**Evaluation Assignment 3 Supervised Learning**

1. Which technique(s) from the following would you use…

* Logistic Regression
* Support Vector Machines
* Bayesian classifier
* K-Nearest Neighbour
* Principal Components Analysis
* Decision Trees

…if you…

Care only about accuracy and not about interpretation?

Support Vector Machines

Care about interpretation more than accuracy?

Decision Trees, Logistic Regression, Bayesian Classifier

Scoring throughput is high?

Logistic Regression, Decision Trees

Want to build a robust and deterministic model?

Support Vector Machines

Each class has potentially multiple sub-classes?

Bayesian Classifier

Want to visualize data that has no class labels?

Principal Components Analysis

Don’t want to build a model at all

K-Nearest Neighbour

2.

* **news.google.com**

○ Classification Problem

Google classifies the news based on many categories like games, technology, movies, science, etc and shows them accordingly

○ Retrieval Problem

We can get the news based on our search on the topic. We will get all the latest trends on the topic retrieved by google.

○ Recommendation Problem

Google recommends news for us. It will suggest the latest news on the topic you like the most. Google learns this from your previous reads on the topics and your search history.

* **youtube.com**

○ Classification Problem

Youtube classifies videos based on many categories like gaming, news, movies, music, science, etc.

○ Retrieval Problem

We can get the videos on the topic we search in youtube from various subscribers and playlists.

○ Recommendation Problem

Youtube will show recommendations on its homepage. This will be available from your previous seen videos and most seen videos. If you search a topic and start seeing a video, then youtube will keep recommendations below the video or the next video on the playlist on the same topic and saves your time.

* **linkedin.com**

○ Retrieval Problem

Linkedin retrieves specific users based on their roles or their qualification, etc. We can also check for job vacancies on a particular domain from LinkedIn.

○ Recommendation Problem

LinkedIn recommends you friends and also notifies you about the job vacancies based on your qualification.

* **ola.com/uber.com**

○ Retrieval Problem

Ola/Uber retrieves many pieces of information like location, distances, nearest drivers, coupons, cab availability, etc.

○ Recommendation Problem

Ola/Uber gives user recommendations on the type of transport he can go and on the surge details. It will give details on the routes user can go to reach his destination.

* **99acres.com/magicbricks.com**

○ Classification Problem

These sites will provide the user the details on different types of houses,

i.e. like 2bedroom, 3bedroom, etc.

○ Retrieval Problem

User can get information about the hospitals, schools, playgrounds, malls nearby the house he wanted to buy for.

○ Regression Problem

User can get different types of houses based on his requirement. He can get the suggestion based on the queries he has.

3.

* **K-nearest Neighbour and k-means**

# Similarities

* K-nearest and k-means are used for analyzing and grouping similar data together
* KNN Algorithm is based on feature similarity and K-means refers to the division of objects into clusters

# Differences

* KNN is a classification technique and K-means is a clustering technique.
* K-means have fixed number of clusters and a variable number of data points for each cluster. KNN has a fixed number of data points for each cluster.

* **PCA and MDS**

# Similarities

* Both PCA and MDS are dimensionality reduction techniques
* If the distance is euclidean distances, then both yield the same outputs
* Both the methods are used to reduce the number of random variables under consideration by obtaining a set of principal variables

# Differences

* Input to PCA is original vectors and input to MDS is pairwise distances between points
* In PCA we preserve variance and in MDS we try to preserve distance ● PCA is a linear method whereas MDS is a non-linear method

* **Classification and Regression**

# Similarities

* Classification and regression comes under supervised machine learning algorithms
* Classification and regression are used to predict possible trends

# Differences

* Regression is applied on numerical data (either continuous or discrete) and classification is applied on categorical data.
* Classification is about predicting a label and regression is about predicting a quantity

* **Bag of words and Market Basket Analysis**

# Similarities

* For Market basket analysis we need a collection of objects and that can be obtained using a bag of words method.
* Bag of words helps to design the vocabulary of known words and how to score the presence of known words which is similar in Market Basket Analysis where the objects are scored.

# Differences

* Bag-of-words modal is a way of representing text data for natural language processing and information retrieval.
* Market basket analysis is one of the key techniques used to uncover the relationships between items.
* Bag of words only gives the numbers whereas Market Basket Analysis will give the predictions and accuracies for selection of items.

* **Naive Bayes and Decision trees**

# Similarities

* If there is a way for Naive Bayes to pick up features on its own then there is not much difference between Naive Bayes and Decision trees.
* If you need to combine Naive Bayes with other statistical techniques to help guide you towards what features best classify then you can use the help of decision trees.

# Differences

* Decision trees are flexible and work for classification and regression problems. Here, the classifiers are built directly and we just need to give the tabled data to it. In Naive Bayes, the classifiers are built by us.
* Decision trees work better with lots of data compared to Naive Bayes.

4.

1. As a Data Engineer and a person from Fintech Startup, I wanted to build a model which will tell how many credit card scores a user can get.

The only data I am getting here is the SMS data of the particular user. Based on all the user's details I need to build a model with high accuracy and less error, which will predict the new user's score.

Firstly, I will analyze his/her patterns of using the credit card. The SMS details will contain the credit card limit, the amount he spent in a month, his due dates, whether he has cleared his dues in time or not, all the timestamps of the above details, etc. So with the help of text analytics and information retrieval, I will retrieve all the above mentions information.

I will start the process by cleaning the text SMS. Then use Named Entity Recognition and Relation Extraction methods to extract the source information for the SMS. If necessary I will use TFIDF to generate useful features from the text. I will make the below columns and keep the user data in it

* + - Customer Mobile Number
    - Monthly Limit
    - Time Stamp
    - Amount he spent
    - Due dates to clear the amount spent
    - Available balance
    - Cleared the due before deadline or not
    - Any Remarks

With all these data I can build a model with the help of Bayesian Classifier (since it can take any sizes of data and creates a robust model). With the help of the model I built, I can easily tell the credit card score of a new user with good accuracy.

1. As a Data Engineer and an employee of WhatsApp, I am asked to build a model which will give the popular person names from the customer base for some viral marketing.

The data I need for analysis is listed below

* + Person Name
  + Number of Active Groups
  + Number of groups as an admin
  + Messages with his name tag in those group
  + Number of messages he gets in a day
  + Status activity
  + Number of people who see his status
  + Number of replies for his status
  + New friends rate

The popularity will be based on the amount of time he spends with people and the amount of time people want to spend for him. This can be known with the help of groups and statuses. From groups, we can find his active participation in the chat and how people are replying to him (based on the name tags). From statuses, we can get his activity rate and based on the number of views and replies we can say how popular and likely he is.

With the help of all the above data on a person, I can build a model and give the popularity score for the person.