ASSIGNMENT NO. 2
DMBI Lab

Q2. Compare the following:

a. **Classification Techniques**
b. **Clustering Techniques**

Classification Techniques

1. **Naïve Bayes** - Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering.

   Advantages: This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

   Disadvantages: Naive Bayes is known to be a bad estimator.

```
from sklearn.naive_bayes import GaussianNB
nb =  GaussianNB()
nb.fit(x_train, y_train)
y_pred=nb.predict(x_test)
```

2. **Decision Tree** - Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

   Advantages: Decision Tree is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data.

   Disadvantages: Decision trees can create complex trees that do not generalize well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

```
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier(max_depth=10, random_state=101,
                               max_features = None, min_samples_leaf = 15)
dtree.fit(x_train, y_train)
y_pred=dtree.predict(x_test)
```

Classification Algorithms	Accuracy	F1-Score
Naïve Bayes	80.11%	0.6005
Decision Tree	84.23%	0.6308

Clustering Techniques

1. **K-means Algorithm** - K-means is a well known partitioning method. Objects are classified as one of the k groups , k chosen as priory. Cluster membership is determined by calculating the centroid of each group and assigning each object to the group with the closest centroid. This approach minimizes the overall within cluster dispersion by iterative reallocation of cluster members.

   In a general sense, a k-partitioning algorithm takes as input a set S of objects and an integer k and outputs a partition of S into subsets S1, S2, S3...Sk. It uses the sum of squares as the optimization criterion Let xri be the rth element of Si, |Si| be the number of elements in Si and d(xri,xsi) be the distance between xri and xsi.

2. **Hierarchical Algorithm** - Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. In contrast, hierarchical algorithms combine or divide existing groups creating a hierarchical structure that reflects the order in which groups are merged or divided.

   In an agglomerative method, which builds the hierarchy by merging, the objects initially belong to a list of singleton sets S1,..., S2, Sn. Then a cost function is used to find the pair of sets {Si, Sj} from the list that is the “cheapest” to merge. Once merged, Si and Sj are removed from the list of sets and replaced with Si U Sj.

   This process iterates until all objects are in a single group. Different variants of agglomerative hierarchical clustering algorithms may use different cost functions.Complete linkage, average linkage, and single linkage methods use maximum, average, and minimum distances between the members of two clusters, respectively.

Q3. Comment on Accuracy calculation of Classification, Regression and Association mining.

Classification

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Accuracy = Number of correct predictionsTotal number of predictions

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

Accuracy = TP+TN / TP+TN+FP+FN
where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Let's try calculating accuracy for the following model that classified 100 tumors as malignant (the positive class) or benign (the negative class):

False Negative (FN): • Reality: Malignant • ML model predicted: Benign • Number of FN results: 8	True Negative (TN): • Reality: Benign • ML model predicted: Benign • Number of TN results: 90
False Negative (FN): • Reality: Malignant • ML model predicted: Benign • Number of FN results: 8	True Negative (TN): • Reality: Benign • ML model predicted: Benign • Number of TN results: 90

Regression

1. **R Square/Adjusted R Square** - R Square measures how much variability in dependent variables can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R Square is calculated by the sum of squares of prediction error divided by the total sum of the square which replaces the calculated prediction with mean. R Square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value. R Square is a good measure to

determine how well the model fits the dependent variables. However, it does not take into consideration the overfitting problem.

2. **Mean Square Error (MSE) / Root Mean Square Error (RMSE)** - While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

MSE is calculated by the sum of squares of prediction error which is real output minus predicted output and then divided by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.

3. **Mean Absolute Error (MAE)** - Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of squares of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Compared to MSE or RMSE, MAE is a more direct representation of the sum of error terms. MSE gives larger penalization to big prediction errors by square it while MAE treats all errors the same.

Q4. Write short notes on the following:

a. **F-P growth Algorithm**
b. **Business Intelligence**

F-P growth Algorithm

FP-tree (Frequent Pattern tree) is the data structure of the FP-growth algorithm for mining frequent itemsets from a database by using association rules. It's a perfect alternative to the apriori algorithm.

Mining patterns from a database have been a research subject; most previous studies suggested an Apriori-like candidate set generation and test approach. But it is pretty slow, and it becomes slower when there are many patterns available in mining. Therefore, FP-tree is proposed. The alternative of the apriori-like algorithm, the frequent-pattern tree(FP-tree) structure, is a tree data structure for storing frequent patterns.

The algorithm is designed to operate on databases containing transactions, such as customers' purchase history on the Amazon website. The purchased item is considered 'frequent'. The similar frequent will share the similar branch of the tree, and when they differ, the nodes will split them. The node identifies a single item from the branch(set of items), and the branch (path) shows the number of occurrences—links between the items called node-link.

This structure helps to find the required frequent set rapidly. Internally FP-growth is an algorithm that does not require candidate generation. It uses an FP-tree data structure that does not require the generation of candidate sets explicitly, making the algorithm work better with large databases.

For support and confidence, FP-tree finds the algorithm of frequent sets. Support is the frequency of an item or set in a database where the confidence is the probability of an item set occurrence with its set of items.

Business Intelligence

Business intelligence (BI) combines business analytics, data mining, data visualization, data tools and infrastructure, and best practices to help organizations to make more data-driven decisions.

Much more than a specific "thing," business intelligence is rather an umbrella term that covers the processes and methods of collecting, storing, and analyzing data from business operations or activities to optimize performance. All of these things come together to create a comprehensive view of a business to help people make better, actionable decisions. Over the past few years, business intelligence has evolved to include more processes and activities to help improve performance. These processes include:

1. **Data mining** - Using databases, statistics and machine learning to uncover trends in large datasets.
2. **Reporting** - Sharing data analysis to stakeholders so they can draw conclusions and make decisions.
3. **Performance metrics and benchmarking** - Comparing current performance data to historical data to track performance against goals, typically using customized dashboards.
4. **Descriptive analytics** - Using preliminary data analysis to find out what happened.
5. **Querying** - Asking the data specific questions, BI pulling the answers from the datasets.
6. **Statistical analysis** - Taking the results from descriptive analytics and further exploring the data using statistics such as how this trend happened and why.
7. **Data visualization** - Turning data analysis into visual representations such as charts, graphs, and histograms to more easily consume data.
8. **Visual analysis** - Exploring data through visual storytelling to communicate insights on the fly and stay in the flow of analysis.

9. **Data preparation** - Compiling multiple data sources, identifying the dimensions and measurements, preparing it for data analysis.

A few ways that business intelligence can help companies make smarter, data-driven decisions:

- Identify ways to increase profit
- Analyze customer behavior
- Compare data with competitors
- Track performance
- Optimize operations
- Predict success
- Spot market trends
- Discover issues or problems

91. classification :

Finance	Travel	Reading	Health	Sex
Yes	No	Yes	No	Male
Yes	Yes	No	No	male
No	Yes	Yes	Yes	Female
No	Yes	No	Yes	Male
Yes	Yes	Yes	Yes	Female
No	No	Yes	No	Female
Yes	No	No	No	Male
Yes	Yes	No	No	Male
No	No	No	Yes	Female
Yes	No	No	No	Male

P(Male) = 6/10 = 3/5

P(Female) = 4/10 = 2/5

Given a test record :

X = (Finance = No, Travel = Yes, Reading = Yes, Health = No)

Therefore,

P(X | class = Male) = P(Finance = No | class = Male) × P (Travel = Yes | class = Male)
× P (Reading = Yes | class = Male) × P (Health = No | class = Male)

= 1/4 × 3/5 × 1/4 × 3/5

= 9/400 = 0.0225

P(X | class = Female) = P (Finance = No | class = Female) × P (Travel = Yes | class = Female)
× P (Reading = Yes | class = Female) × P (Health = No | class = Female)

= 3/4 × 2/5 × 3/4 × 2/5

= 36/400 = 0.09

Since, $P(X|Male) \times P(Male) < P(X|Female) \times P(Female)$

$0.0225 \times 3/5 < 0.09 \times 2/5$

Therefore (Finance = No, Travel = Yes, Reading = Yes, Health = No) has the class of Female.

Clustering:

object 1 tuple = (2, 1, 4, 10)

object 2 tuple = (20, 10, 35, 8)

Euclidean Distance $d_E = \sqrt{\sum_{i=1}^{d} |P_i - Q_i|^2}$

$\therefore d_E = \sqrt{|2-20|^2 + |1-10|^2 + |4-35|^2 + |10-8|^2}$

$= \sqrt{324 + 81 + 961 + 4}$

$= \sqrt{1370}$

$= 37.0135$

Manhattan Distance $d_M = \sum_{i=1}^{d} |P_i - Q_i|$

$\therefore d_M = |2-20| + |1-10| + |4-35| + |10-8|$

$= 18 + 9 + 31 + 2$

$= 62$

Minkowski Distance ($p=3$) $d_{Mi} = \sqrt[3]{\sum_{i=1}^{d} |P_i - Q_i|^3}$

$\therefore d_{Mi} = \sqrt[3]{|2-20|^3 + |1-10|^3 + |4-35|^3 + |10-8|^3}$

$= \sqrt[3]{5832 + 729 + 29791 + 8}$

$= \sqrt[3]{36360}$

$= 33.1289$

Association Mining :

T_id	Items brought
10	Beer, Nuts, Diapers
20	Beer, Coffee, Diapers, Nuts
30	Beer, Diapers, Eggs
40	Beer, Nuts, Eggs, Milk
50	Nuts, Coffee, Diapers, Eggs, Milk

Given association rule :

{ Diapers } => { Coffe, Nuts }

Support (Diapers => Coffe, Nuts)

$$= \frac{\text{Tuples containing Diapers, Coffe and Nuts}}{\text{Total number of tuples}}$$

$$= 1/5 = 20\%$$

Confidence (Diapers => Coffe, Nuts)

$$= \frac{\text{Tuples containing Diapers, Coffe, Nuts}}{\text{Tuples containing Diapers}}$$

$$= 1/4 = 25\%$$