NAME: KSHITIJ VINOD SALI
CLASS & DIV: TE-A
ROLL NO.: 35059

ML
UA-01

Title: Data preperation of heart dataset.

Problem statement: Part-1: Data preperation of heart dataset. Download heart dataset from following link: https://www.kaggle.com/zhaoyingzhu/heartcsv

Perform following operation on given dataset -

a) Find shape of data
b) Find missing values.
c) Find data type of each column
d) Finding out zero's
e) Find mean age of patients
f) Now extract only Age, Sex, ChestPain, RestBP, Chol. Randomly divide dataset in ~~tim~~ training (75%) and testing (25%).

Part-2: Through the diagnosis test I predicted 100 report as COVID positive, but only 45 of those were actually positive. Total 50 people in my sample were actually COVID positive. I have total 500 samples.

Create confusion matrix based on above data and find -

i) Accuracy
ii) Precision
iii) Recall
iv) F-1 score

Objective: Fundamental elements of machine learning how to work on different datasets and prepare data.

Pre-requisites: Python, Discrete Structure.

Experimental structure: Python programming, Jupyter notebook, Google colab.

Theory:

• Part - 1 •

1) Write short theory on types of data (Qualitative and Quantitative).

→ Qualitative data:

i) Associated with numbers.
ii) Implemented when data is numerical.
iii) Collected data can be statistically analyzed.
iv) Examples: Height, Weight, Time, Price, Temperature, etc.

Quantitative data:

i) Associated with details.
ii) Implemented when data can be segregated into well-defined groups.
iii) Collected data can just be observed and not evaluated.
iv) Examples: Scents, Appearance, Beauty, Colors, Flavors, etc.

2) Scales of measurement (Nominal, Ordinal, Interval, Ratio).

→ Nominal Scale:

i) The nominal scale, sometimes called the qualitative type, places non numerical data into categories or classifications.

ii) For example -
   a) Placing cats into breed type. Example: a Persian is a breed of cat.
   b) Putting cities into states: Example: Jacksonville is a city in Florida.
   c) Surveying people to find out if men or women have higher self-esteem.
   d) Finding out if introverts or extroverts are more likely to be philanthropic.

iii) These pieces of information aren't numerical. They are assigned a category (breeds of cat, cities in Florida, men & women, introvert & extrovert) Qualitative variables are measured on the normal scale.

Ordinal Scale:

i) Ordinal scales are made up of ordinal data.

ii) Some examples of ordinal scales:
   a) High school class ranking: 1st, 2nd, 3rd, etc.
   b) Social economic class: working, middle, upper
   c) The likert scale: agree, strongly agree, disagree, etc.

iii) A second example of the ordinal scale: Conducting a survey & ask people to rate their level of satisfaction which choice of following responses:
   a) Extremly satisfied
   b) Satisfied

c) Neither satisfied nor dissatisfied.
d) Dissatisfied.
e) Extremly dissatisfied.

The choices from "extremely satisfied" to "extremely dissatisfied" follow a natural order and are therefore ordinal variables.

Interval:

i) Interval has values of equal intervals that mean something.
ii) For example, a thermometer might have intervals of ten degrees.
iii) Example:- a) Celsius temperature.
b) Fehrenheit temperature.
c) IQ (Intelligance scale)
d) SAT scores.
e) Time on a clock with hands.

Ratio:

i) Ratio, exactly the same as the interval scale except that the zero on the scale means: does not exist.
ii) For example; a weight of zero doesn't exist; an age of zero doesn't exist. On the other hand, temperature is not a ratio scale, because zero exist.
iii) Example -
a) Age
b) Weight
c) Height
d) Sales figures.
e) Ruler measurements.
f) Income earned in a week.
g) Years of education
h) Number of children.

3) Concept of feature:

i) Features is anything that you can measure and build data for.

ii) A feature is a numeric representation of raw data.

iii) For example, the typical length of various animals.

iv) Feature could be numeric, set of characters, Boolean values, or anything else that describes the data.

v) But, for most machine learning mathematics models, features are required to be numeric so that they can be used in various computation.

vi) The features in a data set are also called its dimensions. So a data set having features is called an n-dimensional data set.

vii) For example -

Gender	Marks
Girl	65
Girl	46
Boy	56
Boy	43
Boy	53
Boy	49
Girl	42
Boy	84
Boy	44
Girl	42
Girl	40

Here, only two features are there, hence it is a 2 dimensional dataset.

4) Feature construction:

i) Feature construction process discovers missing information about the relationship between features & expands the feature space by creating additional features.

ii) Hence, if there are n features or dimensions in a data set, after feature construction m more features or dimensions may get added.

iii) So at the end, the data set will become n+m dimensional.

5) Feature Selection:

i) Feature selection is the process of reducing the number of input variables when developing a predictive model.

ii) It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

6) Transformation:

i) Feature transformation is used as an effective tool for dimensionally reduction & hence for boosting learning model performance.

ii) Broadly, there are two distinct goals of feature transformation:

a) Achieving best reconstruction of the original features in the data set.

b) Achieving highest efficiency in the learning task.

iii) Feature transformation could be applied to

numeric feature or non-numeric features such as text & images.

vi) Feature transformation generally involves features construction & feature extraction.

• Part - 2 •

Q. Solve example of COVID and find all parameters - Accuracy, Precision, Recall, F-1 score.

Ans: Total sample = 50.

		Predicted: Positive	Predicted: Negative
Actual →	Positive	45	5
	Negative	55	395

100

TP = 45, TN = 395, FP = 55, FN = 5.

i) Accuracy $= \frac{TP + TN}{TP + TN + FP + FN} = \frac{45 + 395}{45 + 395 + 55 + 5}$

$= \frac{440}{500} = 0.88$

ii) Recall $= \frac{TP}{TP + FN} = \frac{45}{45 + 5} = 0.90$

iii) Precision $= \frac{TP}{TP + FP} = \frac{45}{45 + 55} = 0.45$

iv) F1-score $= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

$= \frac{2 \times 0.45 \times 0.90}{0.45 + 0.90}$

$= 0.6$

Oral Questions.

Q. 1) What are different types of machine learning algorithms?

Ans: The linear types of machine learning algorithms are:

a) Linear Regression.
b) Logistic Regression.
c) Decision Tree
d) SVM
e) Naive Bayes.
f) kNN
g) K-Means.
h) Random Forest.

Q. 2) What is Supervised learning?

Ans:

i) Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output.

ii) The labelled data means some input data is already tagged with the correct output.

iii) In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

Q. 3) What is Unsupervised learning?

Ans:

i) Unsupervised learning is a machine learning technique in which models are not supervised using training dataset.

ii) Instead, models itself find the hidden patterns & insights from the given data.

Q.4) What is Cross-Validation?

Ans:

i) Cross-validation is a statical method used to estimate the performance / accuracy of machine learning models.

ii) It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

iii) In cross-validation, you make a fixed number of folds / partitions of data, run the analysis on each fold & then average the overall error estimate.

Q.5) Explain the difference between classification and regression.

Ans:

Regression	Classification
i) In regression, the output variable must be of continuous nature or real value.	i) In classification, the output variable must be a discrete value.
ii) The task of the regression algorithm is to map the input value (x) and the continuous output variable (y).	ii) The task of classification algorithm is to map the input value (x) with the discrete output variable (y).
iii) Regression algorithm are used with continuous data.	iii) Classification algorithm are used with discrete data.

iv) In regression, we try to find the best fit line, which can predict the output more accurately.	iv) In classification, we try to find the decision boundary, which can divide dataset into different classes.
v) Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	v) Classification algorithm can be used to solve classification problem such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
vi) The regression algorithm can be further divided into Linear & non-linear regression.	vi) The classification algorithm can be divided into binary classifier & multi-class classifier.

Q. 6) Define accuracy, precision & recall & f1-score.

Ans: Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observation.

$$\therefore \text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\therefore \text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to all observations in actual class.

$$\therefore \text{Recall} = \frac{TP}{TP + FN}.$$

F1 - Score - F1 score is the weighted average of precision and recall.

$$\therefore \text{F1 score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}.$$

Q. 7) With reference to feature engineering, explain data scaling and normalization tasks?

Ans:

i) Some features, such as latitude or longitude are bounded in value. Other numeric features, such as counts, may increase without bound.

ii) Models that are smooth functions of the input, such as linear regression, logistic regression or anything that involves a matrix are affected by scale of the input.

iii) If our model is sensitive to the scale of input features, feature scaling could help. As the name suggests, feature scaling changes the scale of feature. It is also called as feature normalization.

iv) Feature scaling is usually done individually to each feature.