**CHAPTER 1**

**INTRODUCTION ABOUT PROJECT**

**Introduction :**

There are various areas in which data mining can be used in financial sectors like customer segmentation and profitability, high risk loan applicants, predicting payment default, marketing, credit analysis, ranking investments, fraudulent transactions, optimizing stock portfolios, cash management and forecasting operations, most profitable Credit Card Customers and Cross Selling. There are many different types of loans you have to take into account when you’re looking to borrow money and it’s important to know your options. Loan categorization refers to the process of evaluation loan collections and assigning loans to groups or grade based on the perceived danger and other related loans properties. The process of continual review and classification of loans enables monitoring the quality of the loan portfolios and to take action to counter fall in the credit quality of the portfolios. It is required for banks to use more complicated internal classification schemes than the more standardized schemes that bank managers need for reporting reasons and that are intended to make easy observing and interbank evaluation. There are many types of loans such as: Open-ended loans are loans that you can have a loan of more and more. Credit cards and lines of credit are the famous types of open-ended loans. You have a credit limit that you can buy with both of these two types of loans. In any time you can purchase automatically your available credit will decreases. since you make expenditure, you're on hand increases permitting you to use the credit more and more. Closed-ended loans, this type of loans cannot be on loan once they’ve been repaid. while you make expenditure on closed ended loans, the balance of the loan became downward. though, you don’t have any existing credit you can employ on closed-ended loans. As an option, if you want to lend more money, you’d have to make application for other loan. widespread types of closed-ended loans involve auto loans, mortgage loans, and student loans.

Secured loans are loans that rely on an asset. In the state of loan failure to pay, the lender can possess the asset and make use of it to cover up the loan. High wellbeing rates for secured loans may be lower than those for unsecured loans. The asset may need to be evaluated before you can have a loan of a secured loan. Unsecured loans lends may be more complicated to get and have higher concern rates. Unsecured loans rely just on your credit history and your revenue to meet the criteria for the loan. If you failed to return back unsecured loan, the lender has to wear out collection alternatives involving debt collector and claim to recover the loan. Conventional loans or mortgage loans are the loans that aren’t insured by a government organization, country Housing Service, or the Veterans management. Conventional loans may be conforming, meaning they tag on the rule set. There are many risks related to bank loans, for the bank and for those who get the loans. The analysis of risk in bank loans need understanding what is the meaning of risk. Risk is denotes to the probability of certain outcomes or the uncertainty of them especially an existing negative threat for trying to achieve a current monetary operation. Risk in bank loans involve: credit risk, the risk that the loan won't be return back on time or at all; liquidity risk, the risk that too many deposits will be withdrawn too quickly, leaving the bank short on immediate cash; and interest rate risk, the risk that the interest rates priced on bank loans will be too low to earn the bank adequate money.

There are two most important goals for data mining prediction and description. Prediction involves using some variables in data set to predict unknown values of other variables and Description concentrates on finding patterns describing the data that can be interpreted by human. Data mining is the process of extracting hidden pattern from large amount of data that used to take a write decisions. The derived knowledge must be new, not obvious, relevant and can be applied in the field where this knowledge has been obtained. It is also the process of extracting useful information from raw data. Data Mining is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated data sets. In present era, Data Mining is becoming popular in banking field because there is a call for efficient analytical methodology for detecting unknown and useful information in banks data. Skills and knowledge are important requirement for achieving Data Mining task since the success and failure of Data Mining is greatly reliant on the person who managing the process due to unavailability of standard framework.

Data mining techniques aid to distinguish between borrowers who pay back loans at the appointed time from those who don't. It also helps to expect when the borrower is at default, whether providing loan to a particular customer will result in bad loans. All processes related to banking sector could be analyzed using data mining to detect the customer. It also helps to analyze whether the customer will make prompt or delay payment if the credit cards are sold to them.

**Existing System:**

Many researches have been conducted based on data mining in the field of financial and banking sector. This section presents briefly some of these techniques which are used in loans risk management and focused on specifying the data mining applications usefulness, these applications are using several data mining techniques such as decision trees and Radial Basis Neural Networks. This study came with in which way to apply these applications in a credit-risk assessment field. McLeod presents Neural networks properties and their fitness for the credit granting process.

**Proposed system:**

We mean by loan evaluation process, the sequence of steps that taken to take diced about granting a loan to the customer or not. When the customer apply for a loan granting application, the bank officer must investigate about what called 5 C’s which are Character (or Credit History), Cash Flow (or Capacity), Collateral, Capitalization and Conditions. It is helpful for evaluation loan application and it regarded as a helpful framework for estimate the credit risk related to a probable creditor.

**Working:**

The process of classification crowd the data set into groups of classes according to their dissimilarity. There are several classification algorithm or classifiers like Naïve Bayes Classifier, Neural Network Classifier, decision Tree Classifier. There are several algorithms in each of this technique which used to produce a model to predict the class of unknown class tables. The major goal of this algorithm is the provision by a model for predicting the class of unknown records.

HARDWARE REQUIREMENTS:

* System : Intel I3 2.4 GHz. Or Advanced
* Hard Disk : 200 GB +
* Monitor : 14’ Colour Monitor. Or Advanced
* Mouse : Optical Mouse.
* Ram : 4GB +
* GRAPHICS CARD : 2GB +

SOFTWARE REQUIREMENTS:

* Operating system : Windows / linux
* Coding Language : Python /anaconda
* Front-End : Python.

**OVERVIEW:**

To ease comparison of the applied machine learning method to other techniques, as well as to allow validation by other researchers, secondary data was used in this project. Two publicly available datasets from the Machine Learning Repository. Basically, industrial risk measures the health and future potential of the industry, management risk measures the organizational structure and managers’ capabilities, financial flexibility is a measure of the company’s cashflow, credibility is a measure of the company reputation and credit scores, competitiveness is a measure of the company’s market position and competitive advantages, and operating risk is a measure of its efficiency in production. As presented in the original paper, the authors’ proposed genetic algorithm method resulted in a classification accuracy of 94%. They also compared it to two other data-mining techniques, namely induction learning and neural networks. Most of the quantitative attributes are financial ratios and econometric indicators as found in majority of existing literature. A complete list of those attributes can be found. The methods for analysing these two datasets are similar and will be explained below. Note that the difference in the quality of results between Dataset 1 and Dataset 2 can be attributed to factors such as different geographic location, different dataset size, different features, and different quality of data.

**Pre-Processing** :

The range of possible values for various input features can vary drastically. For example, gross margin defined by Equation 9 will always be less than one (i.e. below 100%) due to normalization, whereas some financial measures like working capital can theoretically take on any real value (i.e. negative infinity to positive infinity).

One possible solution is to standardize each feature such that all have zero mean and unit variance. To achieve this, the average can simply be subtracted from every training sample and then divided by its standard deviation (Equation 10); however, if the variance of a particular feature is very small (i.e. close to zero), then this division may have numerical issues. An alternative is to simply scale the data to be between a minimum and maximum value of choice: for instance, zero and one.

Dimensionality Reduction Classification with only one, two, or even three dimensional features is generally intuitive, because the classification boundary and training data can be visualized. Unfortunately, financial distress information resides in a higher dimensional feature space. This makes sense because even though the dimensions of many machine learning problems are high, the interesting characteristics typically lie in a lower dimensional manifold. For example, researchers have proposed adding macroeconomic measures to the existing financial ratios for financial distress prediction. As a company’s operation is inevitably affected by the macro environment, part of its effect is already reflected in the company’s financial performance; therefore, when macroeconomic features are included, the information they provide is not totally independent from the other features. Another example is the similarities of some financial ratios. After all, besides reducing the complexity of the following machine learning algorithms by working in a lower dimensional subspace, one of the objectives of this projection should be to maximize the separation between the different classes. One of the methods to achieve this is the Linear Discriminate Analysis (LDA). Class labels are required, thus LDA can be considered a supervised version of PCA. Both PCA and LDA perform a linear projection. When the projection is nonlinear we can estimate the geodesic distance along the manifold and apply multi-dimensional scaling.

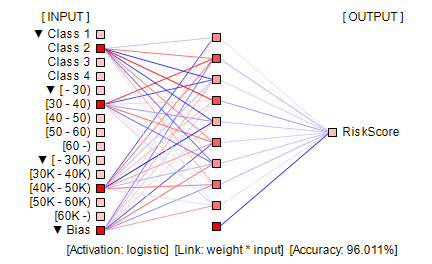


Fig : Process for prediction.

**Learning from Data and Model Selection :**

Different machine learning algorithms have different methods of constructing a model of the real world using the provided data. For instance, a K-D tree learns from the data by partitioning the data and forming a binary tree structure for fast query, while a logistic regression learns by estimating some weight parameters in an optimization framework where some likelihood function is being maximized. In general, this learning from data can either be parametric or non-parametric. Parametric methods will learn some unknown parameters of a model and forget about the data (e.g. logistic regression), while non-parametric methods like GP will have to store all the training data. This can be time-consuming and subjective. An automatic method is to do cross-validation either using a validation dataset or the k-fold scheme. Unless there is an abundance of datasets, doing k-fold cross-validation is typically a better choice because the same data used for training can be used for model selection. In k-fold cross-validation the training dataset is randomly clustered into ‘k’ groups. Group one is used as the validation dataset while all the remaining data is used for training. This process is then repeated with group 2 acting as the validation dataset and all other data is used for training. K-fold cross-validation terminates when all ‘k’ groups had a chance to play the role of a validation dataset. At this point, the model with the highest score or the average of all k-folds can be used to select the best model. For example, this strategy can be used to tune the softness of the margin in SVM, However, not all machine learning methods are probabilistic, such as SVM and K-D tree. Therefore, to have a unified framework for model tuning, cross-validation will be used when comparing different models.

**Accuracy Assessment:**

The machine learning models can have very different characteristics and behaviour, making it difficult to judge which model is performing better. Therefore, a consistent set of tools applicable for assessing the performance of any machine learning model is important. Some of the most popular quality control measures in machine learning are defined below. Care must be taken to ensure this testing dataset was never exposed in any of the pre-processing or training stages in the machine learning pipeline. Also, the testing set should have the same probability distribution as the training set. This can be achieved by randomly choosing a subset of points from the original dataset (e.g. 70% use for training, and 30% used for testing). This single scalar value indicates how well a machine learning algorithm can label companies in a non-recoverable financial crisis as bankrupt and financially healthy companies as not bankrupt.

𝐚𝒄𝒄𝒖𝒓𝒂𝒄𝒚 = 𝟏 𝑵 ∑(𝒚𝒊 𝒑𝒓𝒆𝒅 − 𝒚𝒊 𝒕𝒓𝒖𝒆) 𝑵 𝒊 where 𝜹(𝒙 = 𝟎) = 𝟏 and 𝜹(𝒙 ≠ 𝟎) = 𝟎

Besides the accuracy score, another set of metrics commonly used for quality control in machine learning is the precision, recall, and F1 score. Precision is a measure of how well the algorithm can find true positives. In the case of this thesis, it can be translated into how well the model can predict a company as bankrupt when it is actually going bankrupt. For example, a precision of 100% means that corporations flagged as bankrupt will surely experience bankruptcy in the future with great certainty. Another closely related concept to precision is recall, defined in Measure of how reliable can the classifier identify all true positive samples. For instance, a recall of 50% would indicate that half the bankruptcy candidates have been found while the other half of bankruptcy facing firms were missed by the classifier. Ideally, a good classifier should maximize both precision and recall, unfortunately in reality precision and recall is often a trade-off an econometrics expert would have to make while training the model.

**CHAPTER 2**

**LITERATURE SURVEY**

There are numerous fields where data mining can be used in financial industries such as customer segmentation and profitability, applicants for high-risk loans, predicting default payments, promotions, collateral monitoring, rating assets, irregular sales, managing stock holdings, cash management and estimating activities, most profitable clients with credit cards and cross-selling [2]. Nowadays, the banking sector has become highly competitive. To succeed and expand in an evolving business environment, banks should embrace new and up-to-date technology. It is often regarded as a medium for cost savings and efficient contact with consumers [3]. The bank can cut costs by segmenting clients between bad customers and good customers until it's too late. Through observing transaction trends, banks can monitor fraud transactions until their productivity is impacted [4]. Data mining plays a vital role in the management of transaction data and customer profile in banking. From that, a • Princess May G. Subia is currently pursuing a master degree program in Information Technology in Isabela State University Cauayan City Campus, Philippines. E-mail: princessmaysubia@gmail.com • Angelo C. Galapon, Ph.D., DIT Assist. Prof. of Master degree program in Information Technology in Isabela State University Cauayan City Campus, Philippines. E-mail: angelogalapon110583@gmail.com INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 06, JUNE 2020 ISSN 2277-8616 223 IJSTR©2020 www.ijstr.org user can make an accurate decision using data mining techniques [5]. A data mining technique produces a lot of patterns and rules, typically. Among those rules created should be selected beneficial and interesting [6]. One great example is the credit scoring which aims to divide the applicants into two categories-good credit applicants and poor credit applicants. The former class has excellent chances of repaying financial commitments, and the others have high defaulting risks [7]. At the same time, the return engagement with the right judgment about credit risk is high. It would also be particularly beneficial to see some progress in creating a credible distinction between those who are willing to repay the loan and those who are not [8]. The general principle of credit appraisal is to associate a customer's attributes or qualities with other past borrowers, whose debts they have already paid off. And credit rating is also used to evaluate a list of previous clients and differentiate current and potential credit clients [9]. People who cannot pay back the amount of their loan are considered defaulters. Unforeseen economic conditions may increase the number of defaulters for such years, which in turn may increase the financial institutions' losses. This has a negative impact not only on the institutes but also on the credit record of the client and potential financial stability. Given the increase in default rates, keeping track of the vulnerability of default is critical for both the good of banks and customers [10].

**CHAPTER 3**

**AI & ML**

**3.1 ARTIFICIAL INTELLIGENCE**

Artificial Intelligence is the future of the world. It is expanding rapidly in every industry vertical. Hence, there is a bright future in Artificial Intelligence. AI is a technique that enables machines to mimic human behavior. Artificial Intelligence is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making and translation between languages. AI is the simulation of human intelligence done by machines programmed by us. The machines need to learn how to reason and do some self-correction as needed along the way. Artificial Intelligence is accomplished by studying how human brain thinks, learns, decide, and work while trying to solve a problem, and then using the outcomes of this study as a basis of developing intelligent software and systems. So, Let’s continue this Artificial Intelligence Tutorial and understand it’s importance. Today, a few applications of artificial intelligence seem to bring us closer to the future. The most convincing pieces of evidence are self-driving cars, Google Translate, and Sophia (humanoid robots). In this Artificial Intelligence tutorial, we shall be covering [Machine Learning](https://intellipaat.com/blog/what-is-machine-learning/), [Deep Learning](https://intellipaat.com/blog/tutorial/machine-learning-tutorial/introduction-deep-learning/), [neural networks](https://intellipaat.com/blog/tutorial/machine-learning-tutorial/neural-network-tutorial/), real-life applications of Artificial Intelligence, Python and various packages available in it, TensorFlow, Keras, multilayer perceptron, convolution neural networks, recurrent neural networks, long short-term memory, opencv, and much more.

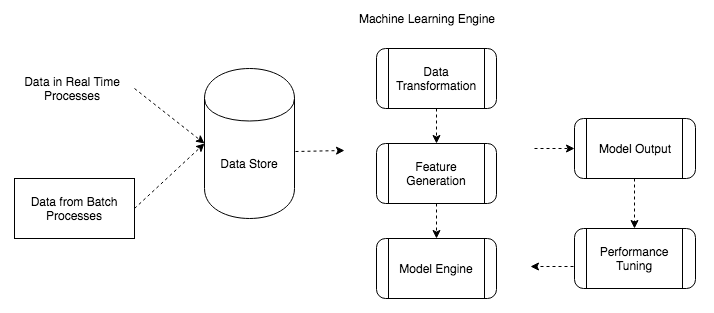


Fig:- Block diagram flow of architecture for [Machine learning systems](https://www.educba.com/machine-learning-system/)

**3.2 WHAT ARE THE GOALS OF ARTIFICIAL INTELLIGENCE**

Creativity and ideas never end as they are limitless. Likewise, there are a lot more things to create, improve, implement, and invent in the field of Artificial Intelligence. Evidently, AI is far from reaching its saturation level of creating new things.

In short, here are the goals of Artificial Intelligence:

Create machines that can replicate human beings

Improve machine efficiency and accuracy

Develop tools to help people solve real-world problems, e.g., robotics for people with disabilities, auto-driving cars to avoid accidents caused by human error, etc.

**3.3 APPLICATIONS OF ARTIFICIAL INTELLIGENCE**

Artificial intelligence machines have the ability to make decisions and when exposed to large amounts of real-world data, they try to learn and improve themselves. To illustrate this, here are some practical applications of artificial intelligence:

Self-Driving Cars: Tesla’s famous self-driving cars is a magnificent real-life application of Artificial Intelligence. These cars have in-built iot sensors for image recognition, forehead collision, spot monitoring, and many more complex mechanisms that allow them to navigate and work in real life.

Google Translate: Google Translate is another great application of Artificial Intelligence. It helps us translate sentences formed in one language to another. It can also translate the entire text on websites, which is possible only because of Artificial Intelligence.

Amazon’s Alexa: Alexa includes a speech recognition system that listens to our voice commands and gives answers. It recognizes our voice and then interprets it as a series of commands and returns the results to us. It uses AVS (Alexa Voice Service), which Amazon provides for free of cost.

Google Maps: Today, without Google Maps, it is impossible to survive in the city. With Google Maps, we can travel from one place to another without any difficulty. All we have to do is open Google Maps and enter our location. Then, its navigation will lead us with the most optimized path to our destination. This is also one of the wonderful applications of artificial intelligence

**3.4 SUBSETS OF ARTIFICIAL INTELLIGENCE**

Artificial Intelligence is an umbrella term. There are two subsets of Artificial Intelligence: Machine Learning and Deep Learning.

**3.5 APPLICATIONS OF MACHINES LEARNING**

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications of ML −

Emotion analysis

Sentiment analysis

Error detection and prevention

Weather forecasting and prediction

Stock market analysis and forecasting

Speech synthesis

Speech recognition

Customer segmentation

Object recognition

Fraud detection

Fraud prevention

Recommendation of products to customer in online shopping

**3.6 TYPES OF MACHINE LEARNING:**

Machine Learning Algorithms can be classified into 3 types as follows –

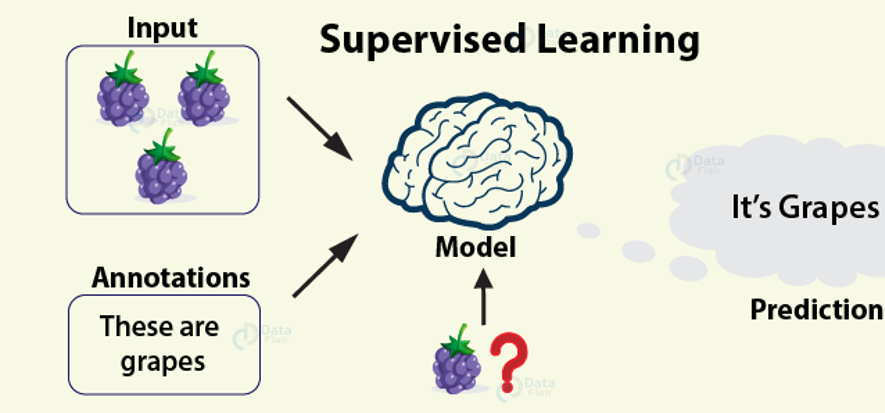
Supervised Learning

Unsupervised Learning

Reinforcement Learning

**3.6.1 Supervised Learning**

Supervised Learning is the most popular paradigm for performing machine learning operations. It is widely used for data where there is a precise mapping between input-output data. The dataset, in this case, is labeled, meaning that the algorithm identifies the features explicitly and carries out predictions or classification accordingly. As the training period progresses, the algorithm is able to identify the relationships between the two variables such that we can predict a new outcome.



Resulting Supervised learning algorithms are task-oriented. As we provide it with more and more examples, it is able to learn more properly so that it can undertake the task and yield us the output more accurately. Some of the algorithms that come under supervised learning are as follows –

**Linear Regression**

In[linear regression](https://data-flair.training/blogs/r-linear-regression-tutorial/), we measure the linear relationship between two or more than two variables. Based on this relationship, we perform predictions that follow this linear pattern.

Random Forest

Random Forests are an ensemble learning method that is for performing classification, regression as well as other tasks through the construction of decision trees and providing the output as a class which is the mode or mean of the underlying individual trees.

**Gradient Boosting**

[Gradient Boosting](https://data-flair.training/blogs/gradient-boosting-algorithm/) is an ensemble learning method that is a collection of several weak decision trees which results in a powerful classifier.

**Support Vector Machine:**

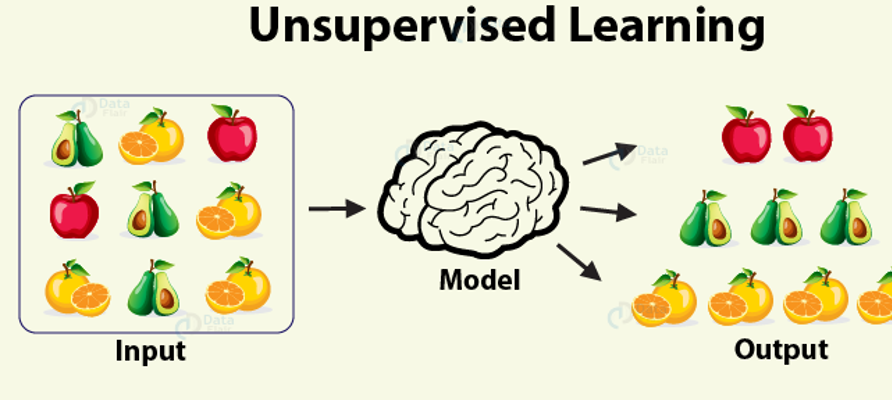
Svms are powerful classifiers that are used for classifying the binary dataset into two classes with the help of hyperplanes.

**Logistic Regression:**

It makes use of a bell-shaped S curve that is generated with the help of logit function to categorize the data into their respective classes.

**3.3.2 Unsupervised Learning:**

In the case of unsupervised learning algorithm, the data is not explicitly labeled into different classes, that is, there are no labels. The model is able to learn from the data by finding implicit patterns. Unsupervised Learning algorithms identify the data based on their densities, structures, similar segments, and other similar features. Unsupervised Learning Algorithms are based on Hebbian Learning. Cluster analysis is one of the most widely used techniques in supervised learning. Let us look at some of the important algorithms that come under Unsupervised Learning.



**Clustering**

Clustering, also known as cluster analysis, is a technique of grouping similar sets of objects in the same group that is different from the objects in other group. Some of the essential clustering techniques are as follows –

**K-means**

The aim of the [k-means clustering algorithm](https://data-flair.training/blogs/k-means-clustering-tutorial/) is to partition the n observations in the data into k clusters such that each observation belongs to the cluster with the nearest mean. This serves as the prototype of the cluster.

**DBSCAN**

This is a clustering method that groups the data based on the density. It groups together the points that are given in the space and marks the outliers in the low-density region.

**Hierarchical clustering**

In this form of clustering, a hierarchy of clusters is built.

**Anomaly Detection**

Anomaly Detection techniques detect outliers in the unlabeled data under an assumption that most of the data examples are normal by observing the instances that fit the remainder of the data set.

**Autoencoders**

Autoencoders are a type of Neural Networks that are used in Unsupervised Learning for representation learning. They are used in denoising and dimensionality reduction.

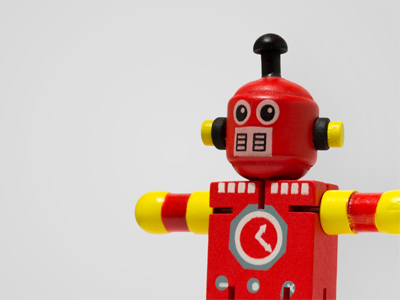
**Deep Belief Network**

It is a generative graphical model which is also a class of neural network designed for unsupervised learning. It is different from the supervised type of neural networks in the sense that it probabilistically reconstructs its inputs to act as feature detectors.

**Principal Component Analysis**

It is a class of unsupervised learning paradigm which is used for reducing the dimensions of the data.

**3.6.3 Reinforcement Learning**

[](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2019/08/reinforcement-learning.jpg)

Reinforcement Learning covers more area of [Artificial Intelligence](https://data-flair.training/blogs/ai-tutorials-home/) which allows machines to interact with their dynamic environment in order to reach their goals. With this, machines and software agents are able to evaluate the ideal behavior in a specific context. With the help of this reward feedback, agents are able to learn the behavior and improve it in the longer run. This simple feedback reward is known as a reinforcement signal.

[](https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2019/08/agent-environment-reinforcement-learning.png)

The agent in the environment is required to take actions that are based on the current state. This type of learning is different from Supervised Learning in the sense that the training data in the former has output mapping provided such that the model is capable of learning the correct answer. Whereas, in the case of reinforcement learning, there is no answer key provided to the agent when they have to perform a particular task. When there is no training dataset, it learns from its own experience.

Based on learning ability

In the learning process, the following are some methods that are based on learning ability −

**Batch Learning**

In many cases, we have end-to-end Machine Learning systems in which we need to train the model in one go by using whole available training data. Such kind of learning method or algorithm is called Batch or Offline learning. It is called Batch or Offline learning because it is a one-time procedure and the model will be trained with data in one single batch. The following are the main steps of Batch learning methods −

Step 1 − First, we need to collect all the training data for start training the model.

Step 2 − Now, start the training of model by providing whole training data in one go.

Step 3 − Next, stop learning/training process once you got satisfactory results/performance.

Step 4 − Finally, deploy this trained model into production. Here, it will predict the output for new data sample.

**Online Learning**

It is completely opposite to the batch or offline learning methods. In these learning methods, the training data is supplied in multiple incremental batches, called mini-batches, to the algorithm. Followings are the main steps of Online learning methods −

Step 1 − First, we need to collect all the training data for starting training of the model.

Step 2 − Now, start the training of model by providing a mini-batch of training data to the algorithm.

Step 3 − Next, we need to provide the mini-batches of training data in multiple increments to the algorithm.

Step 4 − As it will not stop like batch learning hence after providing whole training data in mini-batches, provide new data samples also to it.

Step 5 − Finally, it will keep learning over a period of time based on the new data samples.

Based on Generalization Approach

In the learning process, followings are some methods that are based on generalization approaches −

**Instance based Learning**

Instance based learning method is one of the useful methods that build the ML models by doing generalization based on the input data. It is opposite to the previously studied learning methods in the way that this kind of learning involves ML systems as well as methods that uses the raw data points themselves to draw the outcomes for newer data samples without building an explicit model on training data.

In simple words, instance-based learning basically starts working by looking at the input data points and then using a similarity metric, it will generalize and predict the new data points.

**Model based Learning**

In Model based learning methods, an iterative process takes place on the ML models that are built based on various model parameters, called hyperparameters and in which input data is used to extract the features. In this learning, hyperparameters are optimized based on various model validation techniques. That is why we can say that Model based learning methods uses more traditional ML approach towards generalization.

**CHAPTER 4**

**SOFTWARE REQUIREMENTS & ALGORITHMS**

**4.1 INTRODUCTION TO PYTHON**

Python is a widely used high-level programming language, created by Guido Van Rossum and first released in 1991. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using white space indentation to delimit code blocks rather than curly brackets or keywords) and syntax that allows the programmer to express concepts in fewer lines of code that might be used in programming languages such as C++ or Java. It provides constructs that enable to clear programming on both small and large scales.

Python features a dynamic type system & automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

Python interpreters are available for many operating systems. The reference implementation of Python is open source software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation.

Programming languages have been for ages, and every decade sees the launch of a new language sweeping the developers off their feet. Below are the major features and applications due to which people choose Python as their first programming language:

* Python’s popularity and high salary.
* Python is used in Data Science.
* Python’s scripting and automation.
* Python used with Big Data.
* Python supports Testing.
* Computer Graphics in Python.
* Python used in Artificial Intelligence.
* Python in Web Development.
* Python is portable and extensible.
* Python is simple and easy to learn.

**4.2 HISTORY OF PYTHON:**

Python was conceived in the late 1980’s and its implementation began in December 1989 by Guido van Rossum at Centrum Wiskunde and Informatica (CWI) in the Netherlands as a successor to the ABC language (itself inspired by SETL) capable of exception handling and interfacing with the Amoeba operating system.

Guido Van Rossum remains Python's principal author. His continuing central role in Python’s development is reflected in the title ‘Benevolent Dictator for Life (BDFL)’ given to him by the Python community.

Python 2.0 was released on 16 October 2000 and had many major new features, including cycle-detecting garbage collector. With this release, the development process became more transparent.

Python 3.0 (initially called Python 3000 or py3k) was released on 3 December 2008 after a long testing period. It is a major version of the language that is not backward compatible with previous versions. However, many of its major features have been backported to the backward compatible Python 2.6.x and 2.7.x version series.

Python 2.7's end-of-life date (a.k.a. EOL, sunset date) was initially set at 2015, and then postponed to 2020 out of concern that a large body of existing code could not easily be forward ported to Python 3.x.

Python uses duck typing and has typed objects but untyped variable names. Type constraints are not checked at compile time rather, operations on an object may fail, signifying that the given object is not a suitable type. Despite being dynamically typed, Python is strongly typed, forbidding operations that are not well-defined (for example, adding a number to a string) rather than silently attempting to make sense of them.

Python allows the programmer to define their own types using classes, which are most often used for Object-Oriented Programming (OOP). New instances of classes are constructed by calling the class.

**4.3 FEATURES OF PYTHON:**

* **Easy-to-learn:** Python has few keywords, simple structure, and a clearly defined syntax. It allows the student to pick up the language quickly.
* **Easy-to-read:** Python code is more clearly defined and visible to the eyes.
* **Easy-to-maintain:** Python's source code is fairly easy-to-maintain.
* **A broad standard library:** Python's bulk library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
* **Interactive Mode:** Python has support for an interactive mode which allows the interactive testing and debugging of code snippets.
* **Portable:** Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
* **Extendable:** We can add low-level modules to the Python interpreter. These modules enable the programmers to add or customize their tools to be more efficient.
* **Databases:** Python provides interfaces to all major commercial databases.
* **GUI Programming:** Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X window system of UNIX.
* **Scalable:** Python provides a better structure and support for large programs than shell scripting.
* It supports functional and structured programming methods as well as OOP.
* It can be used as a scripting language or can be compiled to byte-code for building the large applications.
* It provides very high-level dynamic datatypes and supports dynamic type checking.
* It supports automatic garbage collection.

It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

**4.4 INSTALLING OF PYTHON:**

**DOWNLOADING OF PYTHON:**

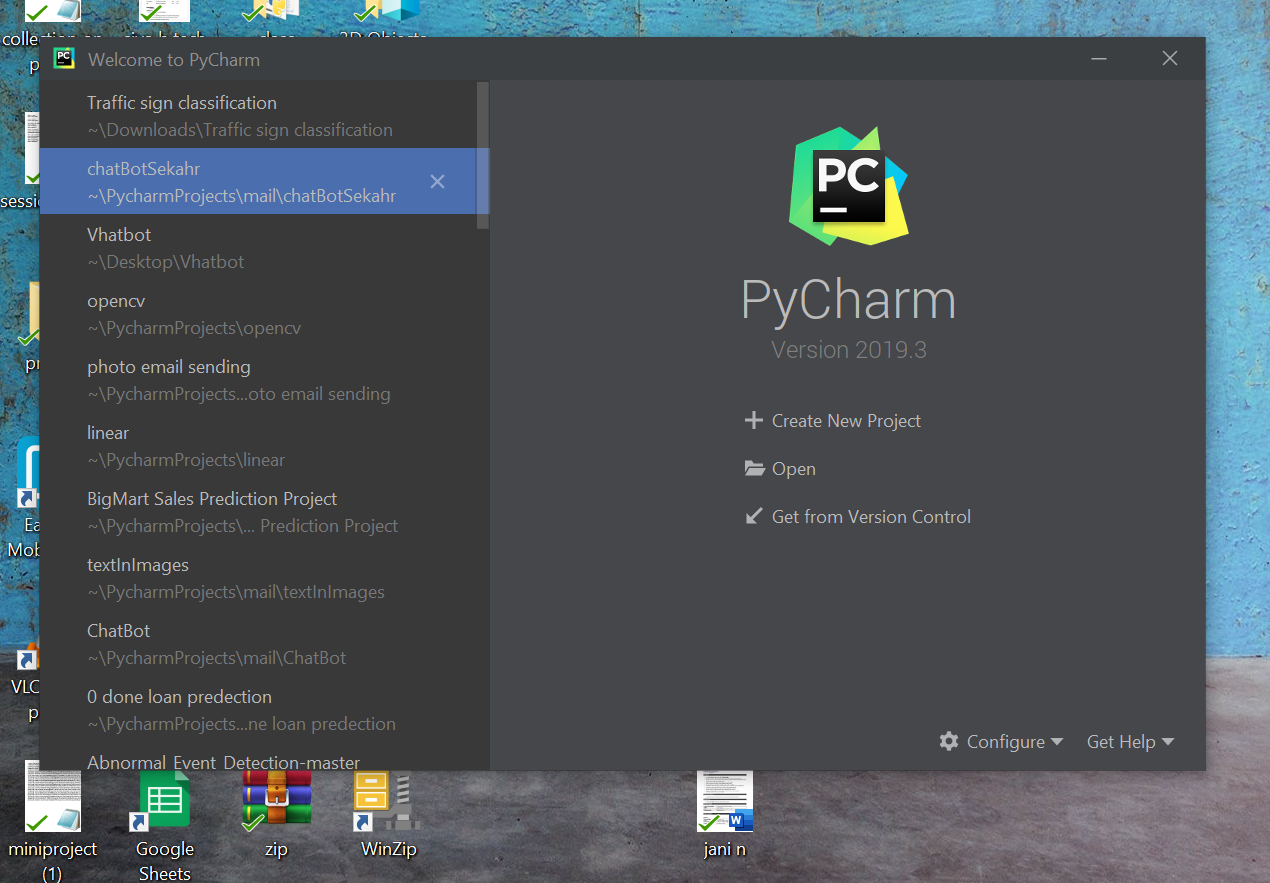
To download Python we have to follow the below steps:

* Go to www.python.org
* Go to the downloads tab.
* Click on the Python version you required to download according to the OS you are using.
* Double click on the downloaded Python version.
* Select the path for Python.
* Click on install now.
* Click on yes.
* Click on close once it was done.

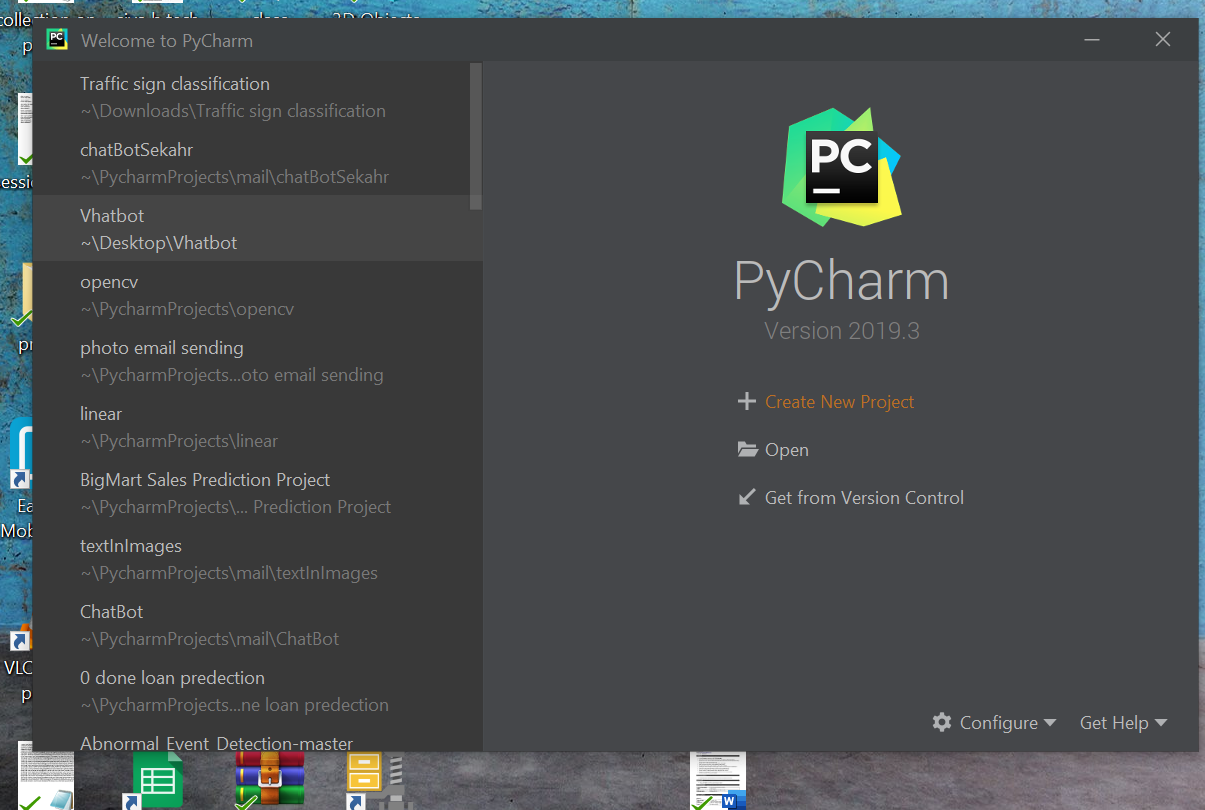
**4.5 INSTALLING OF PYTHON:**

Python distribution is available for a wide variety of platforms. You need to download only the binary code which is applicable for your platform (OS) and install Python. If the binary code for your platform is not available, you need a C compiler to compile the source code manually. Here is a quick overview of installing Python on Windows platform.

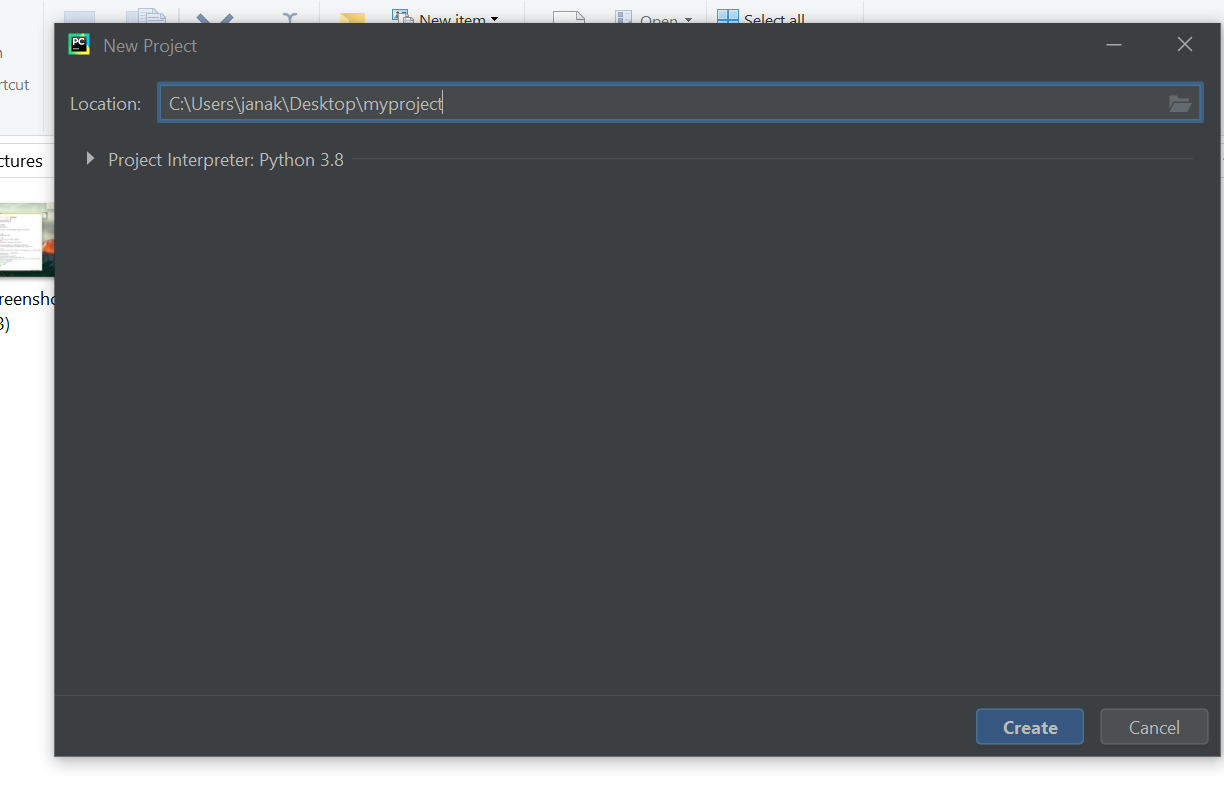
**4.6 PROJECT CREATION STEPS**



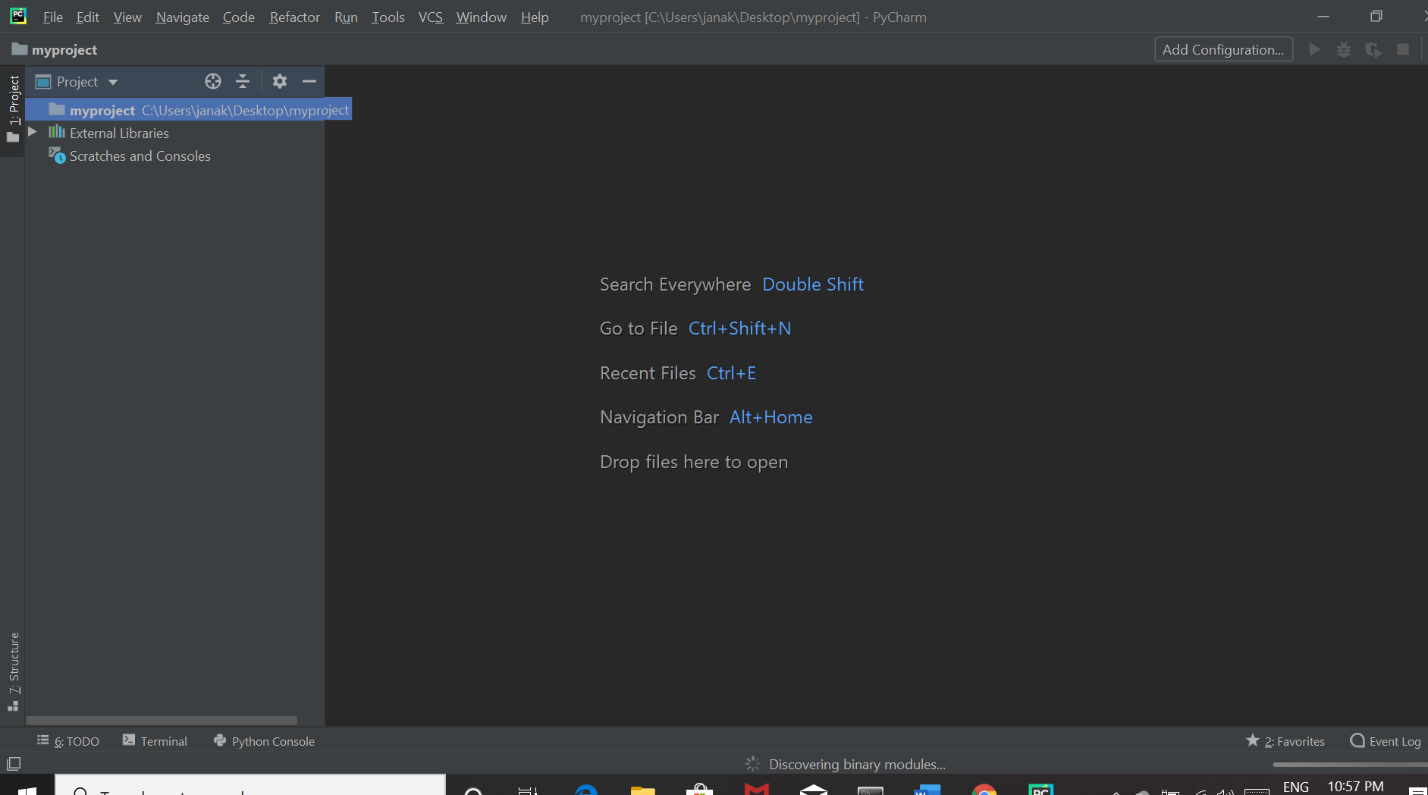
🡪click on create new project



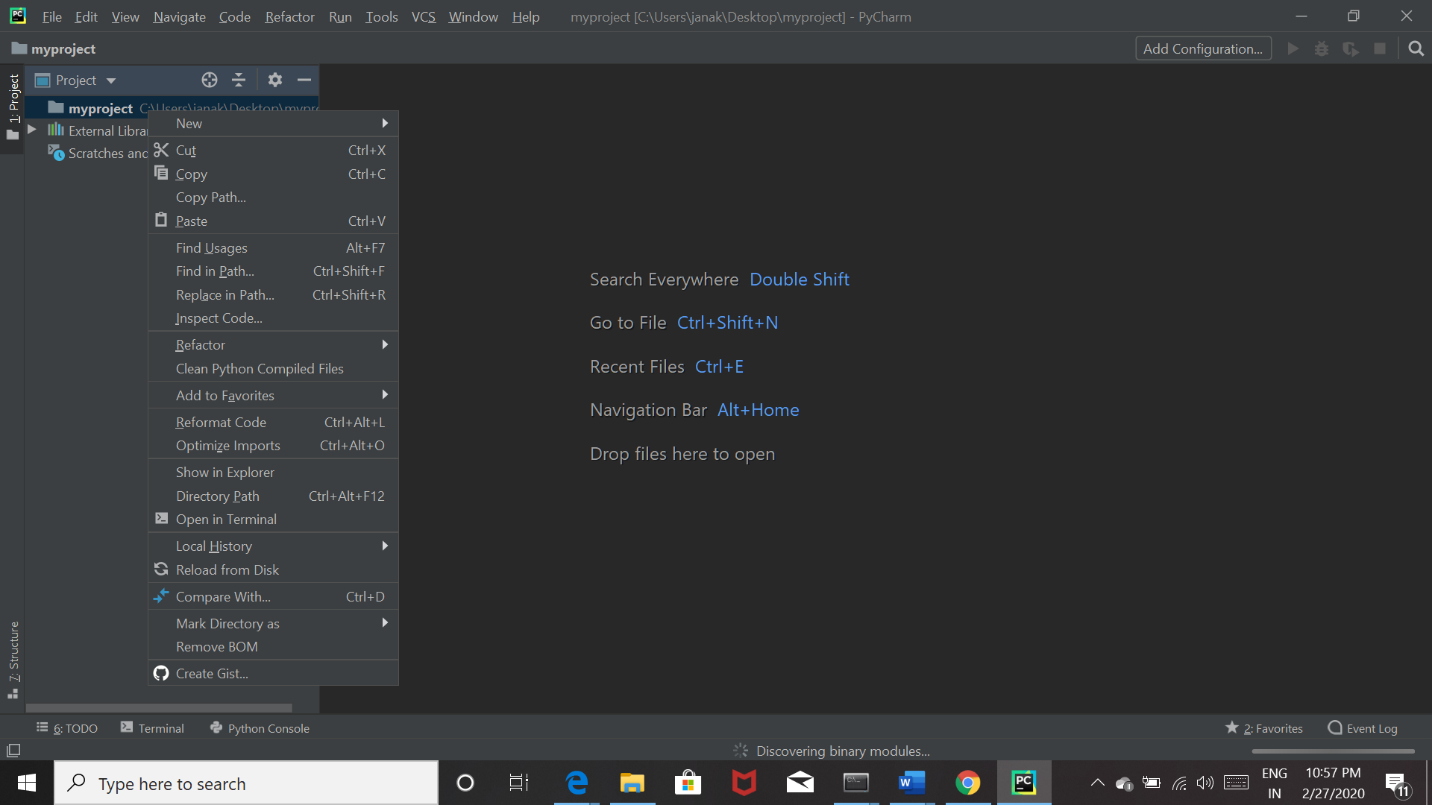
🡪Write project name



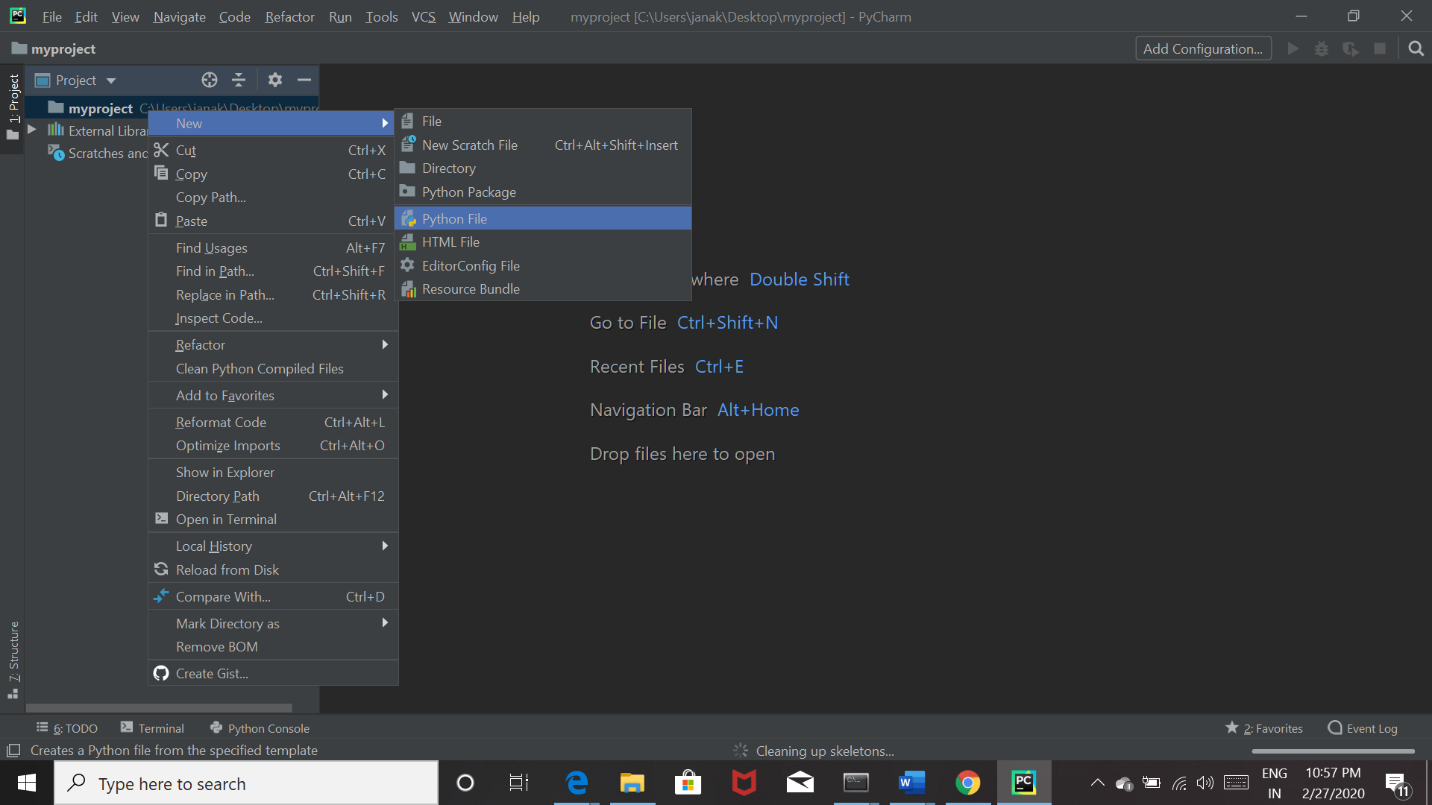
🡪Right Click on project name



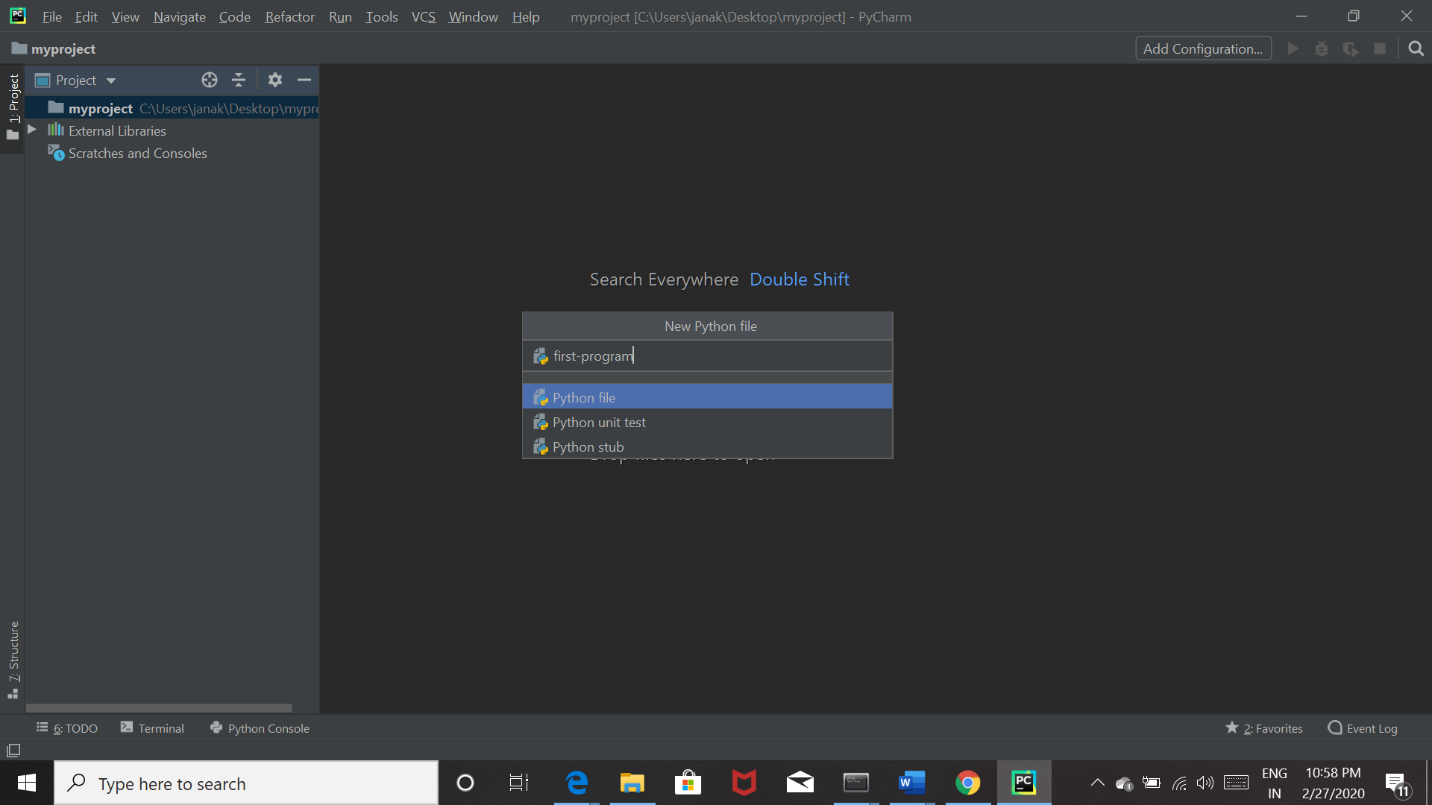
* Click on new



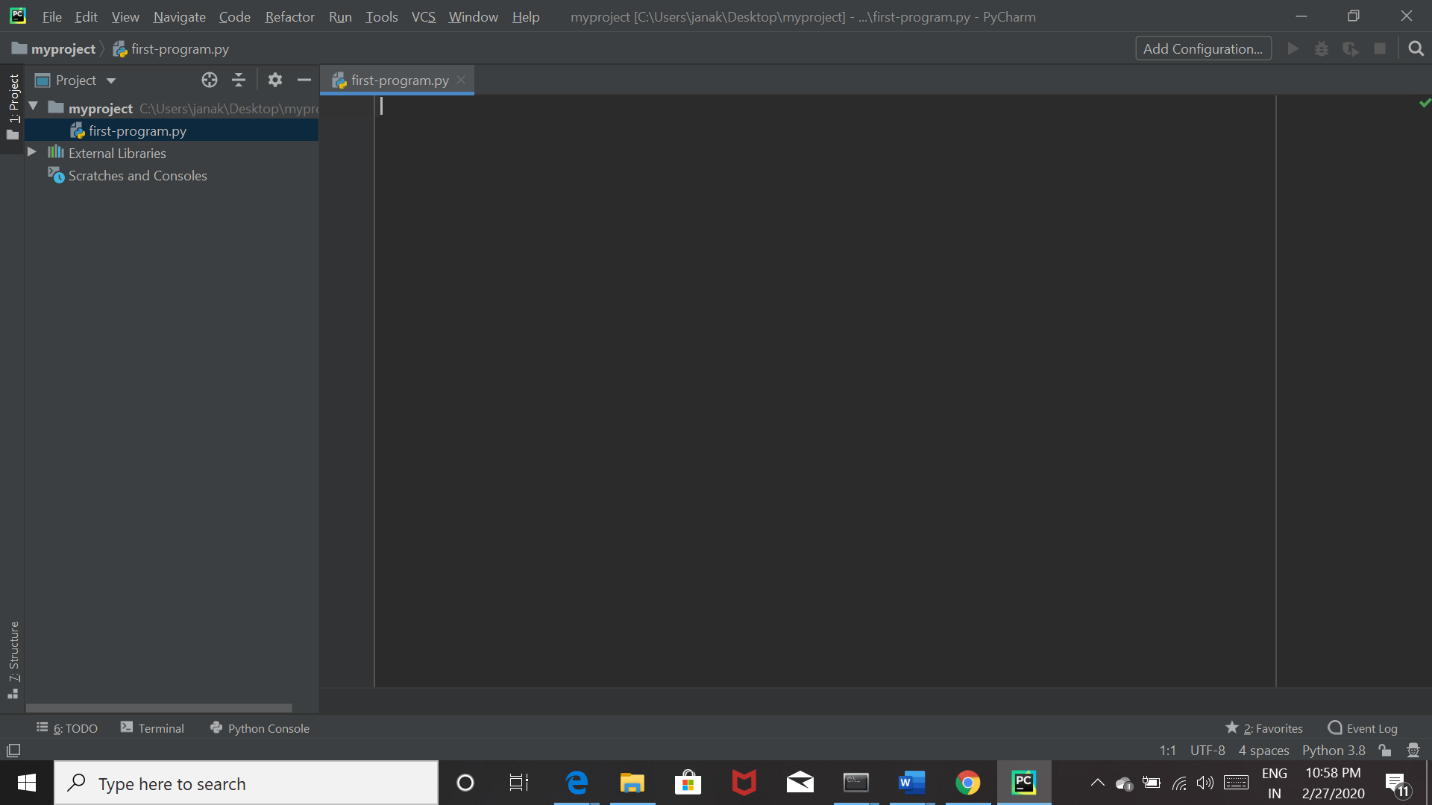
🡪click on python file



🡪write file name



-🡪write the code



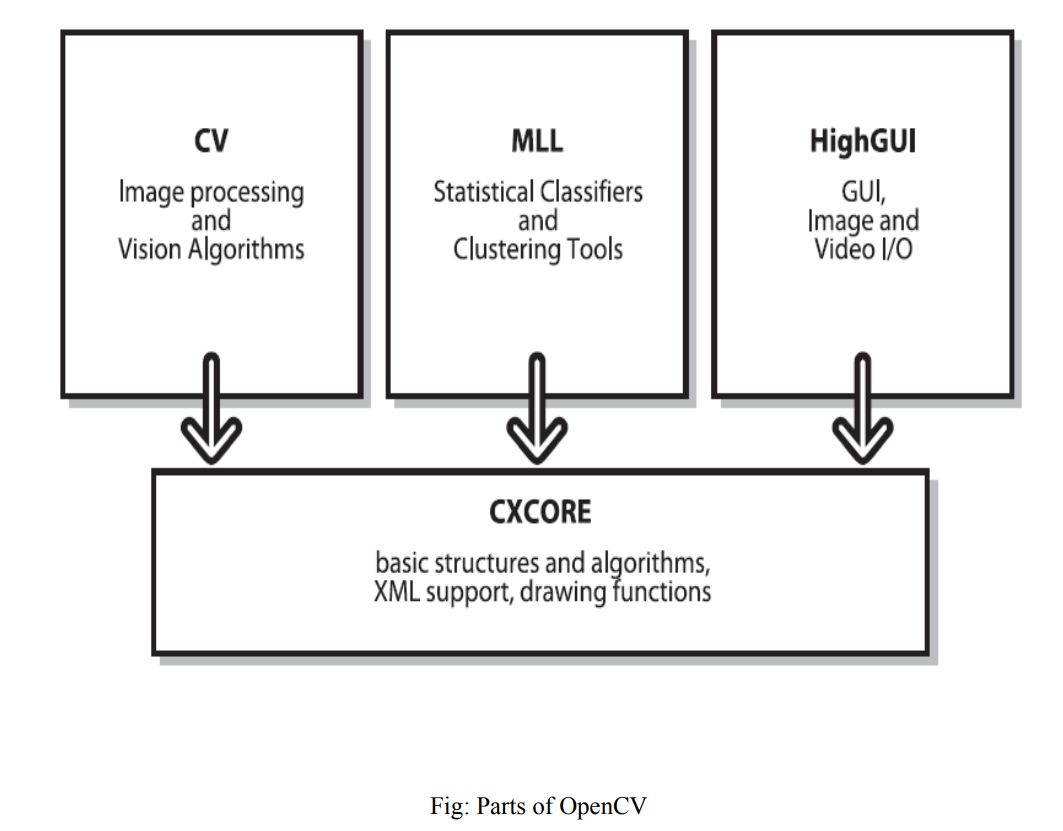
**OpenCV** :

OpenCV is an open source computer vision library accessible in python coding language to code for visionary capabilities of our smart pc. OpenCV was expected for computational capability and having a high focus on ongoing picture location and distinguishing proof. OpenCV is coded with streamlined C and can take work with multicore processors. If we need progressively programmed improvement utilizing Intel models [Intel], you can purchase Intel's Integrated Performance Primitives (IPP) libraries [IPP]. These comprise of low-level schedules in different algorithmic regions which are streamlined. OpenCV consequently utilizes the IPP library, at runtime if that library is introduced. alarm beeps and the speed of the vehicle is automatically reduced.

The Computer’s Vision PC's vision is the change of information from a still, or camcorder into either a depiction or another choice. Each and every such changes are performed to achieve a particular target. A Computer gains a cross section of numbers from a camera or from the circle, and it's just as simple as that. For the most part, there is no worked in example acknowledgment or programmed control of center and gap, no cross-relationship with long periods of experience. Generally, vision frameworks are still reasonably gullible. The Origin of OpenCV OpenCV left an Intel Research action proposed to drive CPU-raised applications.Toward this end, Intel moved various endeavors that included constant beam following and moreover 3D show dividers. One of the product engineers working for Intel at the time was visiting schools. He saw that several top school social affairs, like the MIT Media Lab, used to have well-made similarly as inside open PC vision frameworks—code which was supplied starting with one understudy then onto the next and which gave each resulting understudy an important establishment while building up his own vision application. Rather than rehashing the fundamental capacities from starting, another understudy may begin by adding to that which preceded .

**OpenCV Structure and Content:**

OpenCV left an Intel Research movement planned to drive CPU-raised applications.Toward this end, Intel pushed various endeavors that included continuous beam following and moreover 3D show dividers. One of the product engineers working for Intel at the time was visiting schools. He saw that two or three top school social events, like the MIT Media Lab, used to have well-made similarly as inside open PC vision foundations

****

**Why Open CV?**

**Specific** :OpenCV was planned for picture handling. Each capacity and information structure has been arranged in perspective on an Image Processing application. Then, Matlab, is very conventional. You can get almost everything on the planet by methods for tool compartments. It may be money related tool stash or then again concentrated DNA tool compartments

**• Speedy** : Matlab is just excessively moderate. Matlab itself depended on Java. Similarly Java depended on C. So when we run a Matlab program, our PC gets caught up with attempting to translate and assemble all that convoluted Matlab code. At that point it is transformed into Java, lastly executes the code. In case we use C/C++, we don't waste such time. We direct give machine language code to the PC, and it gets executed. So in the end we get more picture taking care of, and not additionally interpreting. Ensuing to doing some constant picture handling with both Matlab and OpenCV, we typically got low speeds, a point of confinement of around 4-5 outlines arranged each second with Matlab. With OpenCV in any case, we get genuine persistent dealing with at around 30 outlines being handled every second. 22Beyond any confusion we give the prize for speediness – a progressively enigmatic language to handle, yet it's unquestionably of true worth . We can complete a lot more work, as calculate some extremely perplexing arithmetic on pictures utilizing C and still pull off adequatespeeds for your application.

**• Efficient**:

Matlab utilizes just an excessive amount of system assets. With OpenCV, we can pull off as pitiful as 10mb RAM for a constant application. Notwithstanding the way that with the present PCs, the RAM factor is surely not a noteworthy thing to be worried over. In any case, our tiredness identification framework is to be used inside a vehicle in a way that is non-meddlesome and little; so a low handling necessity is vital. Subsequently we shall perceive as to how OpenCV is superior for a real-time drowsiness detection system

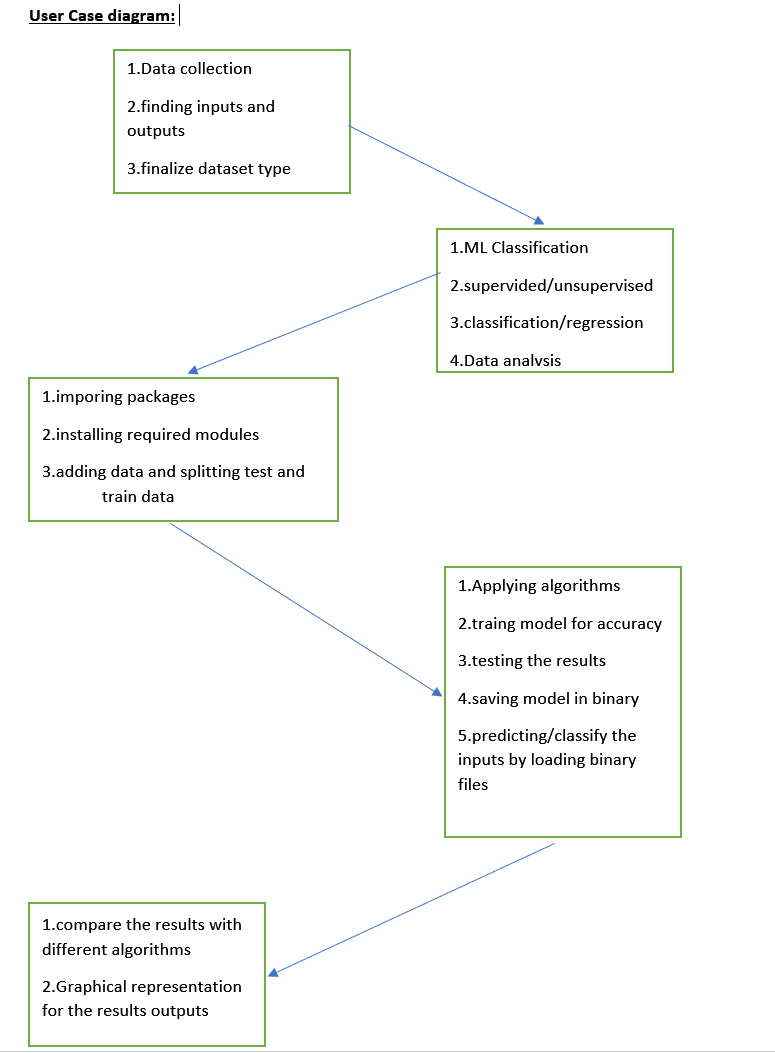
**CHAPTER 5**

**SYSTEM DESIGN & TESTING**

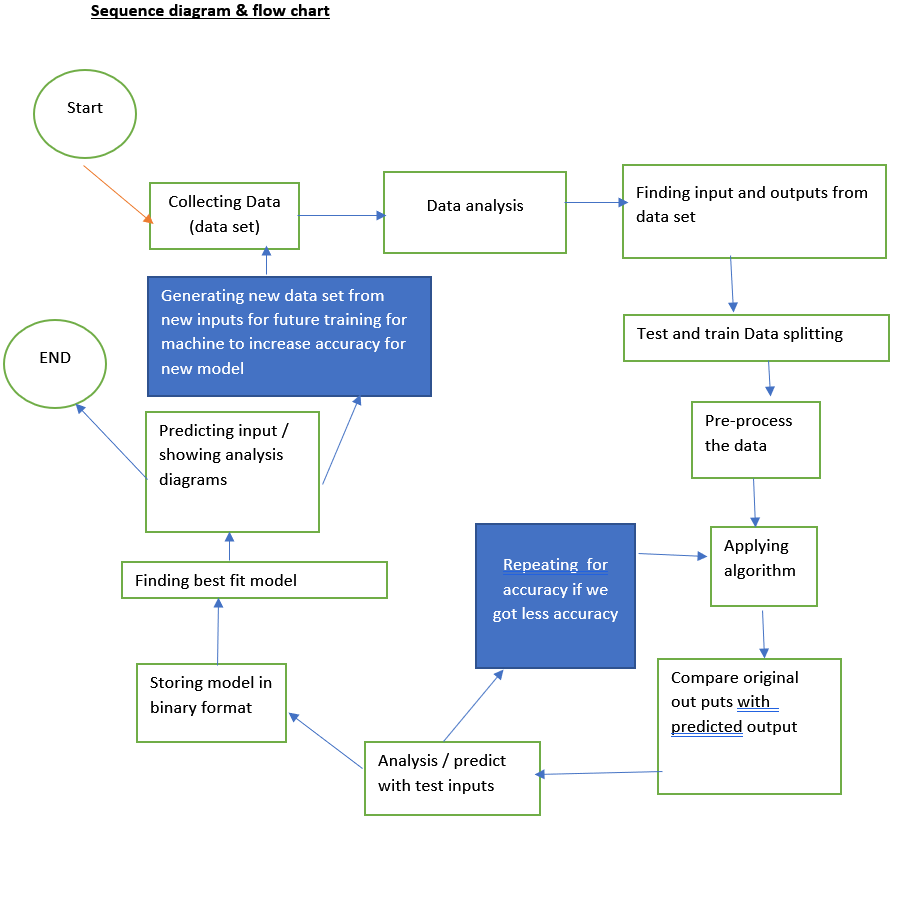
**5.1 INTRODUCTION**

The purpose of the design phase is to plan a solution of the problem specified by the requirement document. It is the process of defining software methods, functions, objects and overall structure and interaction of your code so that the resulting functionality will satisfy your users requirements. It allows you to do the best abstraction, to understand the requirements better and meet them better. This prevents redundancy and increases reusability. This phase is the first step in moving from the problem domain to the solution domain. In other words, starting with what is needed design takes us towards how to satisfy the needs. The design of a system is perhaps the most critical factor affecting the quality of the software; it has a major impact on the later phase, particularly testing, maintenance. The output of this phase is the design document. This document is similar to a blueprint for the solution and is used later during implementation, testing and maintenance. The design activity is often divided into two separate phases System Design and Detailed Design. System Design also called top-level design sign aims to identify the modules that should be in the system, the specifications of these modules, and how they interact with each other to produce the desired results. During detailed design, the internal logic of each of the modules specification in system design is decided. During this phase, the details of the data are usually specified in a high-level design description language, which is independent of the target language in which the software will eventually be implemented. In system design the focus is on identifying the modules, whereas during detailed design the focus is on designing the logic for each of the modules. During the system design activities, Developers bridge the gap between the requirements specification, produced during requirements elicitation and analysis, and the system that is delivered to the user.

**5.2 ALGORITHM & IMPLEMENTATION**

****

**5.3 ALGORITHAM MODEL FLOW DIAGREAM**

****

**5.5 Machine Learning Testing**

First of all, what are we trying to achieve when performing ML testing, as well as any software testing whatsoever?

* Quality assurance is required to make sure that the software system works according to the requirements. Were all the features implemented as agreed? Does the program behave as expected? All the parameters that you test the program against should be stated in the technical specification document.
* Moreover, software testing has the power to point out all the defects and flaws during development. You don’t want your clients to encounter bugs after the software is released and come to you waving their fists. Different kinds of testing allow us to catch bugs that are visible only during runtime.

However, in machine learning, a programmer usually inputs the data and the desired behaviour, and the logic is elaborated by the machine. This is especially true for deep learning. Therefore, the purpose of machine learning testing is, first of all, to ensure that this learned logic will remain consistent, no matter how many times we call the program.

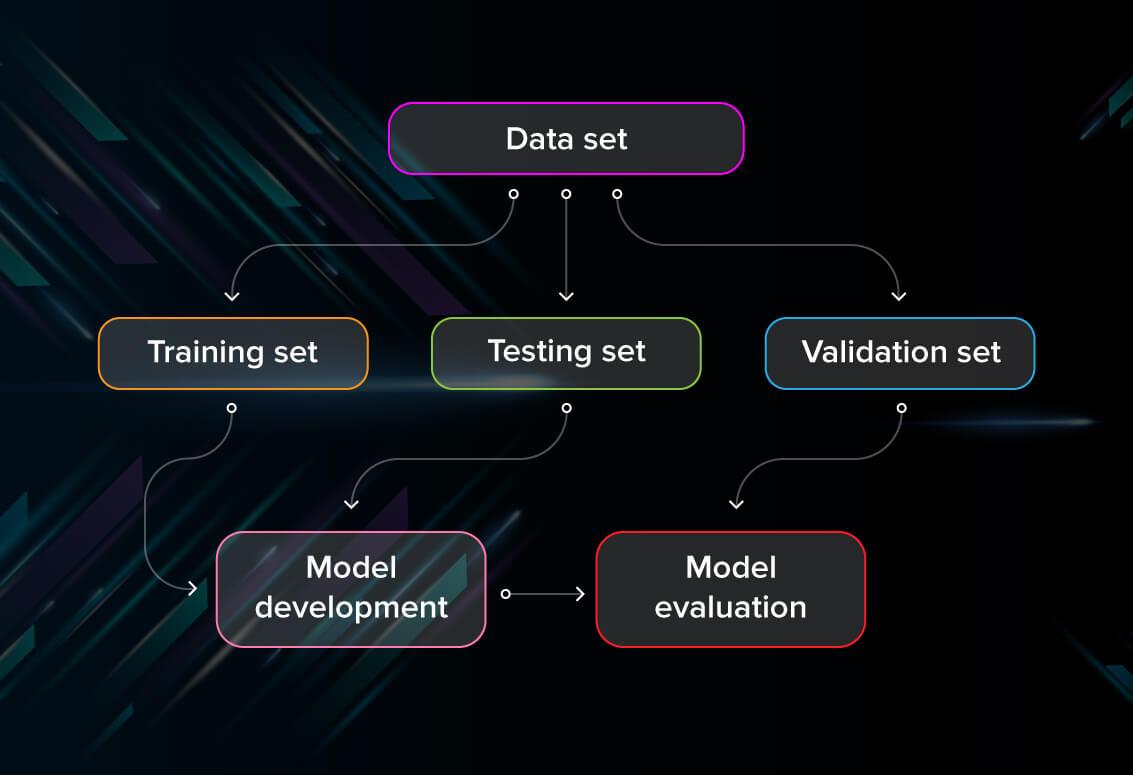
## **Model evaluation in machine learning testing**

Usually, software testing includes:

* **Unit tests.** The program is broken down into blocks, and each element (unit) is tested separately.
* **Regression tests.** They cover already tested software to see if it doesn’t suddenly break.
* **Integration tests.** This type of testing observes how multiple components of the program work together.

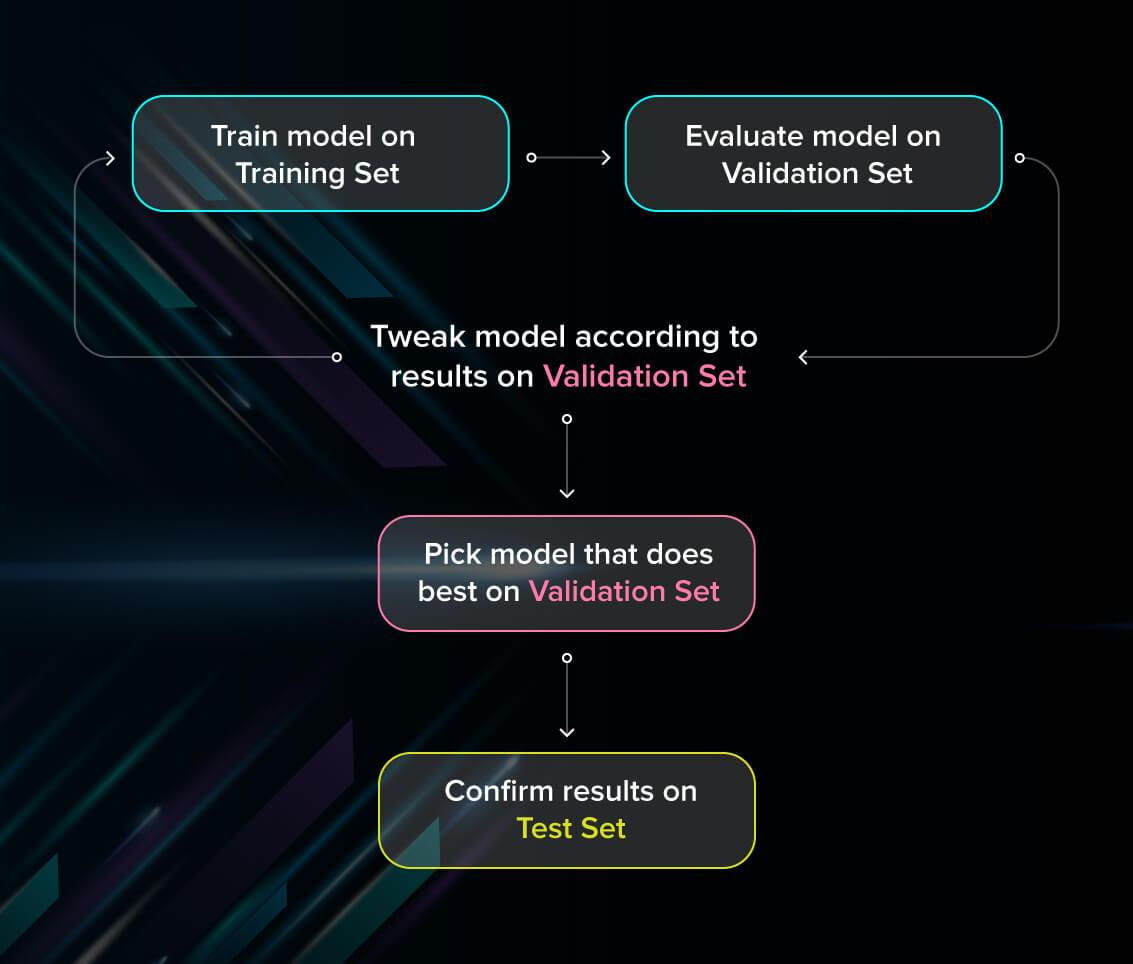
Moreover, there are certain rules that people follow: don’t merge the code before it passes all the tests, always test newly introduced blocks of code, when fixing bugs, write a test that captures the bug.

Machine learning adds up more actions to your to-do list. You still need to follow [ML’s best practices](https://developers.google.com/machine-learning/testing-debugging/common/overview). Moreover, every ML model needs not only to be tested but **evaluated**. Your model should generalize well. This is not what we usually understand by testing, but evaluation is needed to make sure that the performance is satisfactory.



First of all, you split the database into three non-overlapping sets. You use a training set to train the model. Then, to evaluate the performance of the model, you use two sets of data:

* **Validation set.** Having only a training set and a testing set is not enough if you do many rounds of hyperparameter-tuning (which is always). And that can result in overfitting. To avoid that, you can select a small validation data set to evaluate a model. Only after you get maximum accuracy on the validation set, you make the testing set come into the game.
* **Test set (or holdout set).** Your model might fit the training dataset perfectly well. But where are the guarantees that it will do equally well in real-life? In order to assure that, you select samples for a testing set from your training set — examples that the machine hasn’t seen before. It is important to remain unbiased during selection and draw samples at random. Also, you should not use the same set many times to avoid training on your test data. Your test set should be large enough to provide statistically meaningful results and be representative of the data set as a whole.



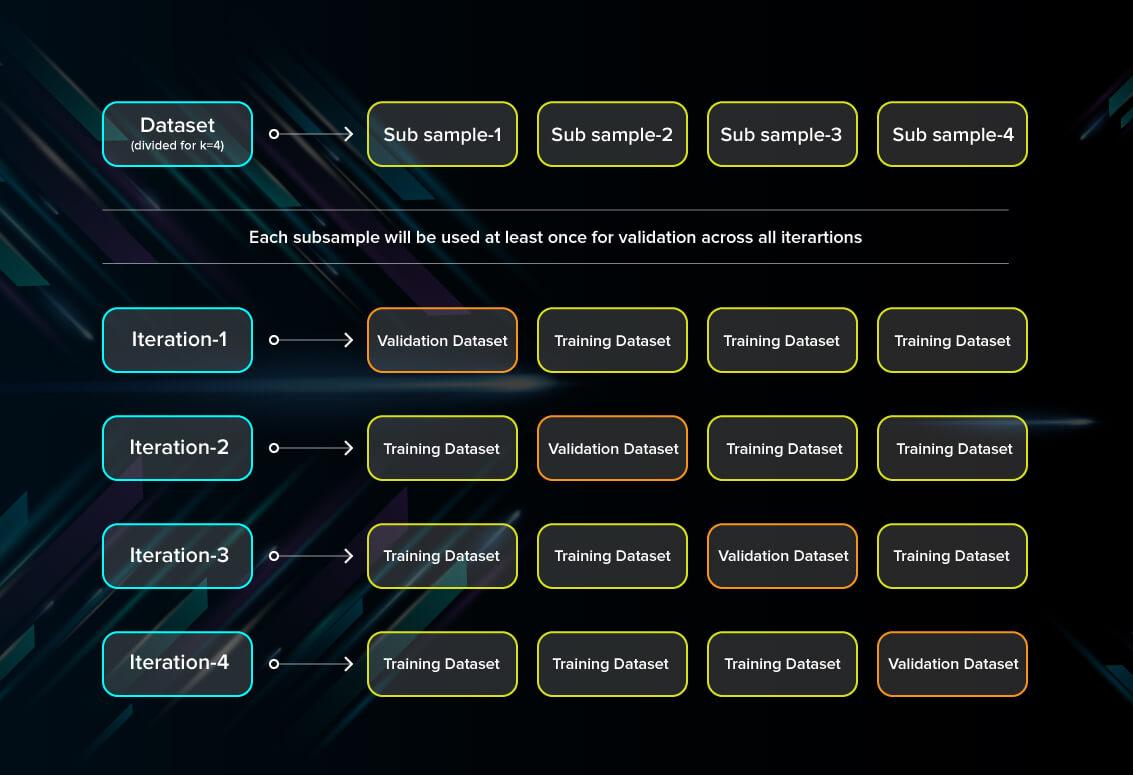
But just as test sets, validation sets “wear out” when used repeatedly. The more times you use the same data to make decisions about hyperparameter settings or other model improvements, the less confident you are that the model will generalize well on new, unseen data. So it is a good idea to collect more data to ‘freshen up’ the test set and validation set.

### Cross-validation

Cross-validation is a model evaluation technique that can be performed even on a limited dataset. The training set is divided into small subsets, and the model is trained and validated on each of these samples.

#### k-fold cross-validation

The most common cross-validation method is called k-fold cross-validation. To use it, you need to divide the dataset into kk subsets (also called folds) and use them kk times. For example, by breaking the dataset into 10 subsets, you will perform a 10-fold cross-validation. Each subset must be used as the validation set at least once.



This method is useful to test the skill of the machine learning model on unseen data. It is so popular because it is simple to apply, works well even with relatively small datasets, and the results you get are generally quite accurate. If you want to learn more about how to cross-validate the model, [check out a more detailed explanation on Medium](https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f).

#### Leave-one-out cross-validation

In this method, we train the model on all the data samples in the set except for one data point that is used to test the model. By repeating this process iteratively, each time leaving a different data point as a testing set, you get to test the performance for all the data.

The benefit of the method is low bias since all the data points are used. However, it also leads to higher variation in testing because we are testing the model against just one data point each time.

### Evaluate models using metrics

Evaluating the performance of the model using different metrics is integral to every data science project. Here is what you have to keep an eye on:

#### Accuracy

Accuracy is a metric for how much of the predictions the model makes are true. The higher the accuracy is, the better. However, it is not the only important metric when you estimate the performance.

Accuracy≡True Positives+True NegativesTrue Positives+False Positives+True Negatives+False NegativesAccuracy≡True Positives+True NegativesTrue Positives+False Positives+True Negatives+False Negatives

#### Loss

Loss describes the percentage of bad predictions. If the model’s prediction is perfect, the loss is zero; otherwise, the loss is greater.

#### Precision

The precision metric marks how often the model is correct when identifying positive results. For example, how often the model diagnoses cancer to patients who really have cancer.

Precision≡True PositivesTrue Positives+False PositivesPrecision≡True PositivesTrue Positives+False Positives

#### Recall

This metric measures the number of correct predictions, divided by the number of results that should have been predicted correctly. It refers to the percentage of total relevant results correctly classified by your algorithm.

Recall≡True PositivesTrue Positives+False NegativesRecall≡True PositivesTrue Positives+False Negatives

#### Confusion matrix

A confusion matrix is an N×NN×N square table, where NN is the number of classes that the model needs to classify. Usually, this method is applied to classification where each column represents a label. For example, if you need to categorize fruits into three categories: oranges, apples, and bananas, you draw a 3×33×3 table. One axis will be the actual label, and the other will be the predicted one.



### Pre-train tests

This type of test is performed early on and allows you to catch bugs before running the model. They do not need training parameters to be run. An example of a pre-train test is a program that checks whether there are any labels missing in your training and validation datasets.

### Post-train tests

These tests are performed on a trained model and check whether it performs correctly. They allow us to investigate the logic behind the algorithm and see whether there are any bugs there. There are three types of tests that report the behavior of the program:

* Invariance tests. Using invariance tests, we can check how much we can change the input without it affecting the performance of the model. We can pair up input examples and check for consistency in predictions. For example, if we run a pattern recognition model on two different photos of red apples, we expect that the result will not change much.
* Directional expectation tests. Unlike invariance tests, directional expectation tests are needed to check how perturbations in input will change the behavior of the model. For example, when building a regression model that estimates the prices of houses and takes square meters as one of the parameters, we want to see that adding extra space makes the price go up.
* Minimum functionality tests. These tests enable us to test the components of the program separately just like traditional unit tests. For example, you can assess the model on specific cases found in your data.

**CHAPTER 7**

**RESULTS**

**\*\* OUTPUT SCREEN SHOTS OF YOUR PROJECT \*\***

**CHAPTER 8**

**CONCULSUION**

Data mining has been used in this paper to construct a sample predictive model defining current borrowers' loan histories that can be used for potential loan application comparisons indicating characteristics of a good or bad loan record based on their credit background and specific demographic profile. It also indicates that pre-processing or cleaning of data which plays a vital role in achieving a higher accuracy rate. During pre-processing, patterns outlined in the figures in this paper may also be used to identify target loan markets, future income-enhancing action plans and reduce default risk, and also to enhance the loan products. Machine Learning can help banks in predicting the future of loan and its status and depends on that they can act in initial days of loan. Using Machine Learning banks can reduce the number of bad loans and from incurring sever losses. Using above discussed methodology bank can easily identify the required information from huge amount of data sets and helps in successful loan prediction to reduce the number of bad loan problems. Data Mining and Machine Learning techniques are very useful to the banking sector for better targeting and acquiring new customers, most valuable customer retention, automatic credit approval and marketing.

**Code :**