

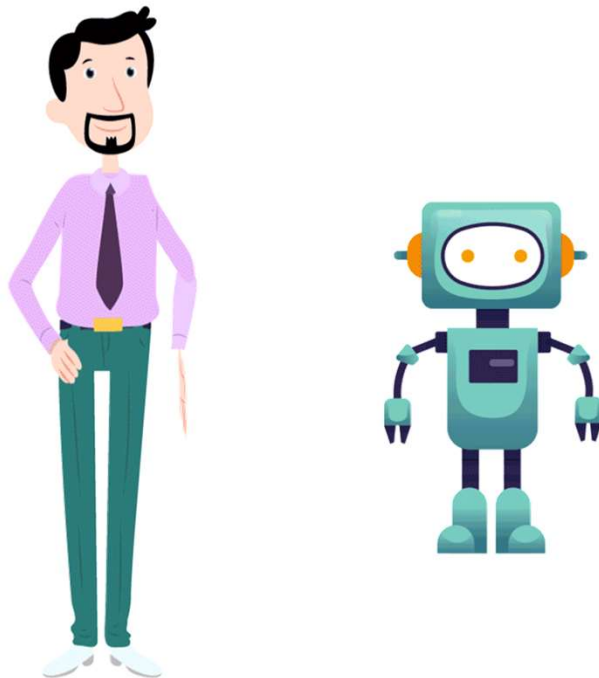
# Day 4 - Machine Learning Practical Applications

- 1. Workshop on applying machine learning to real-world problems.**
- 2. Group work on conceptualising a machine learning project**
- 3. Presentation of group ideas.**
- 4. Feedback session.**

# What is Machine Learning?



Machine Learning is the science of getting computer to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.



# Applying machine learning to real-world problems

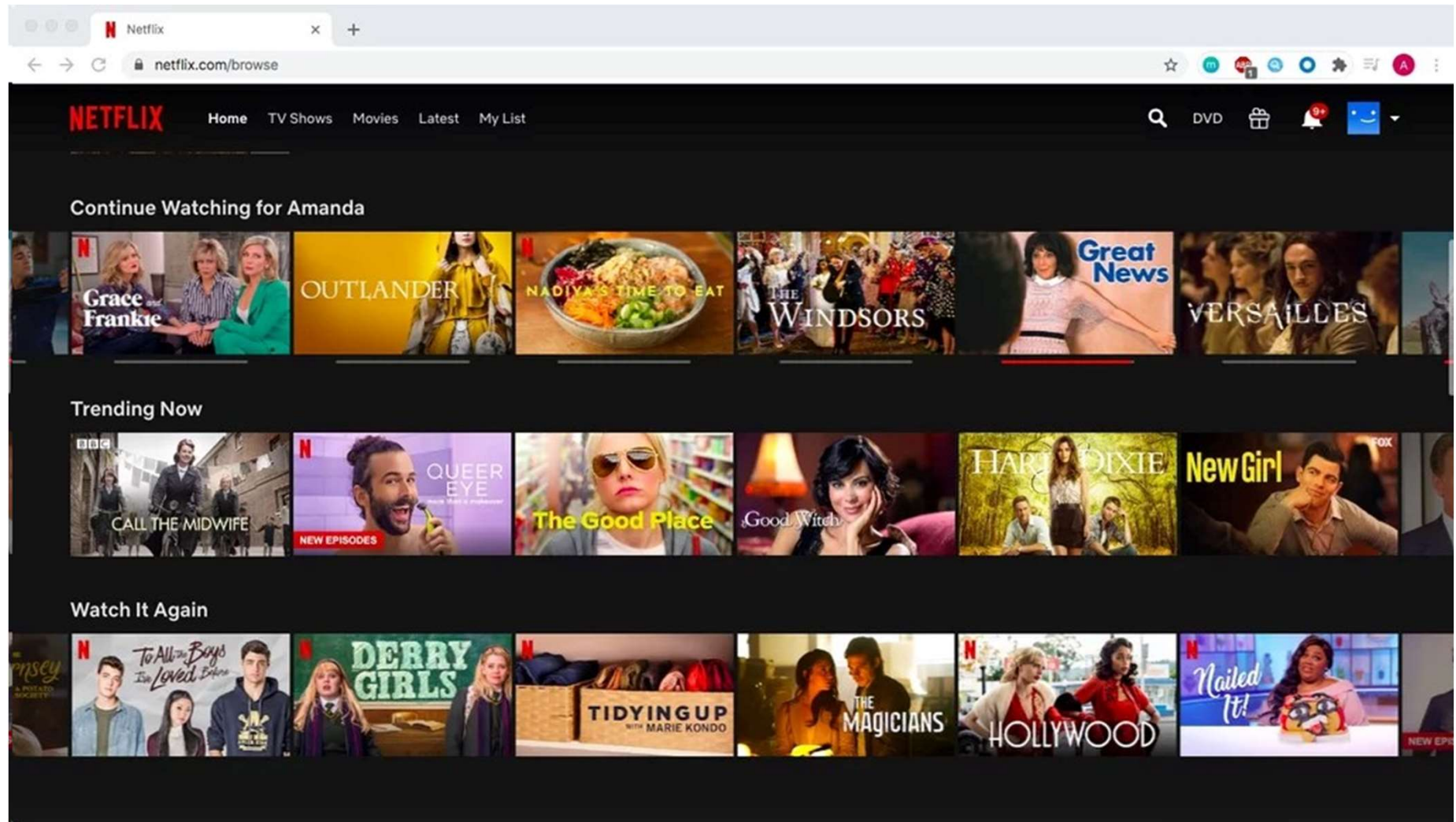


# Recommendation Engines



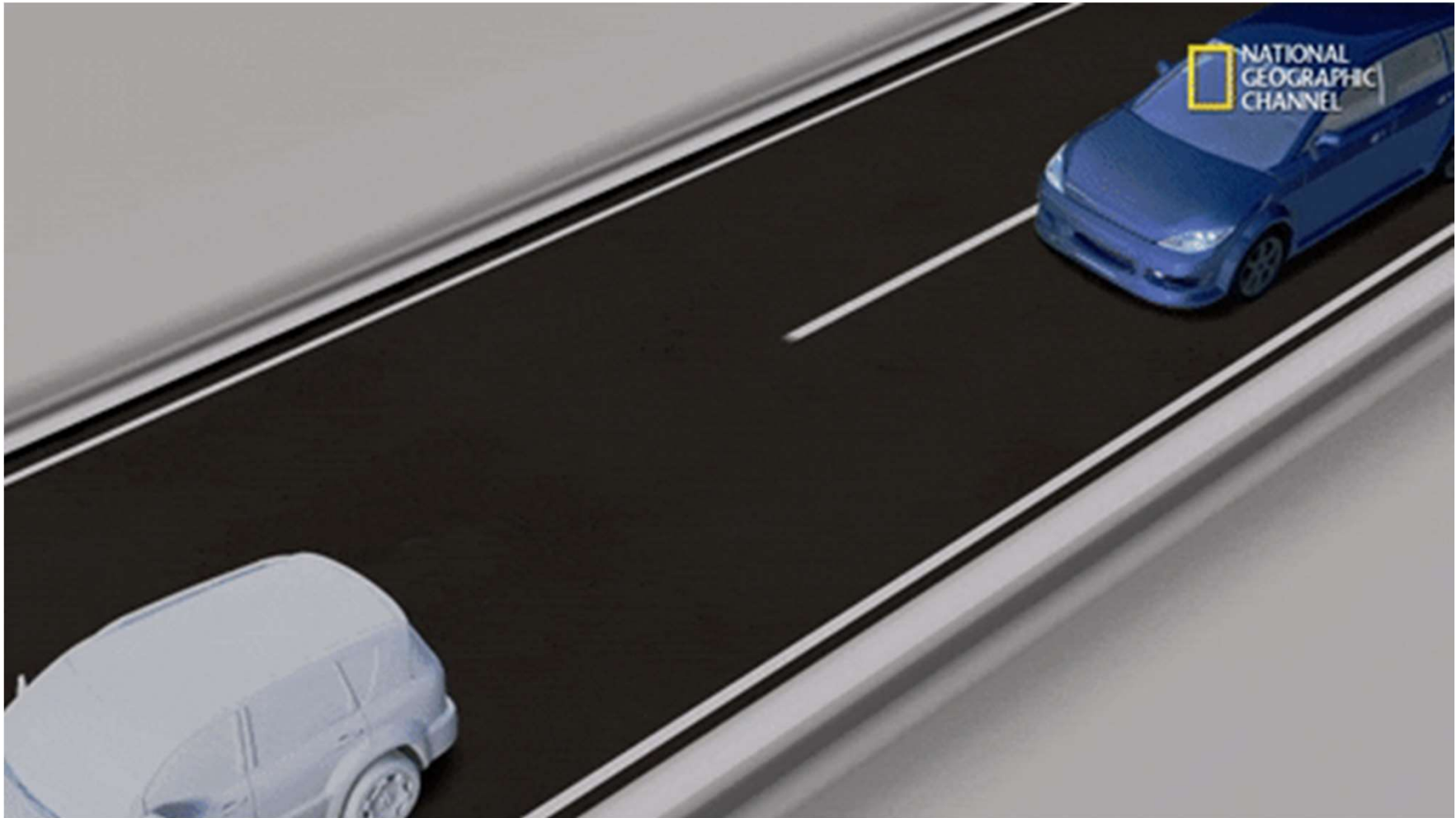
Eg: Netflix Viewing Suggestions

Application Area: Media + Entertainment + Shopping



# Self- Driving Cars

Eg: Tesla Cars use ML to understand surrounding  
Application Area: Automotive + Transportation

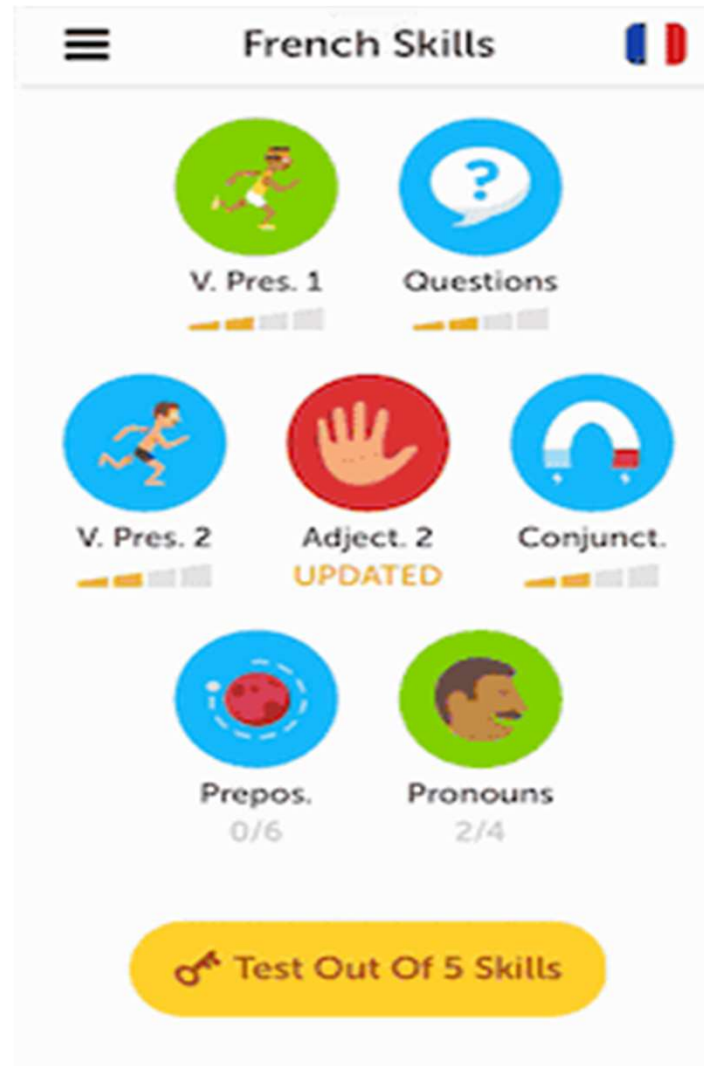


# Gamified Learning and Education



Eg: Duolingo's Mobile Application

Application Area: Learning Language Application

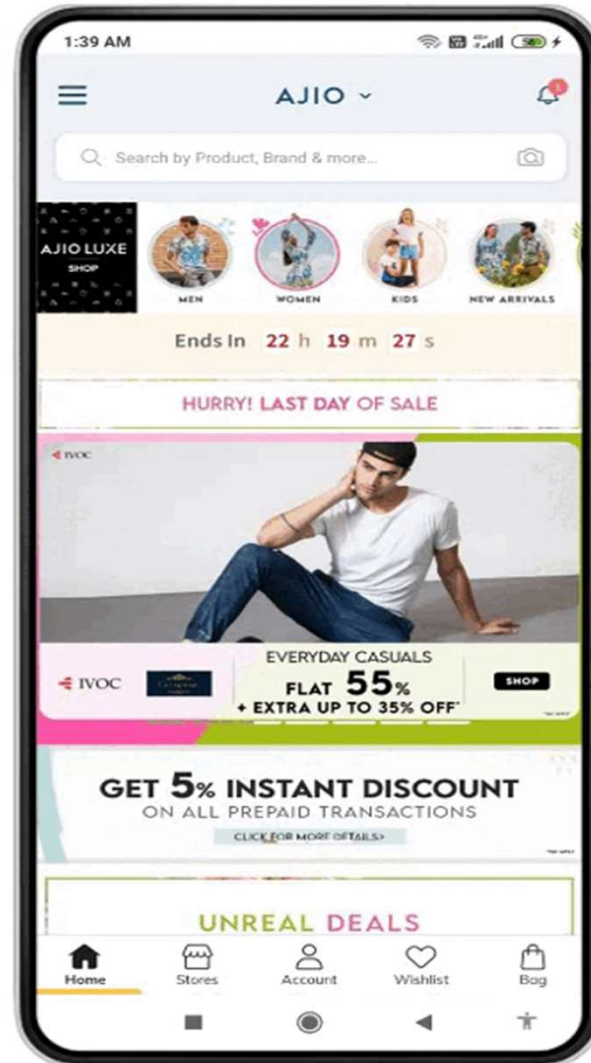


# E-Commerce Websites



Eg: Ajo

Application Area: Fashion E- Commerce

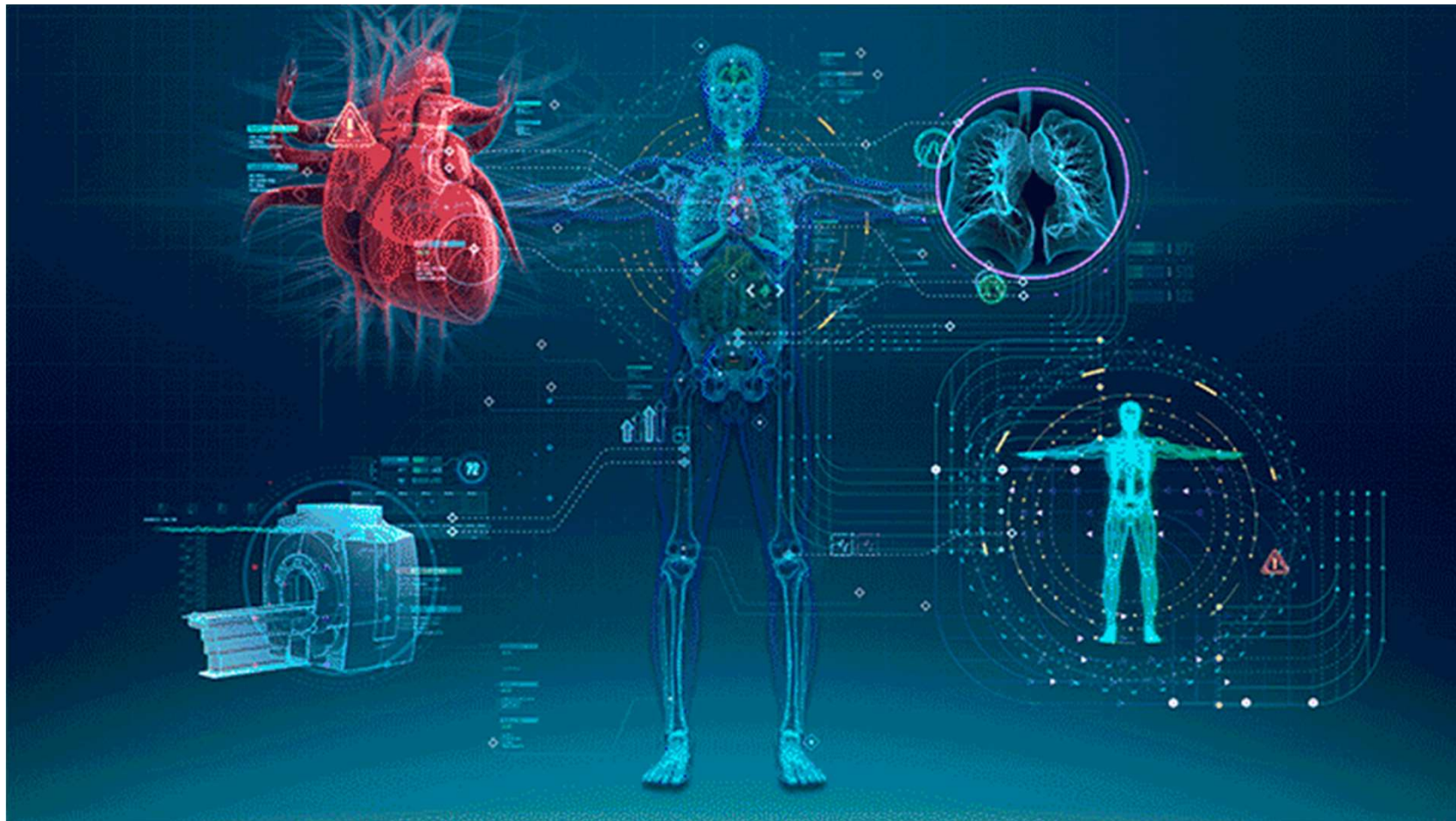




# Medical Diagnosis

Eg: Orderly Health

Application Area: HealthCare




# Getting Your Right Answers



Eg: Quora's Super-Specific Answer Ranking  
Application Area: Search

☰

 Quora Answers

✕

Writesonic's Quora Answer Generator


⚡ Your balance: 15,683 premium words

Question \* 0 / 100

Information 0 / 600  


health benefits of almond include lower blood sugar levels, reduced blood pressure and lower cholesterol levels.

Language  

 English


Quality type  

Premium


3  Outputs

Generate

Each time you click Create, we create 3 unique Quora Answers copies for you. [Check out some examples here](#)



Your copies created by artificial intelligence will appear here.



# Supervised algorithm : Support Vector Machine(SVM)

## Algorithm

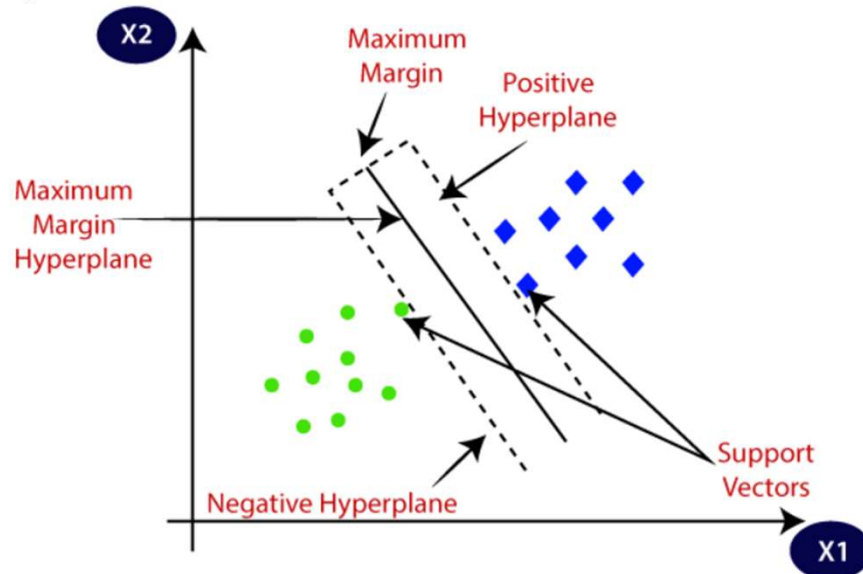


- Used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
- Goal is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
- SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

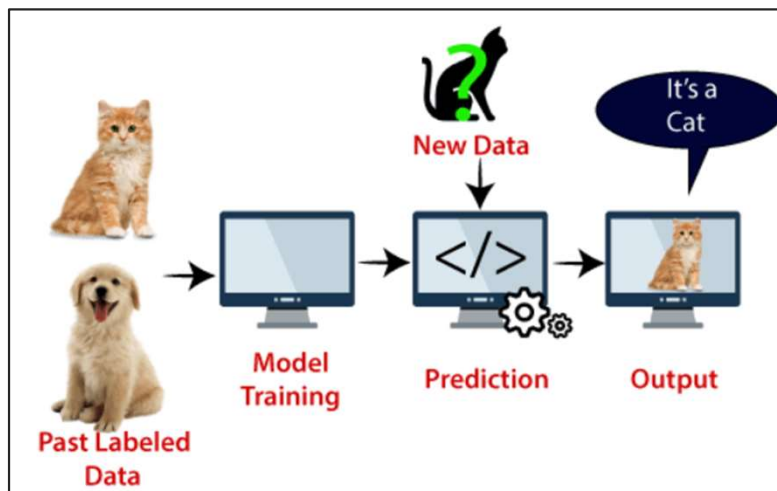
# Supervised algorithm : Support Vector Machine Algorithm



Two different categories that are classified using a decision boundary or hyperplane:



Eg:



SVM algorithm can be used for Face detection, image classification, text categorization, etc.

# Supervised algorithm : Support Vector Machine

## Algorithm



Types of SVM:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.



# Supervised algorithm : Support Vector Machine

## Algorithm



### Hyperplane:

- There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.
- The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.
- We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

### Support Vectors:

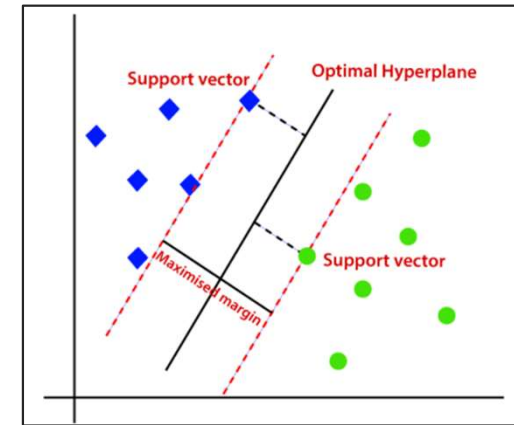
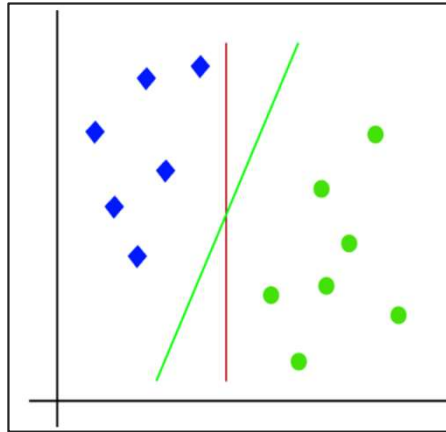
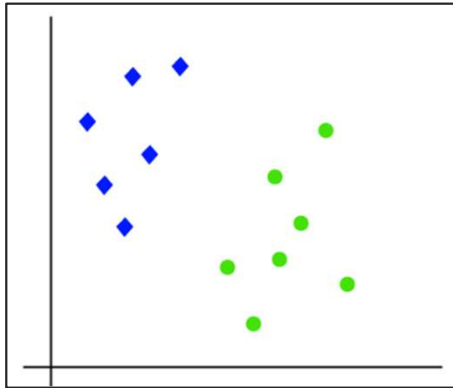
- The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

# Supervised algorithm : Support Vector Machine

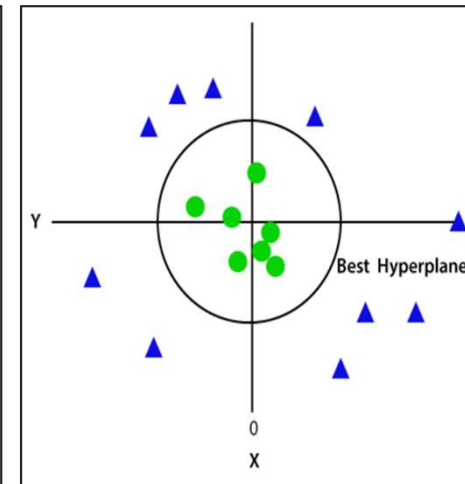
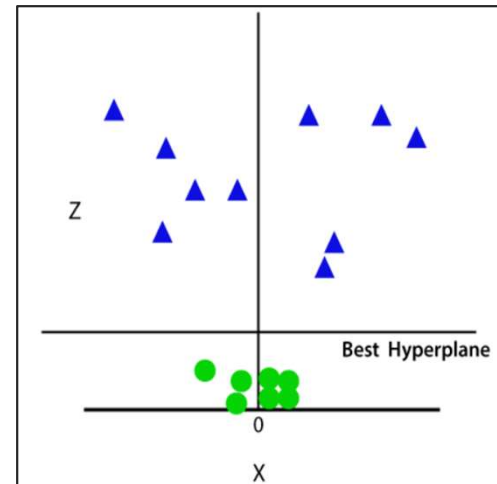
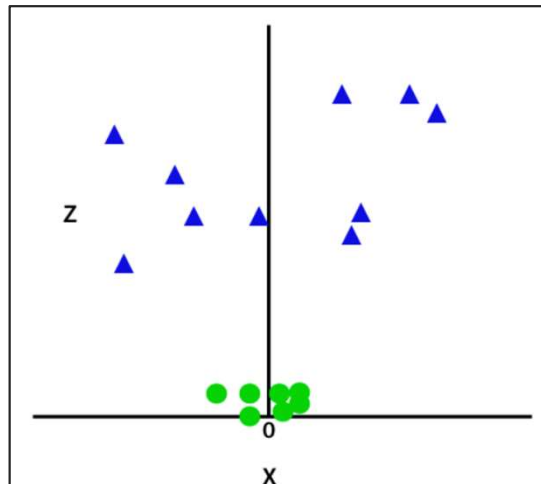
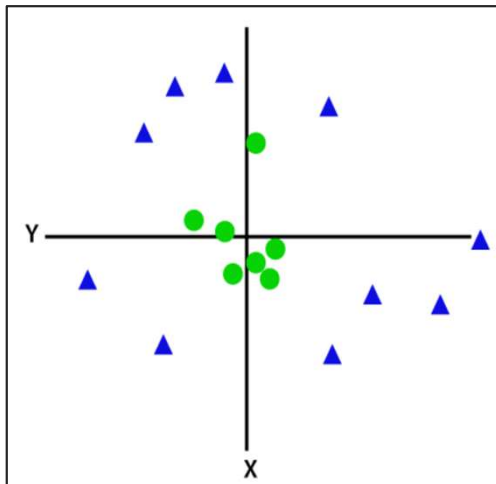
## Algorithm



### Linear SVM:



### Non-linear SVM:



## **SVM Basics:**

[https://colab.research.google.com/drive/1rhvbJxSaOCRsAzwJBqtW7vor5lxo\\_BPi?usp=sharing](https://colab.research.google.com/drive/1rhvbJxSaOCRsAzwJBqtW7vor5lxo_BPi?usp=sharing)

## **SVM basic project using iris data:**

<https://colab.research.google.com/drive/1AiLNJudkhMpglBladcZ4nP8Ean7-dcpb?usp=sharing>

# Supervised algorithm : Naive Bayes



- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.
- Bayes' theorem is to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

**P(A|B) is Posterior probability:** Probability of hypothesis A on the observed event B.

**P(B|A) is Likelihood probability:** Probability of the evidence given that the probability of a hypothesis is true.

## Working of Naïve Bayes' Classifier:

Eg: Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:

1. Convert the given dataset into frequency tables.
2. Generate Likelihood table by finding the probabilities of given features.
3. Now, use Bayes theorem to calculate the posterior probability.



## Types of Naïve Bayes Model:

- **Gaussian:** The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
- **Multinomial:** The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.  
The classifier uses the frequency of words for the predictors.
- **Bernoulli:** The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

# Supervised algorithm : Naive Bayes basic program



[https://colab.research.google.com/drive/1FBph3Bg-He2hlz3p1jE2\\_LMqRrnIBbMB?usp=sharing](https://colab.research.google.com/drive/1FBph3Bg-He2hlz3p1jE2_LMqRrnIBbMB?usp=sharing)

## Supervised algorithm : KNN

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

# Supervised algorithm : KNN

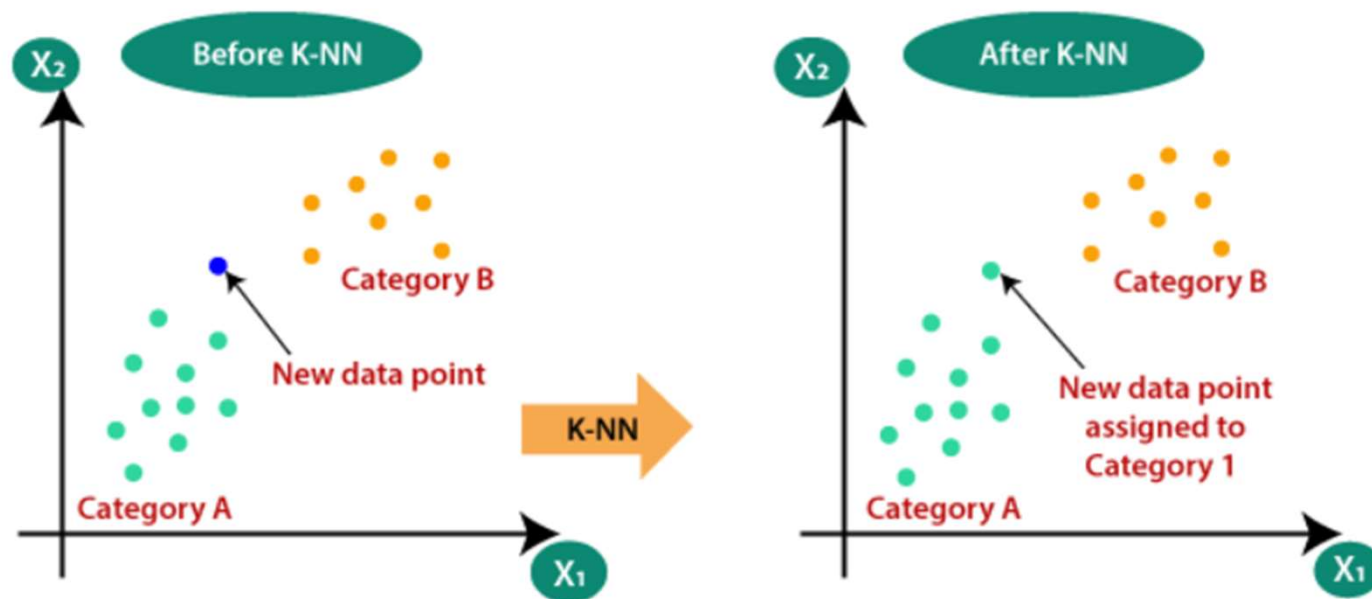
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.



# Supervised algorithm : KNN

## Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point  $x_1$ , so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:

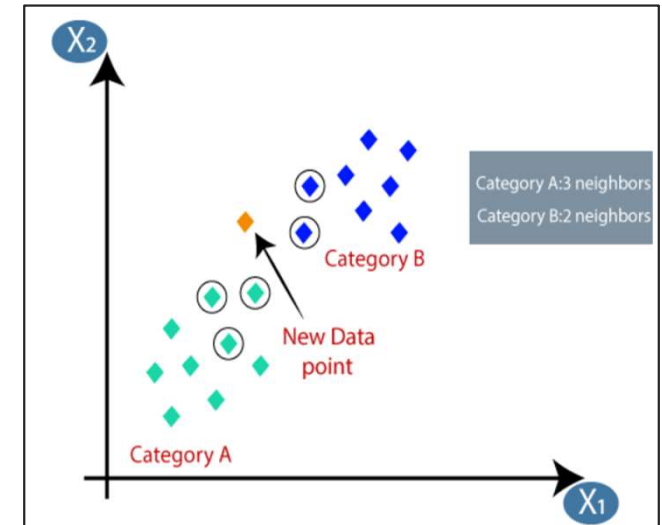
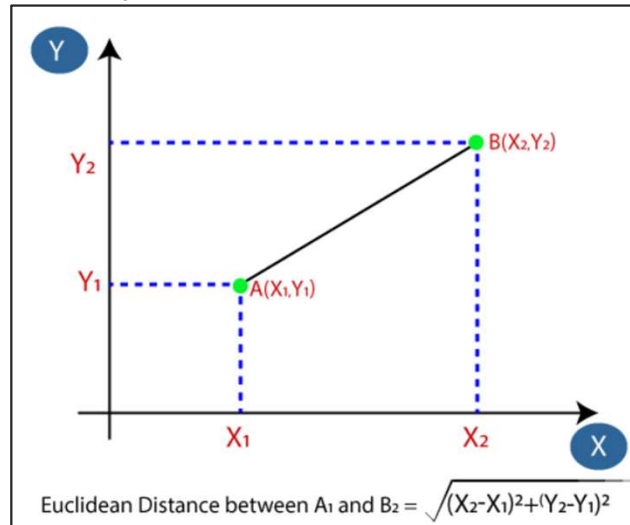
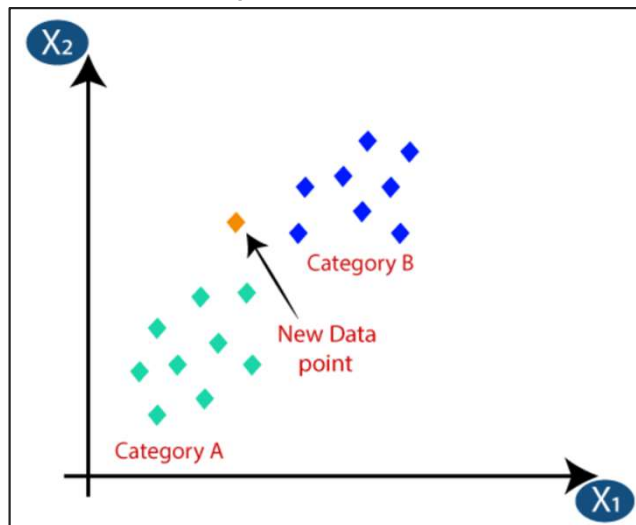




# Supervised algorithm : KNN

## How does K-NN work?

- Step-1: Select the number K of the neighbors
- Step-2: Calculate the Euclidean distance of K number of neighbors
- Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
- Step-4: Among these k neighbors, count the number of the data points in each category.
- Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
- Step-6: Our model is ready.



## How to select the value of K in the K-NN Algorithm?

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

### Basic program using KNN:

<https://colab.research.google.com/drive/1s5OEO56qGoGjAbNiuwpFGHWftFJJssdX?usp=sharing>

# Supervised Learning

## Projects:email\_spam\_detection



Mailing companies like Gmail, Outlook, and Yahoo are heavily investing in their technology to provide security to their users. One possible method is segregating spam emails automatically to avoid phishing attacks. This project demonstrates the capability of Machine learning in the cyber-security domain, where the ML model classifies emails into spam and non-spam categories based on internal textual content. It uses the **KNN classifier** for this task.

### Code:

<https://colab.research.google.com/drive/1BYZU0V94QYa3LRQpYGGWmGWLJwYXvEST?usp=sharing>

### Dataset:

[https://drive.google.com/drive/folders/15\\_CgVHH6bP\\_zDbh28lnjQ0yqG4Zb9TXc?usp=sharing](https://drive.google.com/drive/folders/15_CgVHH6bP_zDbh28lnjQ0yqG4Zb9TXc?usp=sharing)

# Unsupervised Learning Projects: personality prediction



There are mainly five types of human personalities: Openness, Neuroticism, Agreeableness, Extroversion, and Conscientiousness. This project groups persons into these five personalities based on the traits shown on their social media platforms. It uses the most famous **k-means clustering algorithm** in machine learning.

## Code:

<https://colab.research.google.com/drive/1PbrK1u9thpUb3HZMVuCCGgeaymMQGib1?usp=sharing>

## Dataset:

[https://drive.google.com/file/d/1B5plNmEFu81Wvf7GlZkdxynbDbGZdG\\_k/view?usp=sharing](https://drive.google.com/file/d/1B5plNmEFu81Wvf7GlZkdxynbDbGZdG_k/view?usp=sharing)

# Unsupervised Learning : PCA(Principal Component Analysis)



- It is an algebraic technique for converting a set of observations of possibly correlated variables into the set of values of liner uncorrelated variables.
- All principal components are chosen to describe most of the available variance in the variable, and all principal components are orthogonal to each other. In all the sets of the principal component first principal component will always have the maximum variance.

## Objectives of Principal Component Analysis?

- PCA is a nondependent method can be used for reducing attribute space from a larger number of variables of the set to a smaller number of factors.
- It is a dimension reducing technique but with no assurance whether the dimension would be interpretable.
- In PCA, the main job is selecting the subset of variables from a larger set, depending on which original variables will have the highest correlation with the principal amount.



# Unsupervised Learning : PCA(Principal Component Analysis)



**Principal Axis Method:** Principal Component Analysis searches for the linear combination of the variable for extracting maximum variance from the variables. Once the PCA is done with the process, it will move forward to another linear combination which will explain the maximum ratio of the remaining variance, which would lead to orthogonal factors of the sets. This method is used for analysing total variance in the variables of the set.

**Eigen Vector:** It is a nonzero vector that remains parallel after multiplying the matrix. Suppose 'V' is an eigen vector of dimension R of matrix K with dimension  $R \times R$ . If KV and V are parallel. Then the user has to solve  $KV = PV$  where both V and P are unknown for solving eigen vector and eigen value.

**Eigen Value:** It is also known as "characteristic roots" in PCA. This is used for measuring the variance in all the variables of the set, which is reported for by that factor. The proportion of eigen value is the ratio of descriptive importance of the factors concerning the variables. If the factor is low, then it subsidises less to the description of variables.

# PCA(Principal Component Analysis)

---



[https://colab.research.google.com/drive/1VN6bAgGRQ8j5JSdDHHhtf5e1\\_XJZE61E?usp=sharing](https://colab.research.google.com/drive/1VN6bAgGRQ8j5JSdDHHhtf5e1_XJZE61E?usp=sharing)

# Unsupervised Learning Projects: image compress



Image data consumes higher bandwidth, and hence there is a need to reduce the size of the images for transmission. Machine Learning can help with that as well. This project uses **PCA (Principal Component Analysis)**, a dimension reduction technique that compresses the image by 80% with the minimum loss in information.

## Code:

[https://colab.research.google.com/drive/1tzEsXs2LuKPyX1wgBfJuq\\_ZcEHOTtfEn?usp=sharing](https://colab.research.google.com/drive/1tzEsXs2LuKPyX1wgBfJuq_ZcEHOTtfEn?usp=sharing)

## Dataset:

<https://drive.google.com/drive/folders/1dMSHY3U8ltK-2eCjeuhJi64f6nD6Arir?usp=sharing>

# Quiz Time



<https://forms.gle/kemApxHxjwF4dnD9>

# Group work on conceptualising a machine learning project.



## 1. Linear Regression:

<https://colab.research.google.com/drive/11Y0CjJR4bmTBxtnVvSnEJ6HNxtijB9uY?usp=sharing>

## 2. Logistic Regression: <https://colab.research.google.com/drive/1WJ3kuM2D-d-Qpob9Mb5KMWgxhwdgebES?usp=sharing>

Dataset:

<https://drive.google.com/file/d/1EcH07uEBs9oad2xWhB6bmYHiuUfCdGgz/view?usp=sharing>

## 1. KNN:

[https://colab.research.google.com/drive/1NTcj\\_jozaNvMwZrFtaUaFQ\\_57fzNVubH?usp=sharing](https://colab.research.google.com/drive/1NTcj_jozaNvMwZrFtaUaFQ_57fzNVubH?usp=sharing)

## 2. Naive Bayes:

<https://colab.research.google.com/drive/168zYbuHiyd2YvX5kaEEjCO1iYrMQPSUC?usp=sharing>



# References

- I. <https://www.youtube.com/watch?v=4Rl8S7stN5A>
- II. <https://colab.research.google.com/drive/1izGP15oreJ9zFZ4qi8jz9jaK3nuhb8a2?usp=sharing>
- III. <https://www.youtube.com/watch?v=SrY0sTJchHE&t=402s>
- IV. <https://www.youtube.com/watch?v=4jv1pUrG0Zk&t=1878s>
- V. <https://github.com/enjoyalgorithms/Machine-learning-project-code/tree/main>
- VI. <https://medium.com/enjoy-algorithm/top-machine-learning-projects-with-python-code-c83d937050c9>
- VII. <https://www.javatpoint.com/>

# THANKS