^2) or Coefficient of Determination- R^2 measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model.

Classification Model

- Area Under the Receiver Operating Characteristic (ROC AUC)- It measures the area under the receiver operating characteristic curve, which plots the trade-off between true positive rate (recall) and false positive rate at various thresholds.
- Confusion Matrix- A table that shows the number of true positives, true negatives, false positives, and false negatives, providing detailed insights into the performance of a classification model.
- Precision- Measures the ratio of true positive predictions to the total positive predictions, emphasizing the model's ability to avoid false positives.
- Recall- Measures the ratio of true positives to the total actual positives, emphasizing the model's ability to find all relevant instances.
- F1-Score- The harmonic mean of precision and recall, offering a balance between the two metrics.