

# SPPU-BE-COMP-CONTENT - KSKA Git

ML

## ASSIGNMENT-1

classmate

Date :

Page :

Q1

### 1. Supervised Learning

- model learnt from labelled data.
- goal is to predict correct output for new, unseen inputs.
- requires training dataset with labels.
- uses input features to predict target labels.
- evaluated using accuracy, precision, recall
- ex: predicting email is spam (1) or not (0) using past labelled emails.

### 2. Unsupervised Learning:

- model learns from unlabelled data
- Discover hidden patterns or groupings in data
- works without labelled responses.
- used for exploratory data analysis.
- evaluated using clustering metrics.

Q2

### 3. Semi supervised Learning

- uses mix of labelled & unlabelled data.
- combines supervised & unsupervised techniques
- useful when labelling data is expensive or time consuming
- improves model accuracy by leveraging unlabelled data.

# SPPU-BE-COMP-CONTENT - KSKA Git

Q2

## 1) Regression Analysis:

- Models rel<sup>n</sup> b/w dependent variable & 1 or more independent variable
- Linear regression: fits straight line
- Multiple regression: uses several pred<sup>r</sup>
- Evalu<sup>n</sup>:  $R^2$ , MSE

## 2) Classification

- Assigns Data points to discrete categories
- Logistic Regression: predicts probabilities
- Linear ~~Regression~~ Discriminant Analysis: finds linear separ<sup>n</sup> b/w classes
- eval<sup>n</sup>: Accuracy, Precision, Recall.

## 3) Bayesian Methods:

- uses probability theory to update beliefs based on observed data.
- Naive Bayes: Assumed feature independence.
- Bayesian networks: models dependence b/w variables
- Adv: handles uncertainty well.

## 4) Clustering:

- Groups similar data points without preassigned labels
- K means: partitions data into k clusters.
- Hierarchical clustering: creates tree of clusters



# SPPU-BE-COMP-CONTENT - KSKA Git

Date :  
Page :

Q3

min max scaling:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$\min(x) = 18$$

$$\max(x) = 52$$

$$\text{Range} = 52 - 18 = 34$$

$$\text{for } 23 \rightarrow 0.147$$

$$29 \rightarrow 0.324$$

$$52 \rightarrow 1.000$$

$$31 \rightarrow 0.382$$

$$45 \rightarrow 0.794$$

$$19 \rightarrow 0.029$$

$$18 \rightarrow 0$$

$$27 \rightarrow 0.265$$

$$\text{z-score norm}^n = z = \frac{x - \mu}{\sigma}$$

$$\text{mean}(\mu) = \frac{23 + 29 + 52 + 31 + 45 + 19 + 18 + 27}{8}$$

$$= 30.5$$

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

$$\text{sum of squared} = 56.25 + 2.25 + 462.25 + 0.25 + 210.25 + 132.25 + 156.25 + 12.25 = 1032.0$$

$$\sigma = \sqrt{\frac{1032}{8}} = \sqrt{129} = 11.36$$

$$\text{zscore, } 23 \rightarrow -0.66, \quad 29 \rightarrow -0.13$$

$$52 \rightarrow +1.89, \quad 31 \rightarrow +0.04, \quad 45 \rightarrow +1.28$$

$$19 \rightarrow -1.01, \quad 18 \rightarrow -1.10, \quad 27 \rightarrow -0.31$$

# SPPU-BE-COMP-CONTENT - KSKA Git

Q4

## Principal Component Analysis

- Technique used to reduce dimensionality of large datasets while retaining most of var<sup>n</sup> present in original data.
- helps simplifying data, removing noise & improving computational efficiency.

### 1. Standardize Data:

each feature is transformed to have zero mean & variance.

### 2. Compute covariance matrix:

calculated to measure how variables vary together.

### 3. Calculate eigen values & vectors: represents max variance in data, while eigen values indicate amt of variance captured by each principal component.

### 4. Transform original Data:

original data is projected into selected principal components resulting in a reduced feature set with minimal loss of information