**Early Stage Lung Cancer Prediction using various ML Techniques**

**ABSTRACT**

For more than a decade, machine learning techniques have been applied in cancer research. Machine Learning Algorithms (MLa) can now make a significant contribution to lung cancer research. Because LC has the highest death rate in the world, early detection and classification of cancer cells can significantly improve survival rates. Though numerous algorithms for LC prediction are employed in the fields of neurology, radiology, and cancer, MLa excels them due to its accuracy and efficiency. This research initially looks at MLa's workflow methodology for LC early prediction and categorization. The approaches involve input data selection, data preparation, feature selection and extraction, and training and testing data and selecting the best ml technique. e. Second, a survey report of the ML algorithms employed in LC is offered, along with their methodology. Third, performance metrics such as Accuracy, Sensitivity, Specificity, Precision, F1 Score, Root Mean Square Error (RMSE), Confusion Matrix, Area Under the Curve (AUC) – Receiver Operating Characteristics (ROC) curve, Precision-Recall (PR) curve with various MLa are investigated. Finally, this study also covers the parameters used in constructing an efficient and accurate ML model for the early prediction of LC.

**CHAPTER 1: INTRODUCTION**

**Introduction:-**

Lung cancer is mainly triggered by cigarette smoke. Smoke that penetrates into the lungs causes damage to the lung tissue. In nonsmokers, lung cancer may be induced by radon radiation, second hand smoking, air contamination or other causes. Heredity is another source of lung cancer, as well. While lung cancer (malignant growth) is hard to diagnose and cure, it may be avoided or treated in the early stages. Lung malignancy is one among the dangerous cancer forms and is commonly found. Currently an approximate of 2.09 million cases observed against lung cancer. Also depending on the stages of the cancer they assign grades to the cancer.

.

**CHAPTER 2: LITERATURE SURVEY**

**Literature survey 1:-**

**Title:-** “Colon Cancer Detection in Biopsy Images for Indian Population at Different Magnification Factors Using Texture Features”

**Year:-** 2017

**Authors**:- Tina Babu; Tripty Singh; Deepa Gupta; Shahin Hameed

**Abstract**:

This paper presents the identification of colon histology images as normal and malicious ones for the determination of colon cancer. Such classification is very difficult as it takes into account numerous characteristics such as color, several biological interpretable factors. Also this process is subjective and may have some observational variations. In this paper the colon biopsy images are classified as normal and malicious ones based on the texture of the images for the Indian population. Various texture features such as GLCM, LBP, HOG, Gabor, GLRLM, Histogram are extracted from the images with different magnification factors such as 10X, 20X and 40X. Based on these features extracted at different magnification factor, they are trained using classifiers SVM with linear kernel, Naive Bayes and Perceptron. The system has been experimented on 24 normal and 35 malignant biopsy images of colon that were acquired fom Aster Medcity, Kochi. SVM and Perceptron is giving higher accuracy for magnification 10X and 20X.

**Literature survey 2:-**

**Title**: “Cancer Screening On Indian Colon Biopsy Images Using Texture and Morphological Features”

**Year:-** 2018

**Authors**: Tina Babu; Tripty Singh ; Deepa Gupta; Shahin Hameed

**Abstract**:-

Colon Cancer Detection in the biopsy images is the tedious task for the pathologist. Thus an automation in this regard may speed up the process and can assist the pathologist. This work focus on the classification of the colon biopsy images to normal and cancerous ones where the hybrid of the texture and morphological features are taken from the Indian Population at different magnification factors. The hybrid of the texture features are taken from the Harallick features, HOG, Histogram, GLRLM, LBP and Gabor. The classification was carried out with classifiers SVM, Nave Bayes, Adaboost, Multilayer Perceptron and Decision Tree. The system has been tried with two datasets which was taken from Aster Hospital, Kochi and HCG Hospital, Bengaluru where the fused feature set of hybrid texture and morphological feature gave a good performance with the classifiers SVM and Multilayer Perceptron.

**Literature survey 3:-**

**Title:-** “ Investigation of chronic disease correlation using data mining techniques”

**Year:-** 2015

**Authors:-** Dr. Deepa Gupta, Sangita Khare, Dominic, V.a; Sangita Khare; Aggarwal, A.b

**Abstract:-**

A disease is an abnormal condition that affects the structure and function of one or more parts of the body. It may be caused by various factors, external and internal dysfunctions. There is a trend of various chronic diseases in any society. The major concern is that these chronic diseases are leading to many other diseases in future. An attempt to explore the correlation of various chronic diseases has become a necessity. This can be achieved by using data mining techniques, which help to derive knowledge about the affects of a particular chronic disease on the other chronic diseases. Since there is growing trend of diabetes and ischemic heart disease in the society, in this paper the focus is to investigate the effect of these diseases on the other chronic diseases using the ICD9 diagnostic codes. To achieve this goal various types of data mining techniques are used. The conclusion is an optimal set of ICD9 diagnostic codes associated with individuals having diabetes or ischemic heart disease. These codes are then investigated based on the human anatomic systems i.e. Circulatory system, Respiratory system, Nervous system, Musculoskeletal system, Renal system and Neoplasm and their relevance is justified.

**3.1 Existing system**

ML techniques handle the data and find the right model. The main category of ML methods is supervised learning (SL), unsupervised learning (USL), and reinforcement learning (RL). ∙ All SL is a form of classification or Regression. ∙ USL is valuable when the information is uncertain but it needs to be investigated. ∙ RL can be model-free or model-based reinforcement learning.

**3.2 Drawbacks**

* Less amount of accuracy score
* Small level data-set
* Applicable on small level prediction work

**3.3 Proposed system**

To construct an efficient and accurate ML model for early LC, the model can be developed with the following parameters.

* Data should be collected from large and highly qualified authorized centers.
* Data collected should be preprocessed by a powerful technique such that no important data is lost.
* Highly correlated Features with the output should be identified for best results.
* Using the Hybrid ML model, early prediction of LC can produce accurate results.
* Several ML tools and various platforms can be made available for researchers to provide good results.
* There are also many data analytical tool that can provide useful information for future data analysis.

**3.4 Advantages**

* Increasing the accuracy score
* Large amount of feature we are taking for the training and testing.

**3.5 System Requirements**

**Hardware:**

* OS – Windows 7, 8 and 10 (32 and 64 bit)
* RAM – 4GB

**Software:**

* Anaconda Navigator
* in Python Language
* Jupyter Notebook

**3.6 Feasibility study**

Feasibility study in the sense it's a practical approach of implementing the proposed model of system . Here for a machine learning projects .we generally collect the input from online websites and filter the input data and visualize them in graphical format and then the data is divided for training and testing . That training is testing data is given to the algorithms to predict the data .

1. First, we take cervical dataset.

2. Filter dataset according to requirements and create a new dataset which has attribute according to analysis to be done

3. Perform Pre-Processing on the dataset

4. Split the data into training and testing

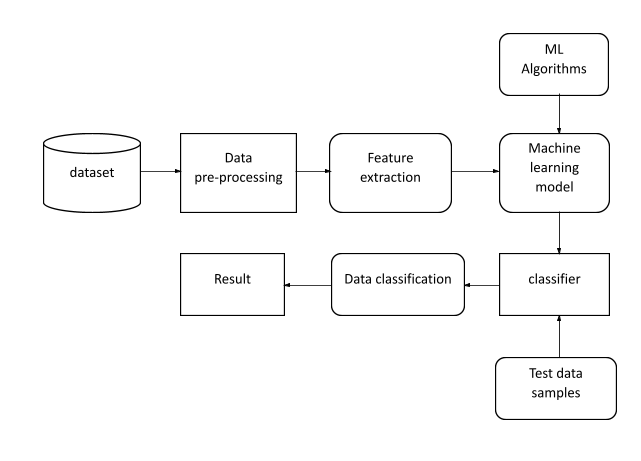
5. Train the model with training data then analyze testing dataset over classification algorithm

6. Finally you will get results as accuracy metrics.

7. Atlast we will creating web page using flask package with pickle python library file.

**CHAPTER 4: SYSTEM DESIGN**

**4.1 System architecture**

****

**4.2 Modules**

1. **DATA COLLECTION**
2. **DATA PRE-PROCESSING**
3. **FEATURE EXTRATION**
4. **EVALUATION MODEL**

**DATA COLLECTION**

Data collection is a process in which information is gathered from many sources which is later used to develop the machine learning models. The data should be stored in a way that makes sense for problem. In this step the data set is converted into the understandable format which can be fed into machine learning models.

Data used in this paper is a set of cervical cancer data with 15 features . This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called *labelled data*.

**DATA PRE-PROCESSING**

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are:

Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.

Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

**FEATURE EXTRATION**

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes.  Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

**EVALUATION MODEL**

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance.

Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs.

**Accuracy** is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

**Features:-**

* Development server and [debugger](https://en.wikipedia.org/wiki/Debugger)
* Integrated support for [unit testing](https://en.wikipedia.org/wiki/Unit_testing)
* [RESTful](https://en.wikipedia.org/wiki/Representational_state_transfer) request dispatching
* Uses Jinja templating
* Support for secure cookies (client side sessions)
* 100% [WSGI](https://en.wikipedia.org/wiki/Web_Server_Gateway_Interface) 1.0 compliant
* [Unicode](https://en.wikipedia.org/wiki/Unicode)-based
* Complete documentation
* [Google App Engine](https://en.wikipedia.org/wiki/Google_App_Engine) compatibility
* Extensions available to extend functionality

**Example:**

The following code shows a simple web application that displays "[Hello World](https://en.wikipedia.org/wiki/Hello_World)!" when visited:

**from** **flask** **import** Flaskapp = Flask(\_\_name\_\_)

@app.route("/")**def** hello() -> str:

**return** "Hello World"

**if** \_\_name\_\_ == "\_\_main\_\_":

app.run(debug=**False**)

**Templates** are files that contain static data as well as placeholders for dynamic data. A template is rendered with specific data to produce a final document. Flask uses the [Jinja](https://jinja.palletsprojects.com/templates/) template library to render templates.

In your application, you will use templates to render [HTML](https://developer.mozilla.org/docs/Web/HTML) which will display in the user’s browser. In Flask, Jinja is configured to autoescape any data that is rendered in HTML templates. This means that it’s safe to render user input; any characters they’ve entered that could mess with the HTML, such as < and > will be escaped with safe values that look the same in the browser but don’t cause unwanted effects.

Jinja looks and behaves mostly like Python. Special delimiters are used to distinguish Jinja syntax from the static data in the template. Anything between {{ and }} is an expression that will be output to the final document. {% and %} denotes a control flow statement like if and for. Unlike Python, blocks are denoted by start and end tags rather than indentation since static text within a block could change indentation.

Each page in the application will have the same basic layout around a different body. Instead of writing the entire HTML structure in each template, each template will extend a base template and override specific sections.



[**g**](https://flask.palletsprojects.com/en/2.1.x/api/#flask.g) is automatically available in templates. Based on if g.user is set (from load\_logged\_in\_user), either the username and a log out link are displayed, or links to register and log in are displayed. [**url\_for()**](https://flask.palletsprojects.com/en/2.1.x/api/#flask.url_for) is also automatically available, and is used to generate URLs to views instead of writing them out manually.

**4.3 UML Diagrams**

The Unified Modeling Language (UML) is used to specify, visualize, modify, construct and document the artifacts of an object-oriented software intensive system under development. UML offers a standard way to visualize a system's architectural blueprints, including elements such as:

* actors
* business processes
* (logical) components
* activities
* programming language statements
* database schemas, and
* Reusable software components.

UML combines best techniques from data modeling (entity relationship diagrams), business modeling (work flows), object modeling, and component modeling. It can be used with all processes, throughout the software development life cycle, and across different implementation technologies. UML has synthesized the notations of the Booch method, the Object-modeling technique (OMT) and Object-oriented software engineering (OOSE) by fusing them into a single, common and widely usable modeling language. UML aims to be a standard modeling language which can model concurrent and distributed systems.

**Sequence Diagram:**

Sequence Diagrams Represent the objects participating the interaction horizontally and time vertically. A Use Case is a kind of behavioral classifier that represents a declaration of an offered behavior. Each use case specifies some behavior, possibly including variants that the subject can perform in collaboration with one or more actors. Use cases define the offered behavior of the subject without reference to its internal structure. These behaviors, involving interactions between the actor and the subject, may result in changes to the state of the subject and communications with its environment. A use case can include possible variations of its basic behavior, including exceptional behavior and error handling.

**Activity Diagrams-:**

Activity diagrams are graphical representations of Workflows of stepwise activities and actions with support for choice, iteration and concurrency.In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

**Usecase diagram:**

UML is a standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems.

UML was created by Object Management Group (OMG) and UML 1.0 specification draft was proposed to the OMG in January 1997.

OMG is continuously putting effort to make a truly industry standard.

UML stands for Unified Modeling Language.

UML is a pictorial language used to make software blue prints

**Class diagram**

The class diagram is the main building block of object-oriented modeling. It is used for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for data modeling.[1] The classes in a class diagram represent both the main elements, interactions in the application, and the classes to be programmed.

In the diagram, classes are represented with boxes that contain three compartments:

The top compartment contains the name of the class. It is printed in bold and centered, and the first letter is capitalized.

The middle compartment contains the attributes of the class. They are left-aligned and the first letter is lowercase.

The bottom compartment contains the operations the class can execute. They are also left-aligned and the first letter is lowercase.

**CHAPTER 5: SYSTEM IMPLEMENTATION**

* 1. **Source code**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df=pd.read\_csv('Cancer.csv')

df.head()

df['Level']

df['Level'].replace('Medium','High',inplace=True)

df['Level']

df.tail()

df.isnull()

df.isnull().any()

df.isnull().sum()

sns.heatmap(df.isnull())

df['Level'].replace('Low',0,inplace=True)

df['Level'].replace('High',1,inplace=True)

df['Level']

df.head()

df.drop(['Patient Id'],axis=1,inplace=True)

df.head()

X=df.drop(['Level'],axis=1)

y=df['Level']

X

X

y

plt.figure(figsize = (10, 5))

sns.countplot(x='Smoking',data=df)

plt.figure(figsize = (10, 10))

sns.countplot(x='ChestPain',data=df)

plt.figure(figsize = (15, 10))

sns.boxplot(x='ChestPain',y='Age',data=df)

plt.figure(figsize = (15, 10))

sns.countplot(x='Smoking',hue='Gender',data=df)

plt.figure(figsize = (10, 10))

df['Age'].plot.hist()

plt.title('Lungs Cancer Detect in the Age')

plt.figure(figsize = (20, 10))

sns.countplot(x='Age',hue='Alcoholuse',data=df)

plt.xlim(0,10)

# Prediction Based On Classification Techniques

# Logistic Regression

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=1)

from sklearn.metrics import f1\_score

from sklearn.metrics import log\_loss

from sklearn.model\_selection import cross\_val\_score

from sklearn.linear\_model import LogisticRegression

model=LogisticRegression()

model.fit(X\_train,y\_train)

y\_pred=model.predict(X\_test)

from sklearn.metrics import confusion\_matrix

confusion\_matrix(y\_test,y\_pred)

from sklearn.metrics import accuracy\_score

accuracy\_score(y\_test,y\_pred)

y\_proba=model.predict\_proba(X\_test)

# KNN

# Modeling

from sklearn.neighbors import KNeighborsClassifier

k = 3

#Train Model and Predict

kNN\_model = KNeighborsClassifier(n\_neighbors=k).fit(X\_train,y\_train)

kNN\_model

# just for sanity chaeck

ypred = kNN\_model.predict(X\_test)

ypred[0:5]

X\_test

X\_test.shape

y\_test

X\_test.info()

from sklearn.neighbors import KNeighborsClassifier

from sklearn.model\_selection import cross\_val\_score

bestscore=0

list=[]

for k in range(3,12):

clf\_knn=KNeighborsClassifier(n\_neighbors=k)

scores=cross\_val\_score(clf\_knn,X,y,cv=10)

score=scores.mean()

list.append(score)

if score>bestscore:

bestscore=score

bestk=k

print("best k is",bestk,"Cross validation score is",bestscore)

from sklearn.neighbors import KNeighborsClassifier

k = 11

model1 = KNeighborsClassifier(n\_neighbors=k).fit(X\_train,y\_train)

model1

y\_pred = model1.predict(X\_test)

accuracy\_score(y\_test,y\_pred)

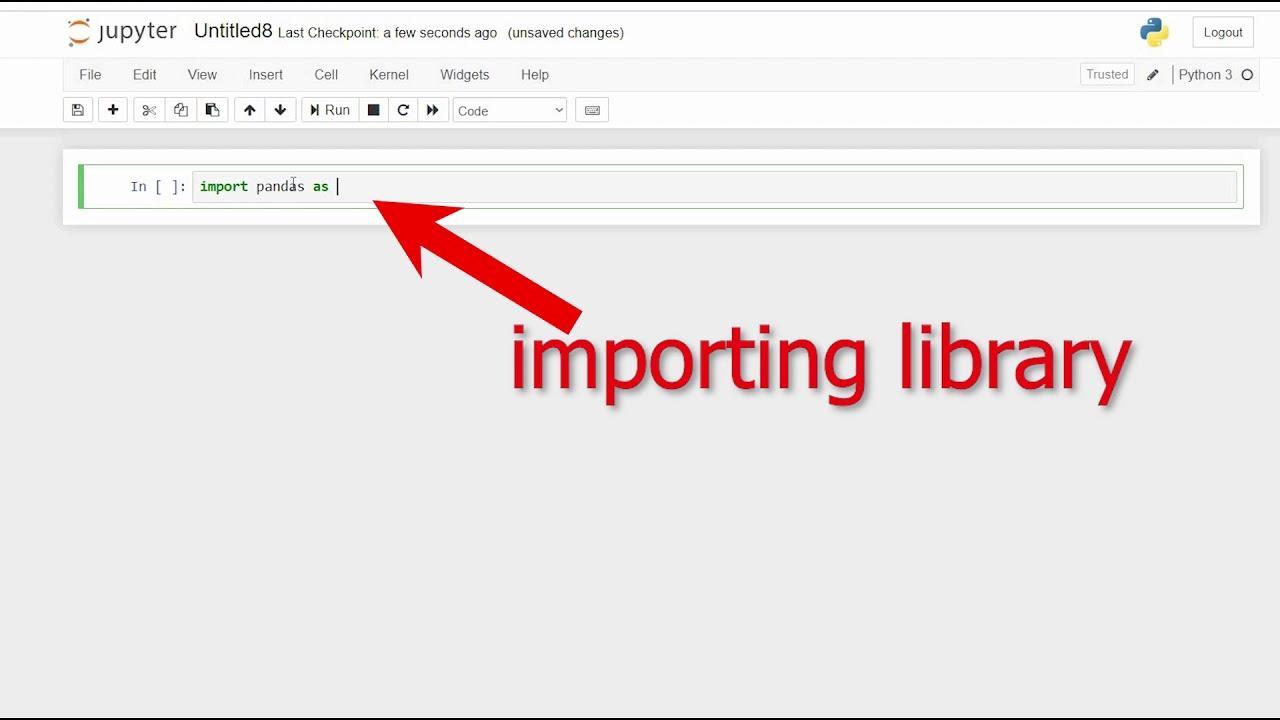
import pickle

pickle.dump(model1, open('model.pkl','wb'))

**CHAPTER 6: RESULTS**

**6.1 Screenshots**

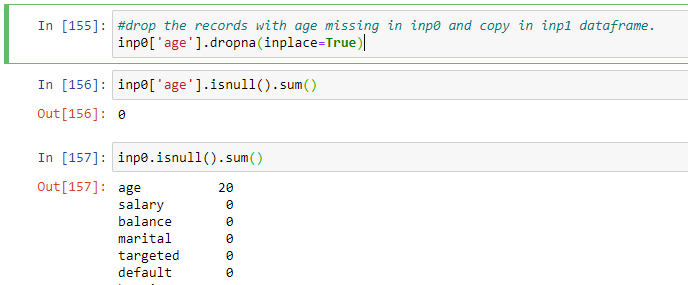
1. **Importing packages**



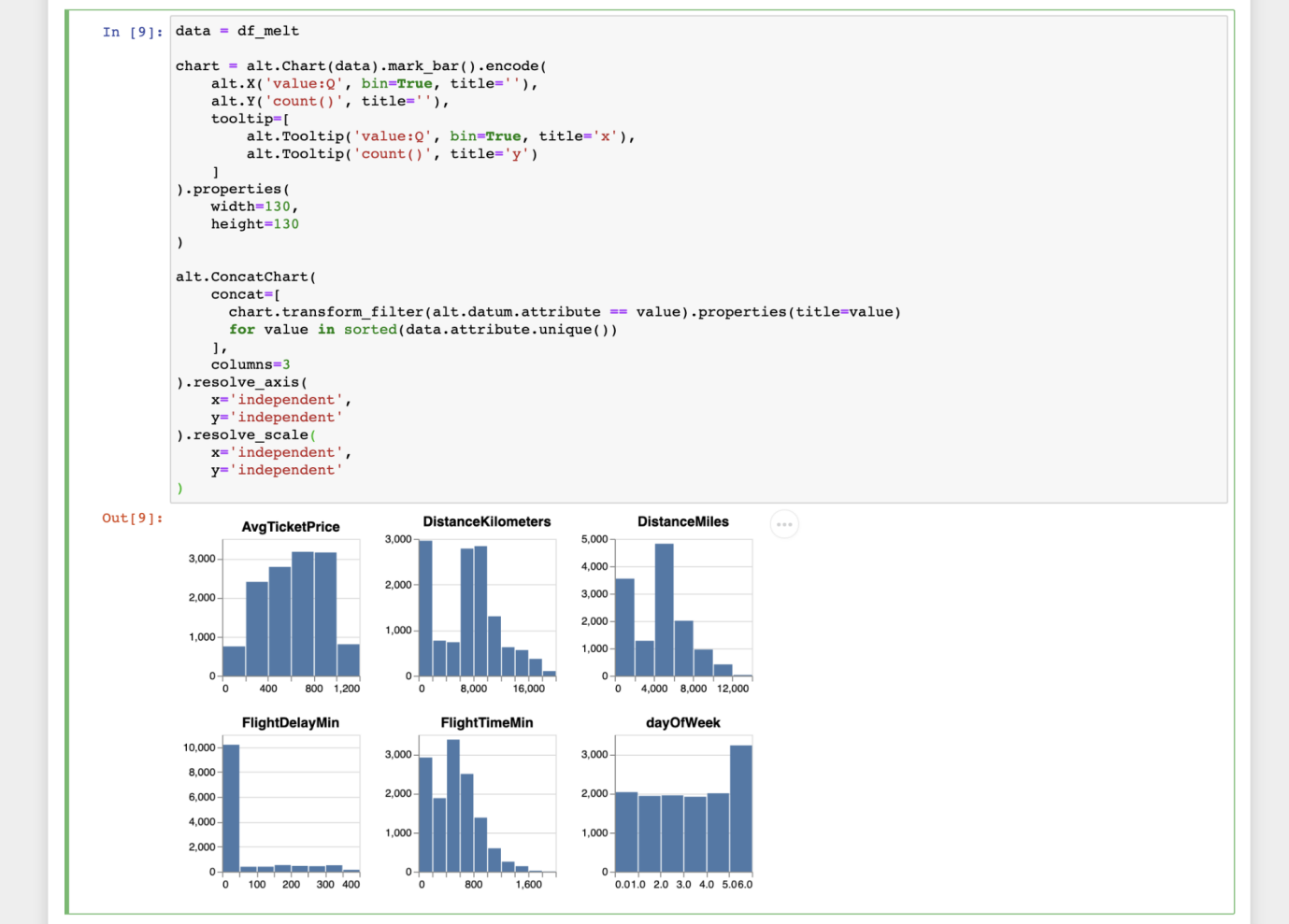
1. **Data Collection**



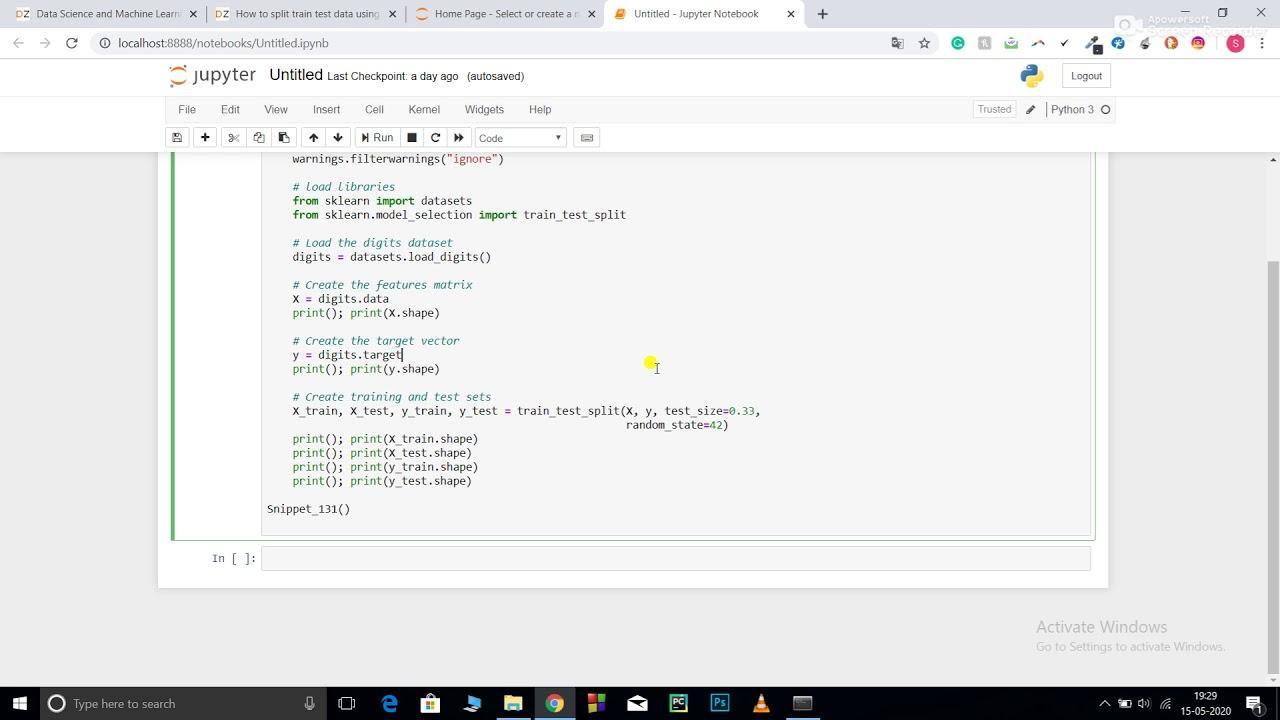
1. **Data preprocessing**



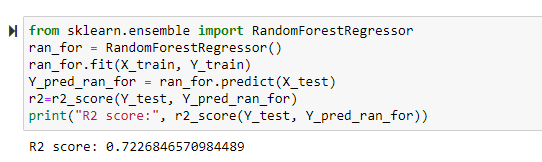
1. **Feature Extraction**



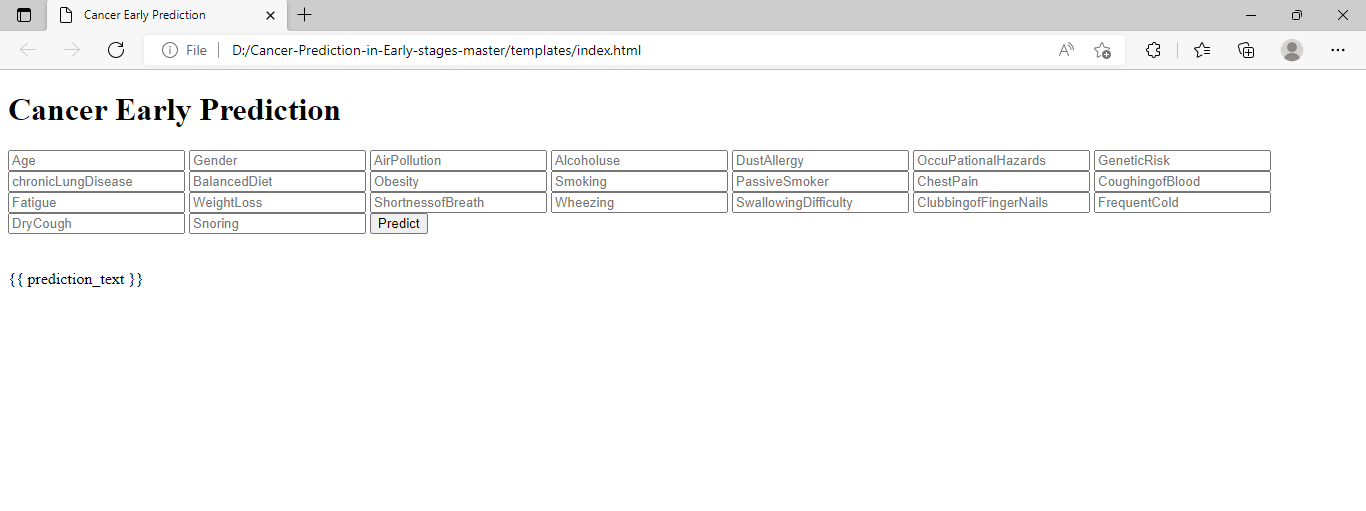
1. **Training and Testing**



1. **Evaluation model**

****

1. **Web page**



**CHAPTER 7: TESTING**

**7.1 Testing**

Software testing is an investigation conducted to provide stakeholders with information about the quality of the product or service under test. Software Testing also provides an objective, independent view of the software to allow the business to appreciate and understand the risks at implementation of the software. Test techniques include, but are not limited to, the process of executing a program or application with the intent of finding software bugs.

Software Testing can also be stated as the process of validating and verifying that a software program/application/product:

* Meets the business and technical requirements that guided its design and Development.
* Works as expected and can be implemented with the same characteristics.

**TESTING METHODS**

**Functional Testing**

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

* Functions: Identified functions must be exercised.
* Output: Identified classes of software outputs must be exercised.
* Systems/Procedures: system should work properly

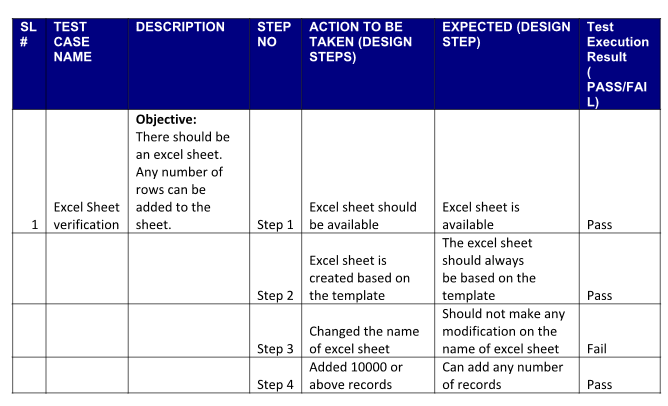
**Integration Testing**

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

Test Case for Excel Sheet Verification:

Here in machine learning we are dealing with dataset which is in excel sheet format so if any test case we need means we need to check excel file. Later on classification will work on the respective columns of dataset .

Test Case 1 :



**RESULTS**

The experiment is carried to predict the best fitting model for lung cancer patient data set which gives the greatest accuracy. The aim of conducting the experiment is to predict the early stage lung cancer in a person utilizing the various machine learning classification algorithms. During the course of the project the research on various papers helped in learning various aspects of the project. The study helped in understanding that SVM is not suitable for data having more noise along with large data sets. Also it is understood that with a large number of trees, the random forest algorithm becomes slow for real time prophecy and it acts as a predictive model like a black box. The study gives an idea about how K-NN works well only with balanced data and is sensitive to outliers. Furthermore, the working of artificial neural networks with numerical data makes it difficult to show the problem to the network and the network duration is unknown. This also helped in the consideration of techniques and features based on various factors.

**CHAPTER 8: CONCLUSION**

**8.1 Conclusion**

The experiment helped in learning various machine learning algorithms, its implementation, uses, pros and cons. It also gave a complete insight on preprocessing techniques and numerous activation functions used in neural networks. Building various models, training and testing them tells that the voting classifier is the best suitable model on this dataset for the early stage prediction of lung cancer in a patient.

**CHAPTER 9: FUTURE ENHANCEMENTS**

**9.1 Future enhancements**

There is a scope of enhancement in the experiment by using various other models like Logistic Regression, Extra trees classifier and boosting methods. The accuracy might vary depending on the volume of the data collected by applying various other preprocessing techniques and tools which can be explored further.

**REFERENCES**

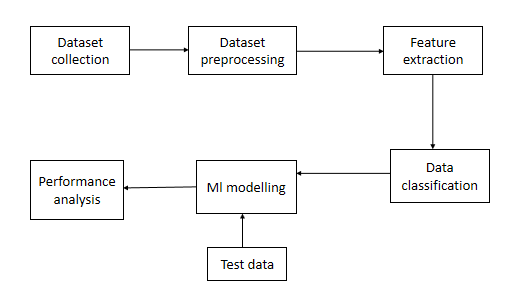
1. T. Babu, T. Singh, D. Gupta and S. Hameed, "Colon Cancer Detection in Biopsy Images for Indian Population at Different Magnification Factors Using Texture Features," 2017 Ninth International Conference on Advanced Computing (ICoAC), Chennai, 2017, pp. 192-197, doi: 10.1109/ICoAC.2017.8441173.
2. T. Babu, D. Gupta, T . Singh, S. Hameed, R. Nayar and R. Veena, "Cancer Screening On Indian Colon Biopsy Images Using Texture and Morphological Features," 2018International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 0175-0181, doi: 10.1109/ICCSP.2018.8524492.
3. Pandey S., M S., Shrivastava A. (2018) Data Classification Using Machine Learning Approach. In: Thampi S., Mitra S., Mukhopadhyay J., Li KC., James A., Berretti S. (eds) Intelligent Systems Technologies and Applications. ISTA 2017. Advances in Intelligent Systems and Computing, vol 683. Springer, Cham.
4. Va Dominic, Dr. Deepa Gupta, Sangita Khare, and Aggarwal, Ab, “Investigation of chronic disease correlation using data mining techniques”, in 2015 2nd International Conference on Recent Advances in Engineering and Computational Sciences, RAECS 2015, 2015.
5. S.Sivakumar, Dr.C.Chandrasekar. Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines, International Journal of Engineering and Technology (IJET)
6. Neesha Jothi, Nur Aini Abdul Rashid, Wahidah Husain, Data Mining in Healthcare - A Review, Procedia Computer Science 72:306-313 · December 2015
7. Dr. S. Senthil, B. Ayshwarya, Lung Cancer Prediction using Feed Forward Back Propagation Neural Networks with Optimal Features, International Journal of Applied Engineering Research, 13(1), pp.318-325.
8. Jennifer P. Cabrera , Lung Cancer Classification Tool using Microarray Data and Support Vector machines, In 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1-6). IEEE
9. S. Sasikala, M. Bharathi, B. R. Sowmiya, Lung Cancer Detection and Classification Using Deep CNN, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-2S December 2018
10. Syed Saba Raoof , M.A.Jabbar, Syed Aley Fathima, Lung Cancer Prediction using Machine Learning: A comprehensive Approach
11. Janee Alam, S., & Hossan, A. Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier. 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)
12. Kakeda, S., Moriya, J., Sato, H., Aoki, T., Watanabe, H., Nakata, Doi, K. 2004. Improved Detection of Lung Nodules on Chest Radiographs Using a Commercial Computer-Aided Diagnosis System. American Journal of Roentgen logy, 182(2), 505– 510. doi:10.2214/ajr.182.2.1820505.
13. Wasudeo Rahane, Yamini Magar, Himali Dalvi ,Anjali Kalane ,Satyajeet Jondhale ,Lung Cancer Detection Using Image Processing and Machine Learning HealthCare, Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India
14. Rohit Y. Bhalerao, Rachana K. Gaitonde, Harsh P. Jani ,Vinit Raut, A novel approach for detection of Lung Cancer using Digital Image Processing and Convolution Neural Networks, 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)

[15] Divya Chauhan, Varun Jaiswal , An Efficient Data Mining Classification Approach for Detecting Lung Cancer Disease.

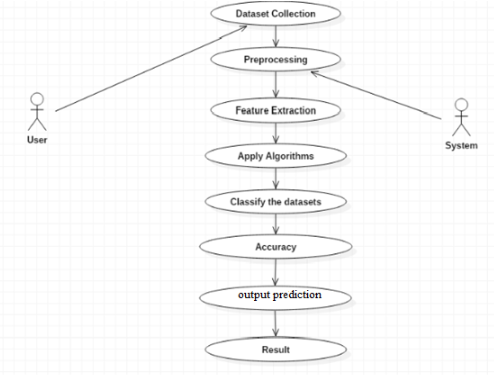
**LIST OF FIGURES**

**Name of the Figure**

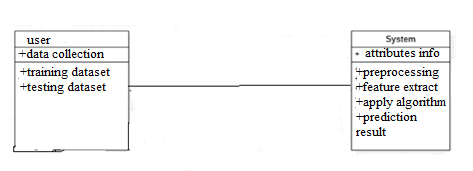
1. System architecture of the model

****

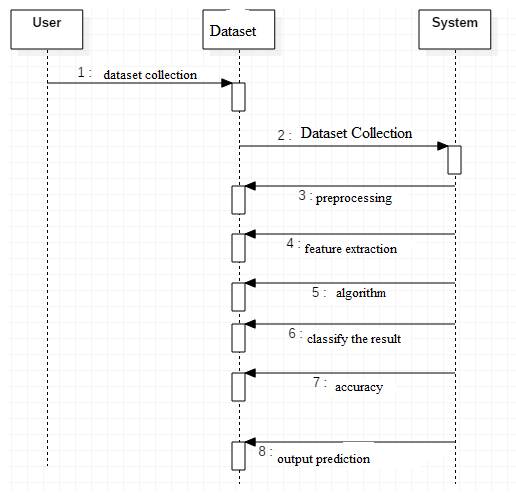
1. Use case Diagram



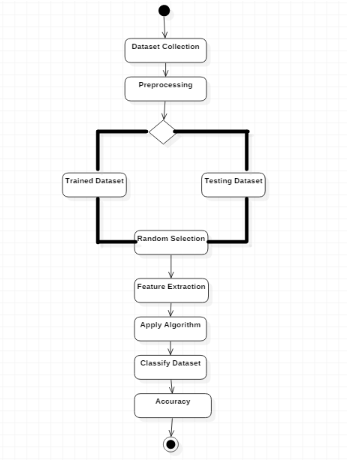
1. Class Diagram



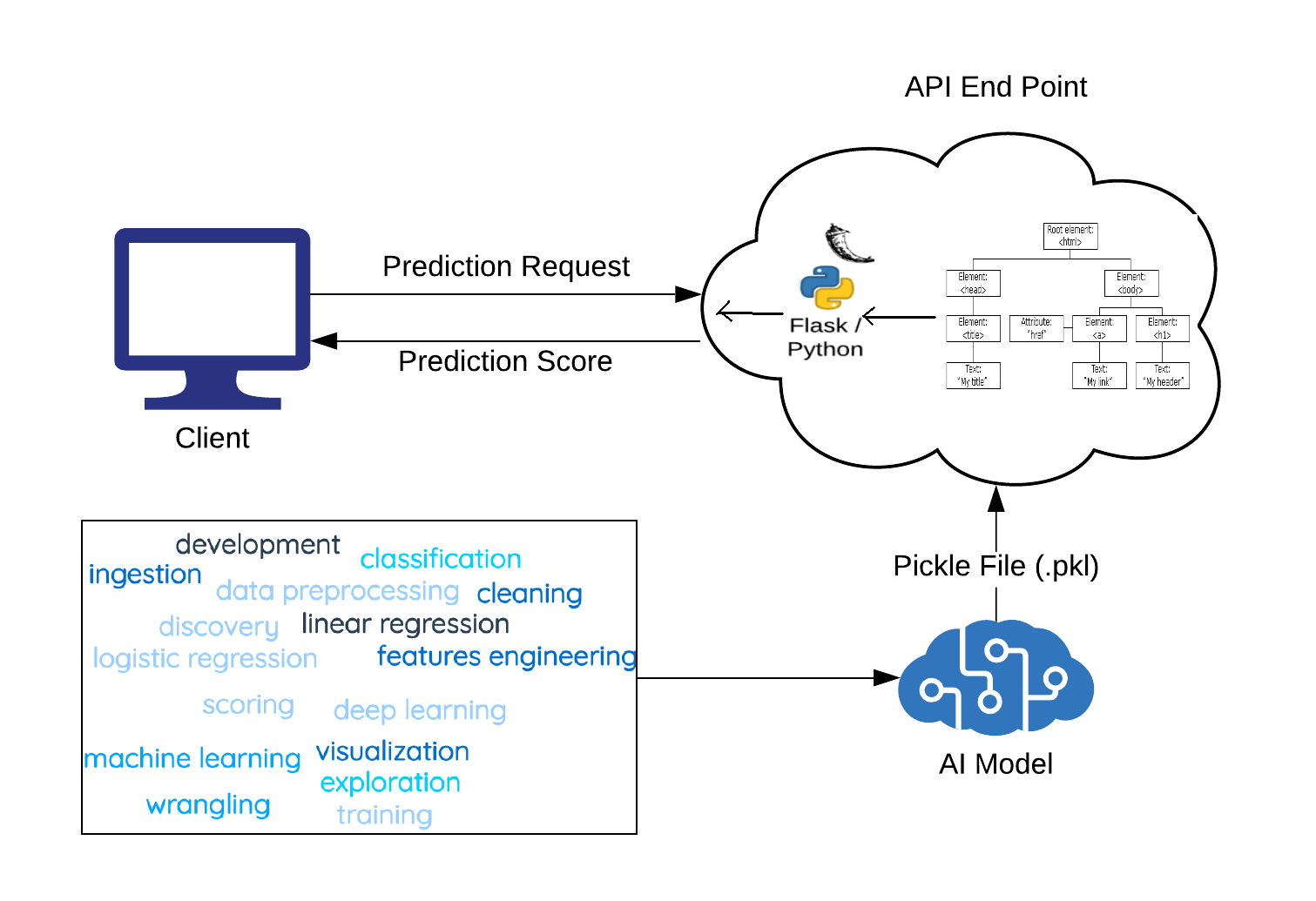
1. Sequence Diagram



1. Activity Diagram



1. Flask



**Domain Specification**

**MACHINE LEARNING**

Machine Learning is a system that can learn from example through self-improvement and without being explicitly coded by programmer. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results.

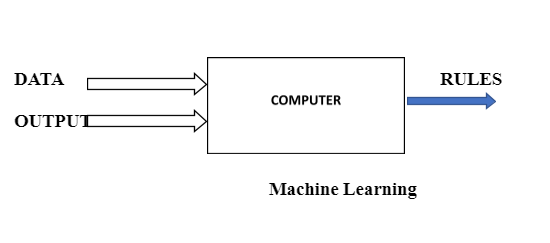
Machine learning combines data with statistical tools to predict an output. This output is then used by corporate to makes actionable insights. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input, use an algorithm to formulate answers.

A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

Machine learning is also used for a variety of task like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

**Machine Learning vs. Traditional Programming**

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.



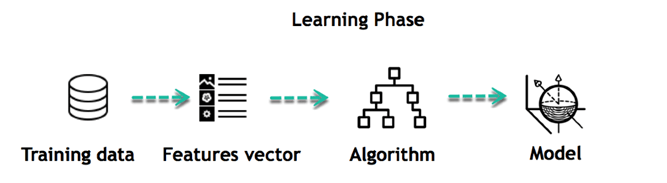
**Machine Learning**

How does Machine learning work?

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similar example, it can figure out the outcome. However, like a human, if its feed a previously unseen example, the machine has difficulties to predict.

The core objective of machine learning is the **learning**and **inference**. First of all, the machine learns through the discovery of patterns. This discovery is made thanks to the **data**. One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a **feature vector.** You can think of a feature vector as a subset of data that is used to tackle a problem.

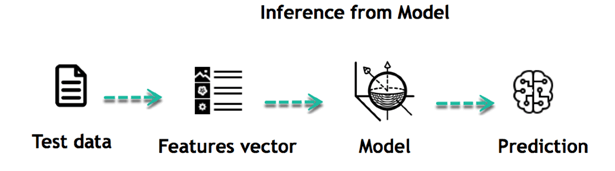
The machine uses some fancy algorithms to simplify the reality and transform this discovery into a **model**. Therefore, the learning stage is used to describe the data and summarize it into a model.



For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant: This is the model

*Inferring*

When the model is built, it is possible to test how powerful it is on never-seen-before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train again the model. You can use the model previously trained to make inference on new data.

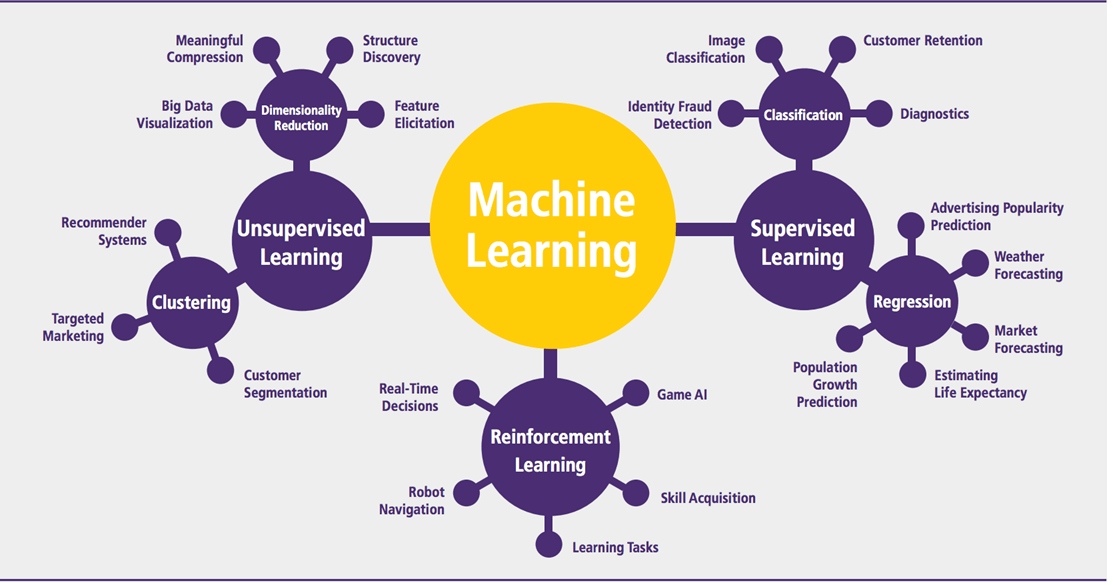


The life of Machine Learning programs is straightforward and can be summarized in the following points:

* Define a question
* Collect data
* Visualize data
* Train algorithm
* Test the Algorithm
* Collect feedback
* Refine the algorithm
* Loop 4-7 until the results are satisfying
* Use the model to make a prediction

Once the algorithm gets good at drawing the right conclusions, it applies that knowledge to new sets of data.

Machine learning Algorithms and where they are used?



Machine learning can be grouped into two broad learning tasks: Supervised and Unsupervised. There are many other algorithms

*Supervised learning*

An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output. For instance, a practitioner can use marketing expense and weather forecast as input data to predict the sales of cans.

You can use supervised learning when the output data is known. The algorithm will predict new data.

There are two categories of supervised learning:

| **Algorithm Name** | **Description** | **Type** |
| --- | --- | --- |
| **Linear regression** | Finds a way to correlate each feature to the output to help predict future values. | Regression |
| **Logistic regression** | Extension of linear regression that's used for classification tasks. The output variable 3is binary (e.g., only black or white) rather than continuous (e.g., an infinite list of potential colors) | Classification |
| **Decision tree** | Highly interpretable classification or regression model that splits data-feature values into branches at decision nodes (e.g., if a feature is a color, each possible color becomes a new branch) until a final decision output is made | Regression Classification |
| **Naive Bayes** | The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an event with the independent probability of each feature that can affect the event. | Regression Classification |
| **Support vector machine** | Support Vector Machine, or SVM, is typically used for the classification task. SVM algorithm finds a hyperplane that optimally divided the classes. It is best used with a non-linear solver. | Regression (not very common) Classification |
| **Random forest** | The algorithm is built upon a decision tree to improve the accuracy drastically. Random forest generates many times simple decision trees and uses the 'majority vote' method to decide on which label to return. For the classification task, the final prediction will be the one with the most vote; while for the regression task, the average prediction of all the trees is the final prediction. | Regression Classification |
| **AdaBoost** | Classification or regression technique that uses a multitude of models to come up with a decision but weighs them based on their accuracy in predicting the outcome | Regression Classification |
| **Gradient-boosting trees** | Gradient-boosting trees is a state-of-the-art classification/regression technique. It is focusing on the error committed by the previous trees and tries to correct it. | Regression Classification |

* Classification task
* Regression task

*Classification*

Imagine you want to predict the gender of a customer for a commercial. You will start gathering data on the height, weight, job, salary, purchasing basket, etc. from your customer database. You know the gender of each of your customer, it can only be male or female. The objective of the classifier will be to assign a probability of being a male or a female (i.e., the label) based on the information (i.e., features you have collected). When the model learned how to recognize male or female, you can use new data to make a prediction. For instance, you just got new information from an unknown customer, and you want to know if it is a male or female. If the classifier predicts male = 70%, it means the algorithm is sure at 70% that this customer is a male, and 30% it is a female.

The label can be of two or more classes. The above example has only two classes, but if a classifier needs to predict object, it has dozens of classes (e.g., glass, table, shoes, etc. each object represents a class)

***Regression***

When the output is a continuous value, the task is a regression. For instance, a financial analyst may need to forecast the value of a stock based on a range of feature like equity, previous stock performances, macroeconomics index. The system will be trained to estimate the price of the stocks with the lowest possible error.

*Unsupervised learning*

In unsupervised learning, an algorithm explores input data without being given an explicit output variable (e.g., explores customer demographic data to identify patterns)

You can use it when you do not know how to classify the data, and you want the algorithm to find patterns and classify the data for you

| **Algorithm** | **Description** | **Type** |
| --- | --- | --- |
| **K-means clustering** | Puts data into some groups (k) that each contains data with similar characteristics (as determined by the model, not in advance by humans) | Clustering |
| **Gaussian mixture model** | A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters | Clustering |
| **Hierarchical clustering** | Splits clusters along a hierarchical tree to form a classification system.  Can be used for Cluster loyalty-card customer | Clustering |
| **Recommender system** | Help to define the relevant data for making a recommendation. | Clustering |
| **PCA/T-SNE** | Mostly used to decrease the dimensionality of the data. The algorithms reduce the number of features to 3 or 4 vectors with the highest variances. | Dimension Reduction |

**Application of Machine learning**

**Augmentation**:

Machine learning, which assists humans with their day-to-day tasks, personally or commercially without having complete control of the output. Such machine learning is used in different ways such as Virtual Assistant, Data analysis, software solutions. The primary user is to reduce errors due to human bias.

**Automation**:

Machine learning, which works entirely autonomously in any field without the need for any human intervention. For example, robots performing the essential process steps in manufacturing plants.

**Finance Industry**

Machine learning is growing in popularity in the finance industry. Banks are mainly using ML to find patterns inside the data but also to prevent fraud.

**Government organization**

The government makes use of ML to manage public safety and utilities. Take the example of China with the massive face recognition. The government uses Artificial intelligence to prevent jaywalker.

**Healthcare industry**

Healthcare was one of the first industry to use machine learning with image detection.

**Marketing**

Broad use of AI is done in marketing thanks to abundant access to data. Before the age of mass data, researchers develop advanced mathematical tools like Bayesian analysis to estimate the value of a customer. With the boom of data, marketing department relies on AI to optimize the customer relationship and marketing campaign.

**Example of application of Machine Learning in Supply Chain**

Machine learning gives terrific results for visual pattern recognition, opening up many potential applications in physical inspection and maintenance across the entire supply chain network.

Unsupervised learning can quickly search for comparable patterns in the diverse dataset. In turn, the machine can perform quality inspection throughout the logistics hub, shipment with damage and wear.

For instance, IBM's Watson platform can determine shipping container damage. Watson combines visual and systems-based data to track, report and make recommendations in real-time.

In past year stock manager relies extensively on the primary method to evaluate and forecast the inventory. When combining big data and machine learning, better forecasting techniques have been implemented (an improvement of 20 to 30 % over traditional forecasting tools). In term of sales, it means an increase of 2 to 3 % due to the potential reduction in inventory costs.

**Example of Machine Learning Google Car**

For example, everybody knows the Google car. The car is full of lasers on the roof which are telling it where it is regarding the surrounding area. It has radar in the front, which is informing the car of the speed and motion of all the cars around it. It uses all of that data to figure out not only how to drive the car but also to figure out and predict what potential drivers around the car are going to do. What's impressive is that the car is processing almost a gigabyte a second of data.

Deep Learning

Deep learning is a computer software that mimics the network of neurons in a brain. It is a subset of machine learning and is called deep learning because it makes use of deep neural networks. The machine uses different layers to learn from the data. The depth of the model is represented by the number of layers in the model. Deep learning is the new state of the art in term of AI. In deep learning, the learning phase is done through a neural network.

**Reinforcement Learning**

Reinforcement learningis a subfield of machine learning in which systems are trained by receiving virtual "rewards" or "punishments," essentially learning by trial and error. Google's DeepMind has used reinforcement learning to beat a human champion in the Go games. Reinforcement learning is also used in video games to improve the gaming experience by providing smarter bot.

One of the most famous algorithms are:

* Q-learning
* Deep Q network
* State-Action-Reward-State-Action (SARSA)
* Deep Deterministic Policy Gradient (DDPG)

**Applications/ Examples of deep learning applications**

**AI in Finance:**The financial technology sector has already started using AI to save time, reduce costs, and add value. Deep learning is changing the lending industry by using more robust credit scoring. Credit decision-makers can use AI for robust credit lending applications to achieve faster, more accurate risk assessment, using machine intelligence to factor in the character and capacity of applicants.

Underwrite is a Fintech company providing an AI solution for credit makers company. underwrite.ai uses AI to detect which applicant is more likely to pay back a loan. Their approach radically outperforms traditional methods.

**AI in HR:**Under Armour, a sportswear company revolutionizes hiring and modernizes the candidate experience with the help of AI. In fact, Under Armour Reduces hiring time for its retail stores by 35%. Under Armour faced a growing popularity interest back in 2012. They had, on average, 30000 resumes a month. Reading all of those applications and begin to start the screening and interview process was taking too long. The lengthy process to get people hired and on-boarded impacted Under Armour's ability to have their retail stores fully staffed, ramped and ready to operate.

At that time, Under Armour had all of the 'must have' HR technology in place such as transactional solutions for sourcing, applying, tracking and onboarding but those tools weren't useful enough. Under armour choose **HireVue**, an AI provider for HR solution, for both on-demand and live interviews. The results were bluffing; they managed to decrease by 35% the time to fill. In return, the hired higher quality staffs.

**AI in Marketing:**AI is a valuable tool for customer service management and personalization challenges. Improved speech recognition in call-center management and call routing as a result of the application of AI techniques allows a more seamless experience for customers.

For example, deep-learning analysis of audio allows systems to assess a customer's emotional tone. If the customer is responding poorly to the AI chatbot, the system can be rerouted the conversation to real, human operators that take over the issue.

Apart from the three examples above, AI is widely used in other sectors/industries.

**Artificial Intelligence**

Difference between Machine Learning and Deep Learning

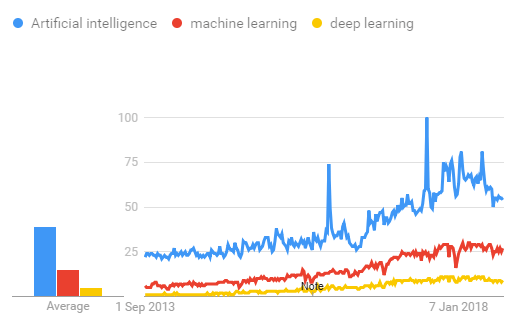
|  | **Machine Learning** | **Deep Learning** |
| --- | --- | --- |
| **Data Dependencies** | Excellent performances on a small/medium dataset | Excellent performance on a big dataset |
| **Hardware dependencies** | Work on a low-end machine. | Requires powerful machine, preferably with GPU: DL performs a significant amount of matrix multiplication |
| **Feature engineering** | Need to understand the features that represent the data | No need to understand the best feature that represents the data |
| **Execution time** | From few minutes to hours | Up to weeks. Neural Network needs to compute a significant number of weights |
| **Interpretability** | Some algorithms are easy to interpret (logistic, decision tree), some are almost impossible (SVM, XGBoost) | Difficult to impossible |

When to use ML or DL?

In the table below, we summarize the difference between machine learning and deep learning.

|  | **Machine learning** | **Deep learning** |
| --- | --- | --- |
| **Training dataset** | Small | Large |
| **Choose features** | Yes | No |
| **Number of algorithms** | Many | Few |
| **Training time** | Short | Long |

With machine learning, you need fewer data to train the algorithm than deep learning. Deep learning requires an extensive and diverse set of data to identify the underlying structure. Besides, machine learning provides a faster-trained model. Most advanced deep learning architecture can take days to a week to train. The advantage of deep learning over machine learning is it is highly accurate. You do not need to understand what features are the best representation of the data; the neural network learned how to select critical features. In machine learning, you need to choose for yourself what features to include in the model.



TensorFlow

the most famous deep learning library in the world is Google's TensorFlow. Google product uses machine learning in all of its products to improve the search engine, translation, image captioning or recommendations.

To give a concrete example, Google users can experience a faster and more refined the search with AI. If the user types a keyword a the search bar, Google provides a recommendation about what could be the next word.

Google wants to use machine learning to take advantage of their massive datasets to give users the best experience. Three different groups use machine learning:

* Researchers
* Data scientists
* Programmers.

They can all use the same toolset to collaborate with each other and improve their efficiency.

Google does not just have any data; they have the world's most massive computer, so TensorFlow was built to scale. TensorFlow is a library developed by the Google Brain Team to accelerate machine learning and deep neural network research.

It was built to run on multiple CPUs or GPUs and even mobile operating systems, and it has several wrappers in several languages like Python, C++ or Java.

In this tutorial, you will learn

**TensorFlow Architecture**

Tensor flow architecture works in three parts:

* Pre processing the data
* Build the model
* Train and estimate the model

It is called Tensor flow because it takes input as a multi-dimensional array, also known as **tensors**. You can construct a sort of **flowchart** of operations (called a Graph) that you want to perform on that input. The input goes in at one end, and then it flows through this system of multiple operations and comes out the other end as output.

This is why it is called TensorFlow because the tensor goes in it flows through a list of operations, and then it comes out the other side.

**Where can Tensor flow run?**

TensorFlow can hardware, and software requirements can be classified into

Development Phase: This is when you train the mode. Training is usually done on your Desktop or laptop.

Run Phase or Inference Phase: Once training is done Tensorflow can be run on many different platforms. You can run it on

* Desktop running Windows, macOS or Linux
* Cloud as a web service
* Mobile devices like iOS and Android

You can train it on multiple machines then you can run it on a different machine, once you have the trained model.

The model can be trained and used on GPUs as well as CPUs. GPUs were initially designed for video games. In late 2010, Stanford researchers found that GPU was also very good at matrix operations and algebra so that it makes them very fast for doing these kinds of calculations. Deep learning relies on a lot of matrix multiplication. TensorFlow is very fast at computing the matrix multiplication because it is written in C++. Although it is implemented in C++, TensorFlow can be accessed and controlled by other languages mainly, Python.

Finally, a significant feature of Tensor Flow is the Tensor Board. The Tensor Board enables to monitor graphically and visually what TensorFlow is doing.

**List of Prominent Algorithms supported by TensorFlow**

* Linear regression: tf. estimator  .Linear Regressor
* Classification :tf. Estimator .Linear  Classifier
* Deep learning classification: tf. estimator. DNN Classifier
* Booster tree regression: tf.estimator.BoostedTreesRegressor
* Boosted tree classification: tf.estimator.BoostedTreesClassifier

**PYTHON OVERVIEW**

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

* **Python is Interpreted:** Python is processed at runtime by the interpreter.You do not need to compile your program before executing it. This is similar to PERL and PHP.
* **Python is Interactive:** You can actually sit at a Python prompt and interactwith the interpreter directly to write your programs.
* **Python is Object-Oriented:** Python supports Object-Oriented style ortechnique of programming that encapsulates code within objects.
* **Python is a Beginner's Language:** Python is a great language for thebeginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

**History of Python**

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, Unix shell, and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

**Python Features**

Python's features include:

**Easy-to-learn:** Python has few keywords, simple structure, and a clearlydefined syntax. This 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 of the library is very portable andcross-platform compatible on UNIX, Windows, and Macintosh.

**Interactive Mode:** Python has support for an interactive mode which allowsinteractive testing and debugging of snippets of code.

**Portable:** Python can run on a wide variety of hardware platforms and has thesame interface on all platforms.

**Extendable:** You can add low-level modules to the Python interpreter. Thesemodules enable programmers to add to 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 andported 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 programsthan shell scripting.

Apart from the above-mentioned features, Python has a big list of good features, few are listed below:

* 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 large applications.
* It provides very high-level dynamic data types and supports dynamic type checking.
* IT supports automatic garbage collection.
* It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

**ANACONDA NAVIGATOR**

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, mac OS and Linux.

   Why use Navigator?

      In order to run, many scientific packages depend on specific versions of other          packages. Data scientists often use multiple versions of many packages, and use multiple environments to separate these different versions.

The command line program conda is both a package manager and an environment manager, to help data scientists ensure that each version of each package has all the dependencies it requires and works correctly.

Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages and update them, all inside Navigator.

**WHAT APPLICATIONS CAN I ACCESS USING NAVIGATOR**?

The following applications are available by default in Navigator:

* Jupyter  Lab
* Jupyter Notebook
* QT Console
* Spyder
* VS Code
* Glue viz
* Orange 3 App
* Rodeo
* RStudio

Advanced conda users can also build your own Navigator applications

How can I run code with Navigator?

The simplest way is with Spyder. From the Navigator Home tab, click Spyder, and write and execute your code.

You can also use Jupyter Notebooks the same way. Jupyter Notebooks are an increasingly popular system that combine your code, descriptive text, output, images and interactive interfaces into a single notebook file that is edited, viewed and used in a web browser.

What’s new in 1.9?

* Add support for **Offline Mode** for all environment related actions.
* Add support for custom configuration of main windows links.

Numerous bug fixes and performance enhancements.

**References**

[1] T. Babu, T. Singh, D. Gupta and S. Hameed, "Colon Cancer

Detection in Biopsy Images for Indian Population at Different Magnification

Factors Using Texture Features," 2017 Ninth International Conference on

Advanced Computing (ICoAC), Chennai, 2017, pp. 192-197, doi:

10.1109/ICoAC.2017.8441173.

[2] T. Babu, D. Gupta, T . Singh, S. Hameed, R. Nayar and

R. Veena, "Cancer Screening On Indian Colon Biopsy Images Using Texture

and Morphological Features," 2018

International Conference on Communication and Signal Processing

(ICCSP), Chennai, 2018, pp. 0175-0181, doi:

10.1109/ICCSP.2018.8524492.

[3] Pandey S., M S., Shrivastava A. (2018) Data Classification Using

Machine Learning Approach. In: Thampi S., Mitra S., Mukhopadhyay J., Li

KC., James A., Berretti S. (eds) Intelligent Systems Technologies and

Applications. ISTA 2017. Advances in Intelligent Systems and Computing,

vol 683. Springer, Cham.

[4] Va Dominic, Dr. Deepa Gupta, Sangita Khare, and Aggarwal, Ab,

“Investigation of chronic disease correlation using data mining techniques”,

in 2015 2nd International Conference on Recent Advances in Engineering

and Computational Sciences, RAECS 2015, 2015.

[5] S.Sivakumar, Dr.C.Chandrasekar. Lung Nodule Detection Using Fuzzy

Clustering and Support Vector Machines, International Journal of

Engineering and Technology (IJET)

[6] Neesha Jothi, Nur Aini Abdul Rashid, Wahidah Husain, Data Mining

in Healthcare - A Review, Procedia Computer Science 72:306-313 ·

December 2015

[7] Dr. S. Senthil, B. Ayshwarya, Lung Cancer Prediction using Feed

Forward Back Propagation Neural Networks with Optimal Features,

International Journal of Applied Engineering Research, 13(1), pp.318-325.

[8] Jennifer P. Cabrera , Lung Cancer Classification Tool using

Microarray Data and Support Vector machines, In 2015 6th International

Conference on Information, Intelligence,

Systems and Applications (IISA) (pp. 1-6). IEEE

[9] S. Sasikala, M. Bharathi, B. R. Sowmiya, Lung Cancer Detection and

Classification Using Deep CNN, International Journal of Innovative

Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075,

Volume-8 Issue-2S December 2018

[10] Syed Saba Raoof , M.A.Jabbar, Syed Aley Fathima, Lung Cancer

Prediction using Machine Learning: A comprehensive Approach

[11] Janee Alam, S., & Hossan, A. Multi-Stage Lung Cancer Detection

and Prediction Using Multi-class SVM Classifier. 2018 International

Conference on Computer, Communication, Chemical, Material and

Electronic Engineering (IC4ME2)

[12] Kakeda, S., Moriya, J., Sato, H., Aoki, T., Watanabe, H., Nakata,

Doi, K. 2004. Improved Detection of Lung Nodules on Chest Radiographs

Using a Commercial Computer-Aided Diagnosis System. American Journal

of Roentgen logy, 182(2), 505– 510.

doi:10.2214/ajr.182.2.1820505.

[13] Wasudeo Rahane, Yamini Magar, Himali Dalvi ,Anjali Kalane

,Satyajeet Jondhale ,Lung Cancer Detection Using Image Processing and

Machine Learning HealthCare, Proceeding of 2018 IEEE International

Conference on Current Trends toward Converging Technologies,

Coimbatore, India

[14] Rohit Y. Bhalerao, Rachana K. Gaitonde, Harsh P. Jani

,Vinit Raut, A novel approach for detection of Lung Cancer using Digital

Image Processing and Convolution Neural Networks, 2019 5th International

Conference on Advanced Computing & Communication Systems

(ICACCS)

[15] Divya Chauhan, Varun Jaiswal , An Efficient Data Mining

Classification Approach for Detecting Lung Cancer Disease.