### ABSTRACT

Data Mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends. In This report we will predict image based.

Python, an interpreted language which was developed by Guido van Rossum came into implementation in 1989. The language supports both objects oriented and procedure-oriented approach. Python is free, open-source software, and consequently anyone can write a library package to extend its functionality. Data science has been an early beneficiary of these extensions, particularly Pandas, the big daddy of them all.

OpenCV was started at Intel in 1999 by Gary Bradsky and the first release came out in 2000. Vadim Pisarevsky joined Gary Bradsky to manage Intel’s Russian software OpenCV team. In 2005, OpenCV was used on Stanley, the vehicle who won 2005 DARPA Grand Challenge. Later its active development continued under the support of Willow Garage, with Gary Bradsky and Vadim Pisarevsky leading the project. Right now, OpenCV supports a lot of algorithms related to Computer Vision and Machine Learning and it is expanding day-by-day.

Currently OpenCV supports a wide variety of programming languages like C++, Python, Java etc and is available on different platforms including Windows, Linux, OS X, Android, iOS etc. Also, interfaces based on CUDA and OpenCL are also under active development for high-speed GPU operations.

OpenCV-Python is the Python API of OpenCV. It combines the best qualities of OpenCV C++ API and Python language.

.

## Machine Learning

Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it makes a prediction on the basis of the model.

The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is trained again and again with an augmented training data set.

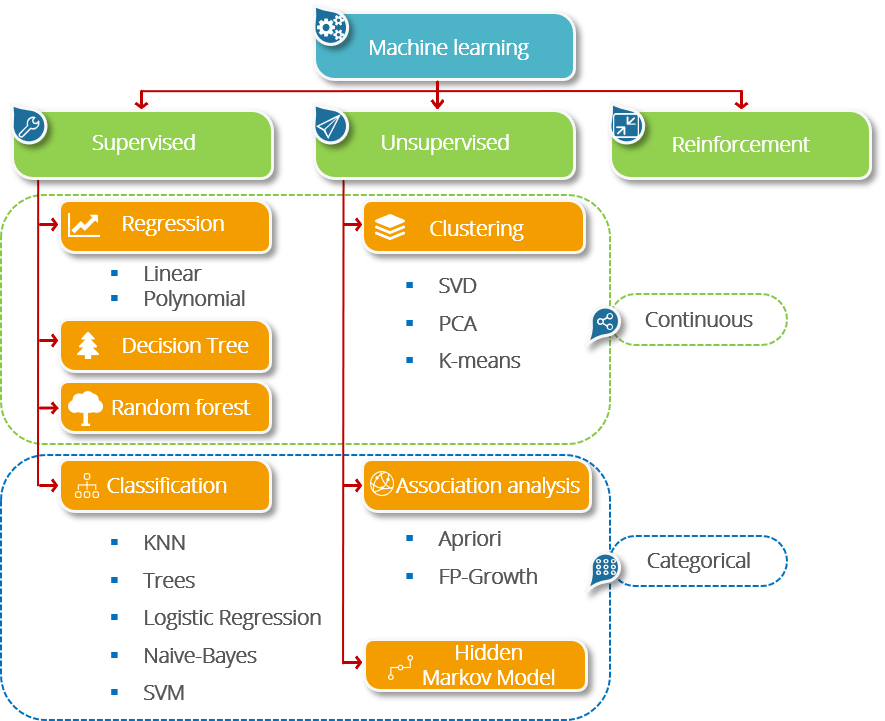
This is just a very high-level example as there are many factors and other steps involved.

## Types of Machine Learning

## Supervised Learning: Supervised Learning is the one, where you can consider the learning is guided by a teacher. We have a dataset which acts as a teacher and its role is to train the model or the machine. Once the model gets trained it can start making a prediction or decision when new data is given to it.

## Unsupervised Learning: The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters in it. What it cannot do is add labels to the cluster, like it cannot say this a group of apples or mangoes, but it will separate all the apples from mangoes. Suppose we presented images of apples, bananas and mangoes to the model, so what it does, based on some patterns and relationships it creates clusters and divides the dataset into those clusters. Now if a new data is fed to the model, it adds it to one of the created clusters.

## Reinforcement Learning: It is the ability of an agent to interact with the environment and find out what is the best outcome. It follows the concept of hit and trial method. The agent is rewarded or penalized with a point for a correct or a wrong answer, and on the basis of the positive reward points gained the model trains itself. And again, once trained it gets ready to predict the new data presented to it.



## ****Classification****

We use the training dataset to get better boundary conditions which could be used to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class. The whole process is known as classification.

**Target class examples:**

* Analysis of the customer data to predict whether he will buy computer accessories **(Target class: Yes or No)**
* Classifying fruits from features like color, taste, size, weight **(Target classes: Apple, Orange, Cherry, Banana)**
* Gender classification from hair length **(Target classes: Male or Female)**

Let’s understand the concept of classification algorithms with gender classification using hair length (by no means am I trying to stereotype by gender, this is only an example). To classify gender **(target class)** using hair length as feature parameter we could train a model using any classification algorithms to come up with some set of boundary conditions which can be used to differentiate the male and female genders using hair length as the training feature. In gender classification case the boundary condition could the proper hair length value. Suppose the **differentiated boundary** hair length value is 15.0 cm then we can say that if hair length is **less than 15.0 cm** then gender could be male or else female.

## ****Basic Terminology in Classification Algorithms****

* **Classifier:** An algorithm that maps the input data to a specific category.
* **Classification model:**A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.
* **Feature:** A feature is an individual measurable property of a phenomenon being observed.
* **Multi-label classification:**Classification task where each sample is mapped to a set of target labels (more than one class). **E.g.: A news article can be about sports, a person, and location at the same time.**

## ****Applications of Classification Algorithms****

* Email spam classification
* Bank customers loan pay willingness prediction.
* Cancer tumor cells identification.
* Sentiment analysis
* Drugs classification
* Facial key points detection
* Pedestrians detection in an automotive car driving

**Project SYNOPSIS**

**Image Identification and Classification Using Decision Tree, Random Forest etc**

Yes, I know you all have worked with Jupyter notebook definitely. This is what I will be showing you the analysis and Comparison of Image Classification using Random Forest Classifier, Decision tree Classifier, Naïve Bayes Classifier and Support Vector Machine using [Python](https://www.edureka.co/blog/python-basics/).  
Before moving on, let’s have a quick look at basic steps through which I proceeded this in Python:

1. [**Installing Libraries**](https://www.edureka.co/blog/snake-game-with-pygame/#install)
2. **Importing Libraries**
3. **Importing Datasets using Libraries**
4. **Extracting Labels and Features**
5. **Training the Models**
6. **Comparing Models**
7. **Predicting Images**

## ****Installing Libraries:****

To do that, you can simply use the following command:

***conda install opencv***

***pip install os***

***pip install numpy***

***pip install -U matplotlib***

Once that is done, then Open Jupyter Notebook Create a new python

And Import the Libraries:

***import opencv as cv2***

***import os as os***

***import numpy as np***

***import matplotlib.pyplot as plt***

## Libraries Used for Data Plotting the Images

#### [**Matplotlib**](https://github.com/matplotlib/matplotlib): This is a standard data science library that helps to generate data visualizations such as two-dimensional diagrams and graphs (histograms, scatterplots, non-Cartesian coordinates graphs). Matplotlib is one of those plotting libraries that are really useful in data science projects — it provides an object-oriented API for embedding plots into applications. It's thanks to this library that Python can compete with scientific tools like MatLab or Mathematica. However, developers need to write more code than usual while using this library for generating advanced visualizations. Note that popular plotting libraries work seamlessly with Matplotlib.

## Libraries/Modules Used for Data Processing and Modeling

#### **NumPy:** NumPy (Numerical Python) is a perfect tool for scientific computing and performing basic and advanced array operations. The library offers many handy features performing operations on n-arrays and matrices in Python. It helps to process arrays that store values of the same data type and makes performing math operations on arrays (and their vectorization) easier. In fact, the vectorization of mathematical operations on the NumPy array type increases performance and accelerates the execution time.

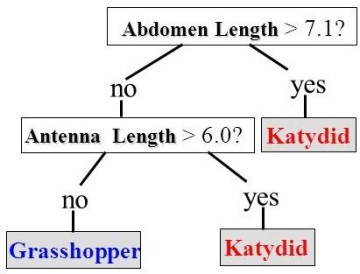
#### **SciPy:** This useful library includes modules for linear algebra, integration, optimization, and statistics. Its main functionality was built upon NumPy, so its arrays make use of this library. SciPy works great for all kinds of scientific programming projects (science, mathematics, and engineering). It offers efficient numerical routines such as numerical optimization, integration, and others in submodules. The extensive documentation makes working with this library really easy.

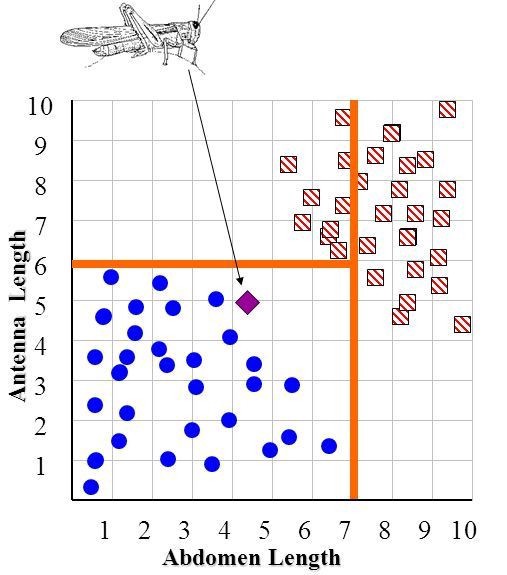
#### **OS:** The OS module in python provides functions for interacting with the operating system. OS, comes under Python’s standard utility modules. This module provides a portable way of using operating system dependent functionality. The \*os\* and \*os.path\* modules include many functions to interact with the file system.

## Models Used for Classification and Identification:

## ****Decision Tree Classifier****

* Using the decision algorithm, we start at the tree root and split the data on the feature that results in the **largest information gain (IG)** (reduction in uncertainty towards the final decision).
* In an iterative process, we can then repeat this splitting procedure at each child node **until the leaves are pure**. This means that the samples at each leaf node all belong to the same class.
* In practice, we may set a **limit on the depth of the tree to prevent overfitting**. We compromise on purity here somewhat as the final leaves may still have some impurity.





Classifying whether an insect is a Grasshopper or a Katydid based on Antenna Length and Abdomen Length.

## ****Advantages of Classification with Decision Trees:****

1. Inexpensive to construct.
2. Extremely fast at classifying unknown records.
3. Easy to interpret for small-sized trees
4. Accuracy comparable to other classification techniques for many simple data sets.
5. Excludes unimportant features.

## ****Disadvantages of Classification with Decision Trees:****

1. Easy to overfit.
2. Decision Boundary restricted to being parallel to attribute axes.
3. Decision tree models are often biased toward splits on features having a large number of levels.
4. Small changes in the training data can result in large changes to decision logic.

## ****Applications of Decision trees in real life:****

1. Biomedical Engineering (decision trees for identifying features to be used in implantable devices).
2. Financial analysis (Customer Satisfaction with a product or service).
3. Astronomy (classify galaxies).
4. System Control.
5. Manufacturing and Production (Quality control, Semiconductor manufacturing, etc).
6. Medicines (diagnosis, cardiology, psychiatry).
7. Physics (Particle detection).

# The Random Forest Classifier

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction (see figure below).



Visualization of a Random Forest Model Making a Prediction

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is:

**A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.**

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. **The reason for this wonderful effect is that the trees protect each other from their individual errors** (as long as they don’t constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So, the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

## Support Vector Machine

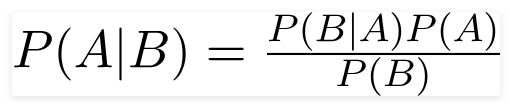
The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.



# Principle of Naive Bayes Classifier:

A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.

# Bayes Theorem:



Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. Here, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

# Types of Naive Bayes Classifier:

## Multinomial Naive Bayes: This is mostly used for document classification problem, i.e. whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

## Bernoulli Naive Bayes: This is similar to the multinomial naive bayes but the predictors are Boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

## Gaussian Naive Bayes: When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

# Conclusion:

Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, recommendation systems etc. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real-life cases, the predictors are dependent, this hinders the performance of the classifier.

**BIBLIOGRAPHY**

* <https://www.kaggle.com/prasunroy/natural-images>
* <https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>
* Udemy Videos
* Working with Panda Library