**Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF**

**ABSTRACT :**

Cervical cancer is frequently a deadly disease, common in females. However, early diagnosis of cervical cancer can reduce the mortality rate and other associated complications. Cervical cancer risk factors can aid the early diagnosis. For better diagnosis accuracy, we proposed a study for early diagnosis of cervical cancer using reduced risk feature set and three ensemble-based classification techniques, i.e., extreme Gradient Boosting (XGBoost), AdaBoost, and Random Forest (RF) along with Firefly algorithm for optimization. Synthetic Minority Oversampling Technique (SMOTE) data sampling technique was used to alleviate the data imbalance problem. Cervical cancer Risk Factors data set, containing 32 risks factor and four targets (Hinselmann, Schiller, Cytology, and Biopsy), is used in the study. The four targets are the widely used diagnosis test for cervical cancer. The effectiveness of the proposed study is evaluated in terms of accuracy, sensitivity, specificity, positive predictive accuracy (PPA), and negative predictive accuracy (NPA). Moreover, Firefly features selection technique was used to achieve better results with the reduced number of features. Experimental results reveal the significance of the proposed model and achieved the highest outcome for Hinselmann test when compared with other three diagnostic tests. Furthermore, the reduction in the number of features has enhanced the outcomes. Additionally, the performance of the proposed models is noticeable in terms of accuracy when compared with other benchmark studies for cervical cancer diagnosis using reduced risk factors data set.

INTRODUCTION :

Cervical cancer is one of the commonly occurring types of cancer in females and mostly develops during their midlives (35 years–44 years) . This type of cancer can be fatal as it does not show clear symptoms in its early stages. Symptoms usually appear in late stages, where it could have spread to other organs like bones, liver, lymph nodes, and lungs. One of the early signs of cervical cancer is when the tube that carries urine from the kidney is blocked. Other late symptoms that can appear are vaginal bleeding, pelvic pain, weight loss, and leg pain .

The risk factors that lead to the development of cervical cancer are hormones containing medicines, birth control pills, smoking, and the number of pregnancies. However, it is believed that human papilloma virus (HPV) is the major factor in developing cervical cancer . HPV is a common sexually transmitted infection; it is usually harmless, but sometimes it may lead to cancer . HPV infection becomes at a higher risk of getting cervical cancer. Furthermore, the probability of getting cervical cancer increases if one possesses more than one risk factor. As the cancer does not show signs in its early stages, regular checkups are required especially for those who have the risk factors. In the developing countries, lack of medical equipment and the cost of conducting checkups could also be a burden. With the advent and advancement of machine learning, it has become possible to find robust solutions for early diagnosis of cancer cases using data-driven approaches.

Various studies have contributed to the field of cervical cancer diagnosis using several classification techniques by using different types of data such as clinical-based, image, and genetic-based data. In our study, we used clinical cervical risk factor data. Two similar studies were conducted by Wu and Zhou and Abdoh et al. they performed the comparative analysis of two feature selection techniques, namely, recursive feature elimination (RFE) and Principal Component Analysis (PCA). The first study used Support Vector Machine (SVM), and the other study used Random Forest (RF). Both studies used the same number of features. Although the data suffered from imbalance, an oversampling was applied to the data in and SMOTE was used in . Both studies identified two risk factors to be removed such as time since the first and last diagnosis of STDs (sexually transmitted diseases), due to a lot of missing entries. Furthermore, the study discovered that less computational cost was an advantage given by both SVM-PCA and SVM-RFE, whereas high computational cost is a limitation to the SVM model. Moreover, STDs, intrauterine device (IUD), hormonal contraceptives, and first sexual intercourse were identified as the highly relevant features . Overall, the outcome of both the studies showed that using 30 features produced highest results. Furthermore, it was found that the SMOTE-RF model performed well for all targets.

Similarly, Lu et al. and Karim and Neehal used ensemble models to estimate the risk of cervical cancer. Both studies performed data cleaning mechanism to replace missing values. The former study used an ensemble classifier with voting strategy using a combination of a private and public data set. The private data set contains 472 records taken from Chinese hospital. The public data set was obtained from the UCI repository; 14 features were used. The private data set was collected using questionnaire. The results revealed that voting ensemble classifier produced better results when compared to Linear Regression, Decision Tree (DT), Multilayer Perceptron (MLP), SVM, and K-NN classifiers. On the other hand, Karim and Neehal study used DT, MLP, and SVM using sequential Minimal Optimization (SMO) and K nearest neighbor (KNN) techniques. Experiments showed that SMO has a better performance in terms of accuracy, precision, recall, and F-measure. Similarly, Ul-Islam et al. used DT, RF, Logistic Model Tree, and ANN for cervical cancer detection. Apriori algorithm was used to identify features that strongly relate to cancer. The study found that age, number of sexual partners, hormonal contraceptives, number of pregnancies, and first sexual intercourse are significant risk factors. Results indicated that RF produced best outcome when compared to the other models.

Al-Wesabi et al. conducted a comparison between different machine learning classifiers such as Gaussian Naïve Bayes (GNB), KNN, DT, LR, and SVM. The outcome of the classifiers was not satisfactory due to the data imbalance. To resolve this problem, undersampling, oversampling, and SMOTETomek were applied. Oversampling had the best result among all three methods. Moreover, a Sequential Feature Selector was applied with both forward and backward versions. Both the Sequential Forward Feature Selector (SFS) and Sequential Backward Feature Selector (SBS) enhanced the performance of the prediction with an accuracy of 95%. After selecting the common features between DT and KNN, the accuracy exceeded 97% for the DT. The results revealed that age, first sexual intercourse, number of pregnancies, smoking, hormonal contraceptives, and STDs: genital herpes were the main predictive features.

Similarly, several studies have been made using deep learning and transfer learning for cervical cancer diagnosis. Fernandes et al. and Adem et al. used deep learning and showed significant outcome in terms of diagnosis accuracy. The study used a loss function that provides a supervised optimization of dimensionality reduction and classification models. The study indicated that it can be useful in examining records of patients if the Biopsy and perhaps other testing results are absent and are capable of classifying successfully whether they have cervical cancer or not. On the other hand, the researchers in used a deep neural network model with softmax function to classify the data sets. The performance of the softmax function with stacked autoencoder was compared with the other machine learning methods (DT, KNN, SVM, Feed Forward NN, and Rotation Forest models). It was found that the softmax function with a stacked autoencoder model produced better outcome classification rate of 97.8%.

Similarly, Fernandes et al. applied transfer learning with partial observability for cancer screenings. The limitation of the study was that several patients were resisting answering some questions for privacy concerns. Challenges were also faced in defining quality as there are multiple readings and it started relying on human preference. Therefore, as an alternative of an ordinal scale, a simple binary scheme was used. Nevertheless, the model performance was considerable.

Conclusively, the finding made after the above-mentioned literature is that the data set found at UCI repository had several missing values; therefore, previous studies have removed at least 2 features. Missing values were due to patient’s concerns regarding their privacy. After removing 2 features due to huge missing value, SVM-PCA seemed to provide satisfactory performance. However, SMO and SMOTE-RF were amongst the best performing models. Another approach to deal with the imbalance in UCI cervical risk factor data set was using oversampling. Deep learning proved to be effective, especially where the Biopsy and possibly other screening results are absent. Age, first sexual intercourse, number of pregnancies, smoking, hormonal contraceptives, IUD, STDs, STDs: genital warts, or HPV infections were identified as the top key features. The significant outcomes made by the machine learning classifiers motivate the need for further investigation and enhancement of the outcomes for the prediction of cervical cancer.

In this study, three ensemble-based classifiers extreme Gradient Boosting, Ada Boost, and RF are used to classify cervical cancer. Cervical Cancer Risk factor data set from UCI machine learning repository was collected at “Hospital Universitario de Caracas” in Caracas, Venezuela . In addition to the importance of correctly classifying cancerous and noncancerous cases, it is also essential to identify key risk factors that contribute to developing cancer. Nature-inspired Firefly feature selection and optimization algorithm was applied. Furthermore, the Synthetic Minority Oversampling Technique (SMOTE) is used to balance the classes of the data as it suffers greatly from imbalanced problem.

EXICITING SYSTEM :

Dataset Description :

cervical cancer risk factors data set used in the study was collected at “Hospital Universitario de Caracas” in Caracas, Venezuela and is available on the UCI Machine Learning repository . It consists of 858 records, with some missing values, as several patients did not answer some of the questions due to privacy concerns. the data set contains 32 risk factors and 4 targets, i.e., the diagnosis tests used for cervical cancer. It contains different categories of feature set such as habits, demographic information, history, and Genomic medical records. Features such as age, Dx: Cancer, Dx: CIN, Dx: HPV, and Dx features contains no missing values. Dx: CIN is a change in the walls of cervix and is commonly due to HPV infection; sometimes, it may lead to cancer if it is not treated properly. However, Dx: cancer variable is represented if the patient has other types of cancer or not. Sometimes, a patient may have more tha the cervical cancer risk factors data set used in the study was collected at “Hospital Universitario de Caracas” in Caracas, Venezuela and is available on the UCI Machine Learning repository . It consists of 858 records, with some missing values, as several patients did not answer some of the questions due to privacy concerns. the data set contains 32 risk factors and 4 targets, i.e., the diagnosis tests used for cervical cancer. It contains different categories of feature set such as habits, demographic information, history, and Genomic medical records. Features such as age, Dx: Cancer, Dx: CIN, Dx: HPV, and Dx features contains no missing values. Dx: CIN is a change in the walls of cervix and is commonly due to HPV infection; sometimes, it may lead to cancer if it is not treated properly. However, Dx: cancer variable is represented if the patient has other types of cancer or not. Sometimes, a patient may have more than one type of cancer. In the data set, some of the patients do not have cervical cancer, but they had the Dx: cancer value true. +erefore, it is not used as a target variable.he data set, some of the patients do not have cervical cancer, but they had the Dx: cancer value true. +erefore, it is not used as a target variable.

Table 1 presents a brief description of each feature with the type. Cervical cancer diagnosis usually requires several tests; this data contains the widely used diagnosis tests as the target. Hinselmann, Schiller, Cytology, and Biopsy are four widely used diagnosis tests for cervical cancer. Hinselmann or Colposcopy is a test that examines the inside of the vagina and cervix using a tool that magnifies the tissues to detect any anomalies . Schiller is a test in which a chemical substance called iodine is applied to the cervix, where it stains healthy cells into brown color and leaves the abnormal cells uncolored, while cytology is a test that examines body cells from uterine cervix for any cancerous cells or other diseases. And Biopsy refers to the test where a small part of cervical tissue is examined under a microscope. Most Biopsy tests can make significant diagnosis.

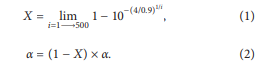
Dataset Preprocessing :

the data set suffers from a huge number of missing values; 24 features out of the 32 contained missing values. Initially, the features with the huge percentage of missing values were removed. STDs: Time since first diagnosis and STDs: Time since last diagnosis features were removed since they have 787 missing values (see Table 2), which is more than half of the data. However, the data imputation was performed for the features with fewer numbers of missing values. +e most frequent value technique was used to impute the remaining missing values. Additionally, the data set also suffers from huge class imbalance. the data set target labels were imbalanced with 35 for the Hinselmann, 74 for Schiller, 44 for Cytology, and 55 Biopsy out of the 858 records as shown in Figure 1. SMOTE was used to deal with class imbalance. SMOTE works by oversampling the minority class by generating new synthetic data for minority instances based on nearest neighbors using the Euclidean Distance between data points . Figure 1 shows the number of records per class labels in the data set.

Firefly Feature Selection :

Dimensionality reduction is one of the effective ways to select the features that improve the performance of the supervised learning model. In the study, we adopted nature-inspired algorithm Firefly for selecting the features that better formulate the problem. Firefly was proposed by Yang and was initially proposed for the optimization. Metaheuristic Firefly algorithm is inspired by fireflies’ and flash lightening capability of a fly. It is a population-based optimization algorithm to find the optimal value or parameter for a target function. In this technique, each fly is pulled out by the glow intensity of the nearby flies. If the intensity of the gleam is extremely low at some point, then the attraction will be declining. Firefly used three rules; that is, (a) all the flies should be of the same gender; (b) the criteria of attractiveness depend upon the intensity of the glow; (c) target function will generate the gleam of the firefly. the flies with less glow will move towards the flies with brighter glow. the brightness can be adjusted using objective function. the same idea is implemented in the algorithm to search the optimal features that can better fit the training model. Firefly is more computationally economical and produced better outcome in feature selection when compared with other metaheuristic techniques like genetic algorithms and particle swarm optimization . the time complexity of firefly is O(n2t) . It uses the light intensity to select the features. Highly relevant features are represented as the features with high intensity light.

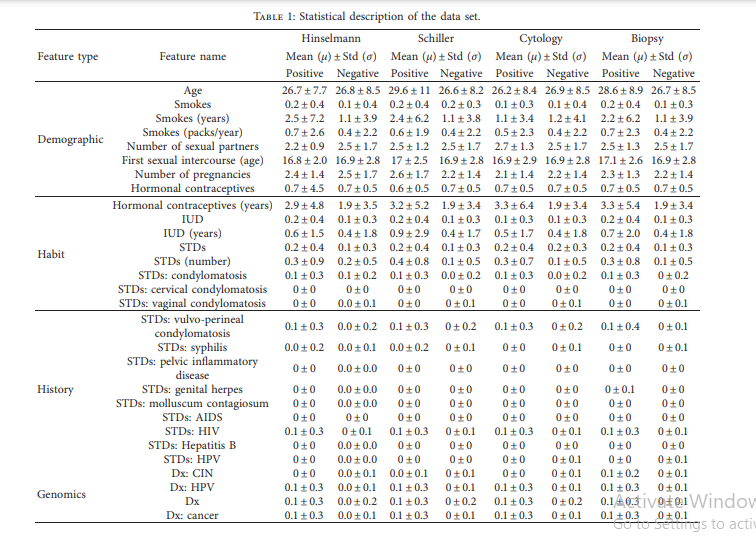
For feature selection, initially, some fireflies will be generated, and each fly will randomly assign the weights to all features. In our study, we generated 50 number of flies (n = 50). +e dimension of the data set is 30. Furthermore, the lower bound was set to − 50, while the upper bound is equal to 50. the maximum generations were 500. Additionally, α (alpha) was initially set to 0.5 and in every subsequent iteration, we used the and to update α (alpha) value.



However, the gamma (c) was set to 1. the number of features selected using Firefly for Hinselmann was 15, for Schiller 13 features, for Cytology 11 features, and 11 features for Biopsy, respectively.

Ensemble-Based Classification Methods :

ensemble-based classification techniques such as Random Forest, Extreme Gradient Boosting, and Ada Boost were used to train the model. the description of these techniques is discussed in the section below.



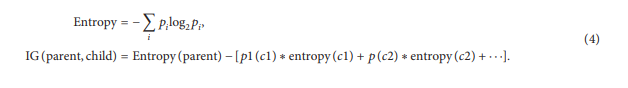
Random Forest :

Random Forest (RF) was first proposed by Breiman in 2001 . Random forest is an ensemble model that uses decision tree as individual model and bagging as ensemble method. It improves the performance of decision tree by adding many trees to reduce the overfitting in the decision tree. RF can be used for both classification and regression. RF generates a random forest that contains decision trees and gets a prediction from each one of them and then selects the best solution with the maximum votes .

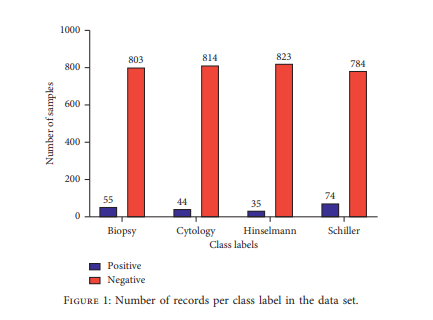
When training a tree, it is important to measure how much each feature decreases the impurity, as the decrease in the impurity indicates the significance of the feature. the tree classification result depends on the impurity measure used. For classification, the measures for impurity are either Gini impurity or information gain and for regression, and the measure for impurity is variance. Training decision tree consists of iteratively splitting the data. Gini impurity decides the best split of the data using the formula.



where p (i) is the probability of selecting a datapoint with class; i.e., Information gain (IG) is also another measure to decide the best split of the data depending on the gain of each feature. the formula that calculates the information gain is given in the following equation:where p (i) is the probability of selecting a datapoint with class; i.e., Information gain (IG) is also another measure to decide the best split of the data depending on the gain of each feature. the formula that calculates the information gain is given in the following equation:where p (i) is the probability of selecting a datapoint with class; i.e., Information gain (IG) is also another measure to decide the best split of the data depending on the gain of each feature. +e formula that calculates the information gain is given in the following equation:



Extreme Gradient Boosting :



eXtreme Gradient Boosting (XGBoost) is a tree-based ensemble technique . XGBoost can be used for classification, regression, and ranking problems. XG boosting is a type of gradient boosting. Gradient Boosting (GB) is a boosting ensemble technique that makes predicators sequentially instead of individually. GB is a method that produces a strong classifier by combining weak classifiers . +e goal of the GB is building an iterative model that optimizes a loss function. It pinpoints the failings of weak learners by using gradients in the loss function

F:\7.PNG

where e denotes the error term. loss function measures how good is the model at fitting the underlying data. the loss function depends on the optimization goal, for regression is a measure of the error between the true and predicated values, whereas, for classification, it measures the how good is a model at classifying cases correctly .this technique takes less time and less iterations, since predictors are learning from the past mistakes of the other predictors.

GB works by teaching a model C to predict values of the form

F:\8.PNG

By minimizing a loss function, e.g., MSE

F:\9.PNG

where i iterates over a training set of size n of true values of the target variable yyʹ = estimated values of C (x) y = true values & n =number of instances in y. Considering a GB model with M phases and m as a single phase being (1 ≤ m ≤M), to improve some deficient model Fm, a new estimator hm (x) is added. Therefore

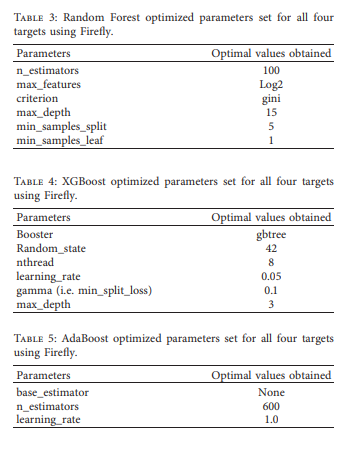


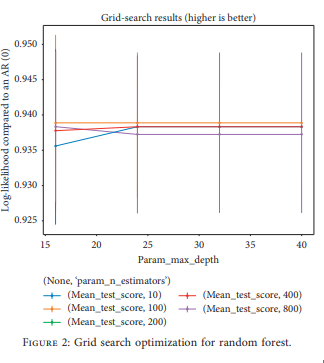
Estimator h will be fitted to Y − Fm(x), which is the difference between the true value and the predicated value, i.e., the residual. thus, we attempt to adjust the errors of the previous model (Fm)

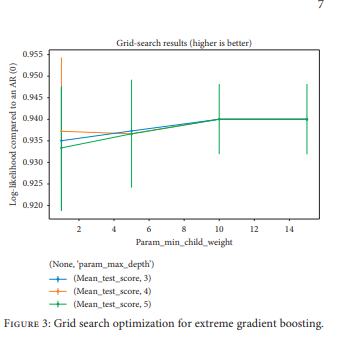
XGBoost is better than Ada boost in terms of speed and performance. It is highly scalable and runs 10 times faster as compared to the other traditional single machine learning algorithms. XGBoost handles the sparse data and implements several optimization and regularization techniques. Moreover, it also uses the concept of parallel and distributed computing.

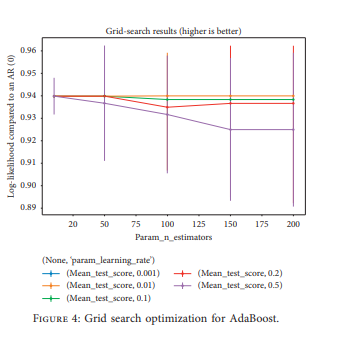
PROPOSED SYSTEM :

the model was implemented in Python language 3.8.0 release using Jupyter Notebook environment. Ski-learn library was used for the classifiers along with other needed built-in tools, while separate library (xgboost 1.2.0) was used for XGBoost ensemble. +ere is K-fold cross validation with K =10 for partitioning the data into training and testing. Five evaluation measures such as accuracy, sensitivity (recall), specificity (precision), positive predictive accuracy (PPA), and negative predictive accuracy (NPA) were used. Sensitivity and specificity are focused more during the study due to the application of the proposed model. Accuracy denotes the percentage of correctly classified cases, sensitivity measures the percentage of positives cases that were classified as positives, and specificity refers to the percentage of negative cases that were classified as negatives. Moreover, the criteria for the selection of the performance evaluating. measures depend upon the measures used in the benchmark studies. Two sets of experiments were conducted for each target using selected features by using Firefly feature selection algorithm and 30 features for four targets. +e SMOTE technique was applied to generate synthetic data. +e results of model are presented in section below



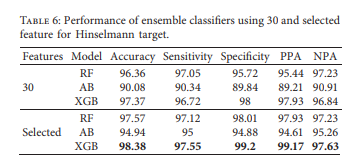






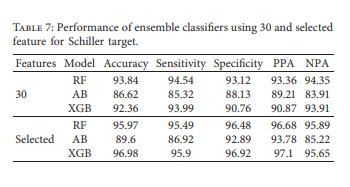
Hinselmann :

Table 6 presents the accuracy, sensitivity, specificity, PPA, and NPA for the RF, AdaBoost, and XGBoost models, respectively, using SMOTE for Hinselmann test target class. +e number of selected features for Hinselmann was 15. XGBoost outperformed the other classifiers for both feature sets. However, the performance of XGBoost with selected feature is better when compared with 30 features. +e model produces an accuracy of 98.83, sensitivity of 97.5, specificity of 99.2, PPA of 99.17, and NPA of 97.63, respectively



Schiller :

Table 7 presents the outcomes for the Schiller test. Like Hinselmann target, XGBoost with selected features outperformed that of Schiller, respectively. However, the outcomes achieved by the model for Schiller are lower when compared with Hinselmann target class. the performanceof RF and XGBoost is similar with selected feature for Schiller with a minor difference. the number of features selected by Firefly for Schiller was 13.



HARDWARE & SOFTWARE REQUIREMENTS:

HARD REQUIRMENTS **:**

* System    :   Pentium IV 2.4 GHz.
* Hard Disk  :   40 GB.
* Floppy Drive :   1.44 Mb.
* Monitor   :   15 VGA Colour.
* Mouse    :   Logitech.
* Ram    :   512 MB.

SOFTWARE REQUIRMENTS :

* Operating system   : Windows 7 Professional.
* Coding Language  : Java, Swing
* Front End Tool : Netbeans 7.0
* Database : My Sql .
* Server : Apache Tomcat 7.0.14.0
* Back End Tool : SQLYog

**3.4 SYSTEM STUDY**

**FEASIBILITY STUDY**

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

* ECONOMICAL FEASIBILITY
* TECHNICAL FEASIBILITY
* SOCIAL FEASIBILITY

**ECONOMICAL FEASIBILITY**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

### TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

**SOCIAL FEASIBILITY**

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

**4.SYSTEM DESIGN**

**4.1 UML DIAGRAMS**

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

**GOALS:**

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

**USE CASE DIAGRAM:**

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



**CLASS DIAGRAM:**

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

****

**SEQUENCE DIAGRAM:**

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

****

**ACTIVITY DIAGRAM:**

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.



**4.3 IMPLEMENTATION:**

**MODULES:**

1.Add Product Details

To build project I used some sample products image to train product identification models

2.Train Model

In this Module screen train model generated with 100% accuracy and now show product to web cam.

3. Add/Remove Product from basket

To allow application to identify product image and then show in text area and if we again show same product then application will remove from text area

**5.SOFTWARE ENVIRONMENT**

# What is Python :-

Below are some facts about Python.

Python is currently the most widely used multi-purpose, high-level programming language.

Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.

Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.

Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard library which can be used for the following –

* + [Machine Learning](https://www.geeksforgeeks.org/machine-learning/)
  + GUI Applications (like Kivy, Tkinter, PyQt etc. )
  + Web frameworks like Django (used by YouTube, Instagram, Dropbox)
  + Image processing (like Opencv, Pillow)
  + Web scraping (like Scrapy, BeautifulSoup, Selenium)
  + Test frameworks
  + Multimedia

### Advantages of Python :-

Let’s see how Python dominates over other languages.

#### 1. Extensive Libraries

Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don’t have to write the complete code for that manually.

#### 2. Extensible

As we have seen earlier, Python can be**extended to other languages**. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

#### 3. Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add **scripting capabilities**to our code in the other language.

#### 4. Improved Productivity

The language’s simplicity and extensive libraries render programmers**more productive** than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

#### 5. IOT Opportunities

Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet Of Things. This is a way to connect the language with the real world.

#### 6. Simple and Easy

When working with Java, you may have to create a class to print **‘Hello World’**. But in Python, just a print statement will do. It is also quite **easy to learn, understand,** and**code.** This is why when people pick up Python, they have a hard time adjusting to other more verbose languages like Java.

#### 7. Readable

Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and **indentation is mandatory.** This further aids the readability of the code.

#### 8. Object-Oriented

This language supports both the **procedural and object-oriented**programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the **encapsulation of data** and functions into one.

#### 9. Free and Open-Source

Like we said earlier, Python is **freely available.** But not only can you[**download Python**](https://data-flair.training/blogs/install-python-windows/) for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

#### 10. Portable

When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn’t the same with Python. Here, you need to**code only once**, and you can run it anywhere. This is called **Write Once Run Anywhere (WORA)**. However, you need to be careful enough not to include any system-dependent features.

#### 11. Interpreted

Lastly, we will say that it is an interpreted language. Since statements are executed one by one, **debugging is easier** than in compiled languages.

Any doubts till now in the advantages of Python? Mention in the comment section.

### Advantages of Python Over Other Languages

#### 1. Less Coding

Almost all of the tasks done in Python requires less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don’t have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python to beginners.

#### 2. Affordable

Python is free therefore individuals, small companies or big organizations can leverage the free available resources to build applications. Python is popular and widely used so it gives you better community support.

**The 2019 Github annual survey showed us that Python has overtaken Java in the most popular programming language category.**

#### 3. Python is for Everyone

Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and [**machine learning**](https://data-flair.training/blogs/machine-learning-tutorials-home/), automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

### Disadvantages of Python

So far, we’ve seen why Python is a great choice for your project. But if you choose it, you should be aware of its consequences as well. Let’s now see the downsides of choosing Python over another language.

#### 1. Speed Limitations

We have seen that Python code is executed line by line. But since [Python](https://www.python.org/) is interpreted, it often results in **slow execution**. This, however, isn’t a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

#### 2. Weak in Mobile Computing and Browsers

While it serves as an excellent server-side language, Python is much rarely seen on the **client-side**. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is called **Carbonnelle**.

The reason it is not so famous despite the existence of Brython is that it isn’t that secure.

#### 3. Design Restrictions

As you know, Python is **dynamically-typed**. This means that you don’t need to declare the type of variable while writing the code. It uses **duck-typing**. But wait, what’s that? Well, it just means that if it looks like a duck, it must be a duck. While this is easy on the programmers during coding, it can**raise run-time errors**.

#### 4. Underdeveloped Database Access Layers

Compared to more widely used technologies like **JDBC (Java DataBase Connectivity)** and **ODBC (Open DataBase Connectivity)**, Python’s database access layers are a bit underdeveloped. Consequently, it is less often applied in huge enterprises.

#### 5. Simple

No, we’re not kidding. Python’s simplicity can indeed be a problem. Take my example. I don’t do Java, I’m more of a Python person. To me, its syntax is so simple that the verbosity of Java code seems unnecessary.

This was all about the Advantages and Disadvantages of Python Programming Language.

**History of Python : -**

What do the alphabet and the programming language Python have in common? Right, both start with ABC. If we are talking about ABC in the Python context, it's clear that the programming language ABC is meant. ABC is a general-purpose programming language and programming environment, which had been developed in the Netherlands, Amsterdam, at the CWI (Centrum Wiskunde &Informatica). The greatest achievement of ABC was to influence the design of Python.Python was conceptualized in the late 1980s. Guido van Rossum worked that time in a project at the CWI, called Amoeba, a distributed operating system. In an interview with Bill Venners1, Guido van Rossum said: "In the early 1980s, I worked as an implementer on a team building a language called ABC at Centrum voor Wiskunde en Informatica (CWI). I don't know how well people know ABC's influence on Python. I try to mention ABC's influence because I'm indebted to everything I learned during that project and to the people who worked on it."Later on in the same Interview, Guido van Rossum continued: "I remembered all my experience and some of my frustration with ABC. I decided to try to design a simple scripting language that possessed some of ABC's better properties, but without its problems. So I started typing. I created a simple virtual machine, a simple parser, and a simple runtime. I made my own version of the various ABC parts that I liked. I created a basic syntax, used indentation for statement grouping instead of curly braces or begin-end blocks, and developed a small number of powerful data types: a hash table (or dictionary, as we call it), a list, strings, and numbers."

**What is Machine Learning : -**

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data.

Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models tunable parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain.Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

**Categories Of Machine Leaning :-**

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning.

Supervised learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classification tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section.

Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

## Need for Machine Learning

Human beings, at this moment, are the most intelligent and advanced species on earth because they can think, evaluate and solve complex problems. On the other side, AI is still in its initial stage and haven’t surpassed human intelligence in many aspects. Then the question is that what is the need to make machine learn? The most suitable reason for doing this is, “to make decisions, based on data, with efficiency and scale”.

Lately, organizations are investing heavily in newer technologies like Artificial Intelligence, Machine Learning and Deep Learning to get the key information from data to perform several real-world tasks and solve problems. We can call it data-driven decisions taken by machines, particularly to automate the process. These data-driven decisions can be used, instead of using programing logic, in the problems that cannot be programmed inherently. The fact is that we can’t do without human intelligence, but other aspect is that we all need to solve real-world problems with efficiency at a huge scale. That is why the need for machine learning arises.

## Challenges in Machines Learning :-

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are −

**Quality of data** − Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.

**Time-Consuming task** − Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.

**Lack of specialist persons** − As ML technology is still in its infancy stage, availability of expert resources is a tough job.

**No clear objective for formulating business problems** − Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.

**Issue of overfitting & underfitting** − If the model is overfitting or underfitting, it cannot be represented well for the problem.

**Curse of dimensionality** − Another challenge ML model faces is too many features of data points. This can be a real hindrance.

**Difficulty in deployment** − Complexity of the ML model makes it quite difficult to be deployed in real life.

## 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

# How to Start Learning Machine Learning?

Arthur Samuel coined the term **“Machine Learning”** in 1959 and defined it as a **“Field of study that gives computers the capability to learn without being explicitly programmed”.**

And that was the beginning of Machine Learning! In modern times, Machine Learning is one of the most popular (if not the most!) career choices. According to [Indeed](http://blog.indeed.com/2019/03/14/best-jobs-2019/), Machine Learning Engineer Is The Best Job of 2019 with a 344% growth and an average base salary of **$146,085** per year.

But there is still a lot of doubt about what exactly is Machine Learning and how to start learning it? So this article deals with the Basics of Machine Learning and also the path you can follow to eventually become a full-fledged Machine Learning Engineer. Now let’s get started!!!

### How to start learning ML?

This is a rough roadmap you can follow on your way to becoming an insanely talented Machine Learning Engineer. Of course, you can always modify the steps according to your needs to reach your desired end-goal!

### Step 1 – Understand the Prerequisites

In case you are a genius, you could start ML directly but normally, there are some prerequisites that you need to know which include Linear Algebra, Multivariate Calculus, Statistics, and Python. And if you don’t know these, never fear! You don’t need a Ph.D. degree in these topics to get started but you do need a basic understanding.

#### (a) Learn Linear Algebra and Multivariate Calculus

Both Linear Algebra and Multivariate Calculus are important in Machine Learning. However, the extent to which you need them depends on your role as a data scientist. If you are more focused on application heavy machine learning, then you will not be that heavily focused on maths as there are many common libraries available. But if you want to focus on R&D in Machine Learning, then mastery of Linear Algebra and Multivariate Calculus is very important as you will have to implement many ML algorithms from scratch.

#### (b) Learn Statistics

Data plays a huge role in Machine Learning. In fact, around 80% of your time as an ML expert will be spent collecting and cleaning data. And statistics is a field that handles the collection, analysis, and presentation of data. So it is no surprise that you need to learn it!!!  
Some of the key concepts in statistics that are important are Statistical Significance, Probability Distributions, Hypothesis Testing, Regression, etc. Also, Bayesian Thinking is also a very important part of ML which deals with various concepts like Conditional Probability, Priors, and Posteriors, Maximum Likelihood, etc.

#### (c) Learn Python

Some people prefer to skip Linear Algebra, Multivariate Calculus and Statistics and learn them as they go along with trial and error. But the one thing that you absolutely cannot skip is [Python](https://www.geeksforgeeks.org/python-programming-language/)! While there are other languages you can use for Machine Learning like R, Scala, etc. Python is currently the most popular language for ML. In fact, there are many Python libraries that are specifically useful for Artificial Intelligence and Machine Learning such as [Keras](https://keras.io/" \t "_blank), [TensorFlow](https://www.tensorflow.org/" \t "_blank), [Scikit-learn](https://scikit-learn.org/stable/" \t "_blank), etc.

So if you want to learn ML, it’s best if you learn Python! You can do that using various online resources and courses such as [**Fork Python**](https://practice.geeksforgeeks.org/courses/fork-python) available Free on GeeksforGeeks.

### Step 2 – Learn Various ML Concepts

Now that you are done with the prerequisites, you can move on to actually learning ML (Which is the fun part!!!) It’s best to start with the basics and then move on to the more complicated stuff. Some of the basic concepts in ML are:

#### (a) Terminologies of Machine Learning

* **Model –**A model is a specific representation learned from data by applying some machine learning algorithm. A model is also called a hypothesis.
* **Feature –**A feature is an individual measurable property of the data. A set of numeric features can be conveniently described by a feature vector. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like color, smell, taste, etc.
* **Target (Label) –**A target variable or label is the value to be predicted by our model. For the fruit example discussed in the feature section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.
* **Training –**The idea is to give a set of inputs(features) and it’s expected outputs(labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.
* **Prediction –**Once our model is ready, it can be fed a set of inputs to which it will provide a predicted output(label).

#### (b) Types of Machine Learning

* **Supervised Learning –**This involves learning from a training dataset with labeled data using classification and regression models. This learning process continues until the required level of performance is achieved.
* **Unsupervised Learning –**This involves using unlabelled data and then finding the underlying structure in the data in order to learn more and more about the data itself using factor and cluster analysis models.
* **Semi-supervised Learning –**This involves using unlabelled data like Unsupervised Learning with a small amount of labeled data. Using labeled data vastly increases the learning accuracy and is also more cost-effective than Supervised Learning.
* **Reinforcement Learning –**This involves learning optimal actions through trial and error. So the next action is decided by learning behaviors that are based on the current state and that will maximize the reward in the future.

### Advantages of Machine learning :-

#### 1. Easily identifies trends and patterns -

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

#### 2. No human intervention needed (automation)

With ML, you don’t need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwares; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

#### 3. Continuous Improvement

As [**ML algorithms**](https://data-flair.training/blogs/machine-learning-algorithms/) gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

#### 4. Handling multi-dimensional and multi-variety data

Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

#### 5. Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

### Disadvantages of Machine Learning :-

#### 1. Data Acquisition

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

#### 2. Time and Resources

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

#### 3. Interpretation of Results

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

#### 4. High error-susceptibility

[**Machine Learning**](https://en.wikipedia.org/wiki/Machine_learning) is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

**Python Development Steps : -**

Guido Van Rossum published the first version of Python code (version 0.9.0) at alt.sources in February 1991. This release included already exception handling, functions, and the core data types of list, dict, str and others. It was also object oriented and had a module system.  
Python version 1.0 was released in January 1994. The major new features included in this release were the functional programming tools lambda, map, filter and reduce, which Guido Van Rossum never liked.Six and a half years later in October 2000, Python 2.0 was introduced. This release included list comprehensions, a full garbage collector and it was supporting unicode.Python flourished for another 8 years in the versions 2.x before the next major release as Python 3.0 (also known as "Python 3000" and "Py3K") was released. Python 3 is not backwards compatible with Python 2.x. The emphasis in Python 3 had been on the removal of duplicate programming constructs and modules, thus fulfilling or coming close to fulfilling the 13th law of the Zen of Python: "There should be one -- and preferably only one -- obvious way to do it."Some changes in Python 7.3:

* Print is now a function
* Views and iterators instead of lists
* The rules for ordering comparisons have been simplified. E.g. a heterogeneous list cannot be sorted, because all the elements of a list must be comparable to each other.
* There is only one integer type left, i.e. int. long is int as well.
* The division of two integers returns a float instead of an integer. "//" can be used to have the "old" behaviour.
* Text Vs. Data Instead Of Unicode Vs. 8-bit

**Purpose :-**

We demonstrated that our approach enables successful segmentation of intra-retinal layers—even with low-quality images containing speckle noise, low contrast, and different intensity ranges throughout—with the assistance of the ANIS feature.

**Python**

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace.

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

* 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 interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

**Modules Used in Project :-**

**Tensorflow**

TensorFlow is a [free](https://en.wikipedia.org/wiki/Free_software) and [open-source](https://en.wikipedia.org/wiki/Open-source_software) [software library for dataflow and differentiable programming](https://en.wikipedia.org/wiki/Library_(computing)) across a range of tasks. It is a symbolic math library, and is also used for [machine learning](https://en.wikipedia.org/wiki/Machine_learning) applications such as [neural networks](https://en.wikipedia.org/wiki/Neural_networks). It is used for both research and production at [Google](https://en.wikipedia.org/wiki/Google).‍

TensorFlow was developed by the [Google Brain](https://en.wikipedia.org/wiki/Google_Brain) team for internal Google use. It was released under the [Apache 2.0](https://en.wikipedia.org/wiki/Apache_License) [open-source license](https://en.wikipedia.org/wiki/Open-source_license) on November 9, 2015.

**Numpy**

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

* A powerful N-dimensional array object
* Sophisticated (broadcasting) functions
* Tools for integrating C/C++ and Fortran code
* Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

**Pandas**

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

**Matplotlib**

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and [IPython](http://ipython.org/) shells, the [Jupyter](http://jupyter.org/) Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the [sample plots](https://matplotlib.org/tutorials/introductory/sample_plots.html) and [thumbnail gallery](https://matplotlib.org/gallery/index.html).

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

**Scikit – learn**

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. **Python**

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace.

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

* 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 interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

**Install Python Step-by-Step in Windows and Mac :**

Python a versatile programming language doesn’t come pre-installed on your computer devices. Python was first released in the year 1991 and until today it is a very popular high-level programming language. Its style philosophy emphasizes code readability with its notable use of great whitespace.

The object-oriented approach and language construct provided by Python enables programmers to write both clear and logical code for projects. This software does not come pre-packaged with Windows.

## How to Install Python on Windows and Mac :

There have been several updates in the Python version over the years. The question is how to install Python? It might be confusing for the beginner who is willing to start learning Python but this tutorial will solve your query. The latest or the newest version of Python is version 3.7.4 or in other words, it is Python 3.

**Note:** The python version 3.7.4 cannot be used on Windows XP or earlier devices.

Before you start with the installation process of Python. First, you need to know about your **System Requirements**. Based on your system type i.e. operating system and based processor, you must download the python version. My system type is a **Windows 64-bit operating system**. So the steps below are to install python version 3.7.4 on Windows 7 device or to install Python 3. [Download the Python Cheatsheet here.](https://myelearninghub.com/python-cheat-sheet/)The steps on how to install Python on Windows 10, 8 and 7 are **divided into 4 parts** to help understand better.

### Download the Correct version into the system

**Step 1:** Go to the official site to download and install python using Google Chrome or any other web browser. OR Click on the following link: [**https://www.python.org**](https://www.python.org/)



Now, check for the latest and the correct version for your operating system.

**Step 2:** Click on the Download Tab.

****

**Step 3:** You can either select the Download Python for windows 3.7.4 button in Yellow Color or you can scroll further down and click on download with respective to their version. Here, we are downloading the most recent python version for windows 3.7.4

****

**Step 4:** Scroll down the page until you find the Files option.

**Step 5:** Here you see a different version of python along with the operating system.



• To download Windows 32-bit python, you can select any one from the three options: Windows x86 embeddable zip file, Windows x86 executable installer or Windows x86 web-based installer.

•To download Windows 64-bit python, you can select any one from the three options: Windows x86-64 embeddable zip file, Windows x86-64 executable installer or Windows x86-64 web-based installer.

Here we will install Windows x86-64 web-based installer. Here your first part regarding which version of python is to be downloaded is completed. Now we move ahead with the second part in installing python i.e. Installation

**Note:** To know the changes or updates that are made in the version you can click on the Release Note Option.

### Installation of Python

**Step 1:** Go to Download and Open the downloaded python version to carry out the installation process.



**Step 2:** Before you click on Install Now, Make sure to put a tick on Add Python 3.7 to PATH.



**Step 3:** Click on Install NOW After the installation is successful. Click on Close.



With these above three steps on python installation, you have successfully and correctly installed Python. Now is the time to verify the installation.

**Note:** The installation process might take a couple of minutes.

### Verify the Python Installation

**Step 1:** Click on Start

**Step 2:** In the Windows Run Command, type “cmd”.



**Step 3:** Open the Command prompt option.

**Step 4:** Let us test whether the python is correctly installed. Type **python –V** and press Enter.



**Step 5:** You will get the answer as 3.7.4

**Note:** If you have any of the earlier versions of Python already installed. You must first uninstall the earlier version and then install the new one.

### Check how the Python IDLE works

**Step 1:** Click on Start

**Step 2:** In the Windows Run command, type “python idle”.



**Step 3:** Click on IDLE (Python 3.7 64-bit) and launch the program

**Step 4:** To go ahead with working in IDLE you must first save the file. **Click on File > Click on Save**



**Step 5:** Name the file and save as type should be Python files. Click on SAVE. Here I have named the files as Hey World.

**Step 6:** Now for e.g. **enter print**

**6.SYSTEM TEST**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

### TYPES OF TESTS

**Unit testing**

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

**Integration testing**

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

**Functional test**

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:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

**System Test**

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

**White Box Testing**

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

**Black Box Testing**

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

**Unit Testing**

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

**Test strategy and approach**

Field testing will be performed manually and functional tests will be written in detail.

**Test objectives**

* All field entries must work properly.
* Pages must be activated from the identified link.
* The entry screen, messages and responses must not be delayed.

**Features to be tested**

* Verify that the entries are of the correct format
* No duplicate entries should be allowed
* All links should take the user to the correct page.

# 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.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

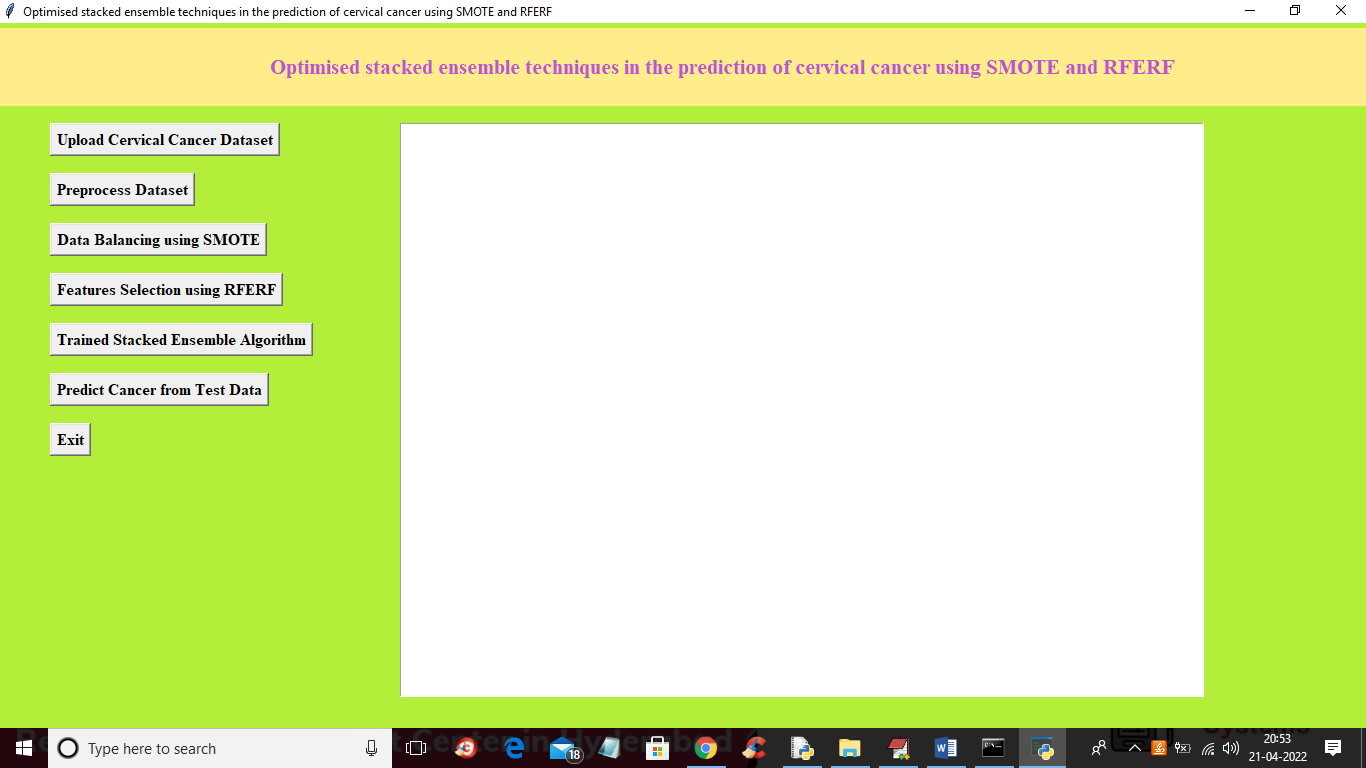
**Acceptance Testing**

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

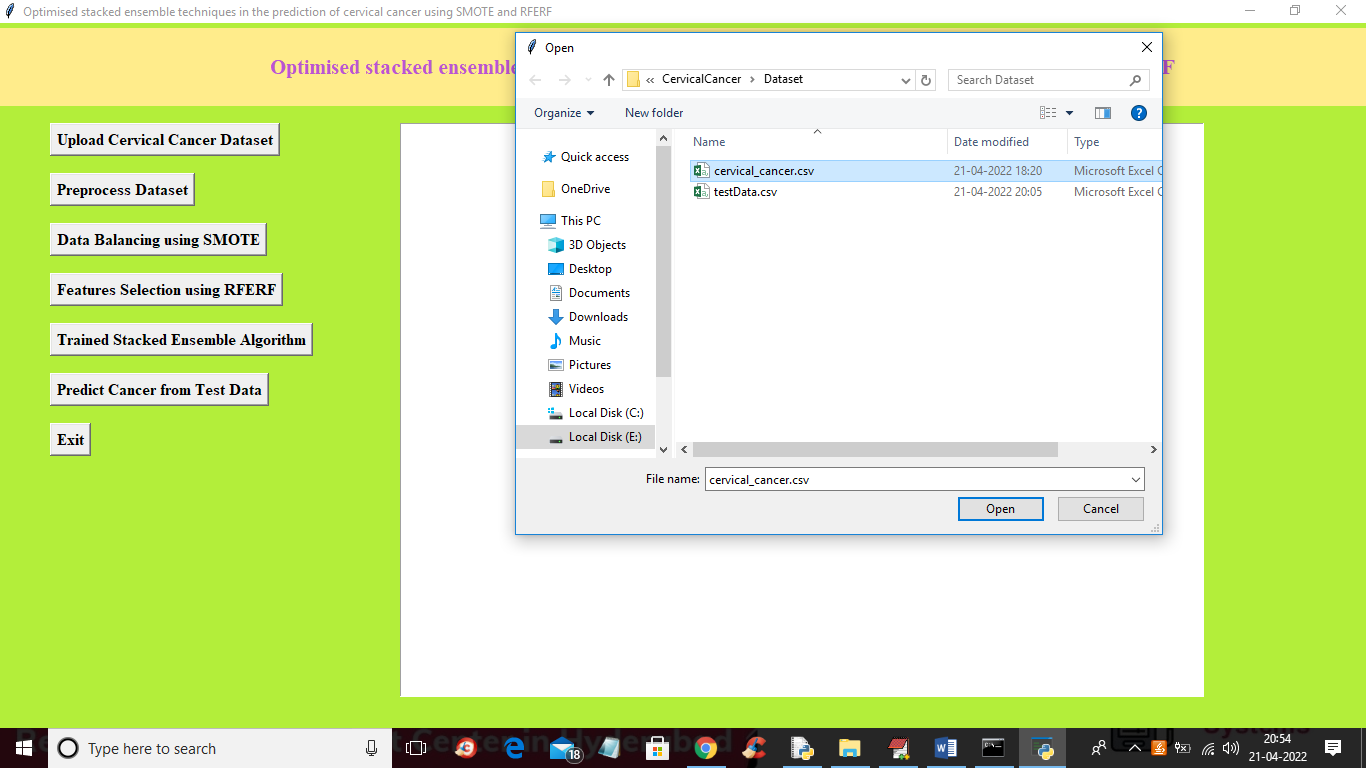
**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

SCREEN SHOTS :

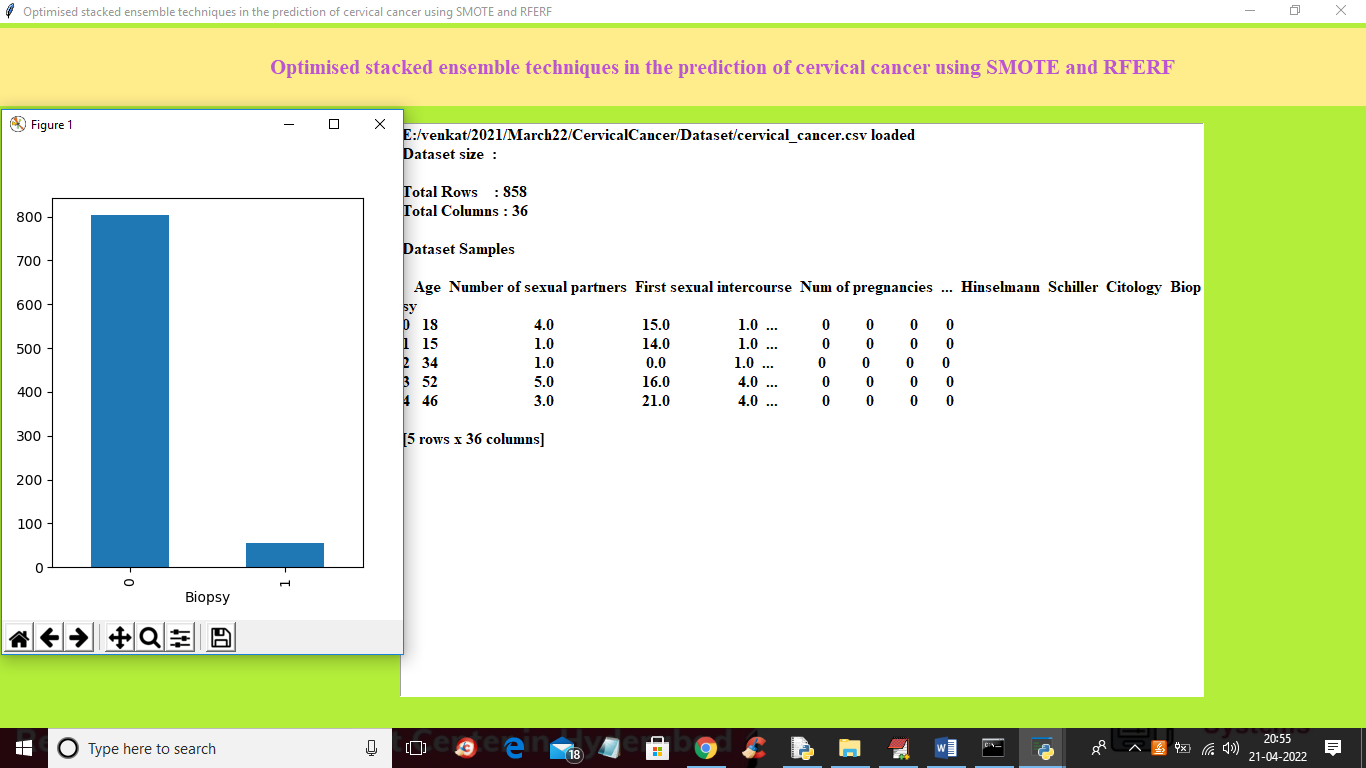
To run project double click on ‘run.bat’ file to get below screen



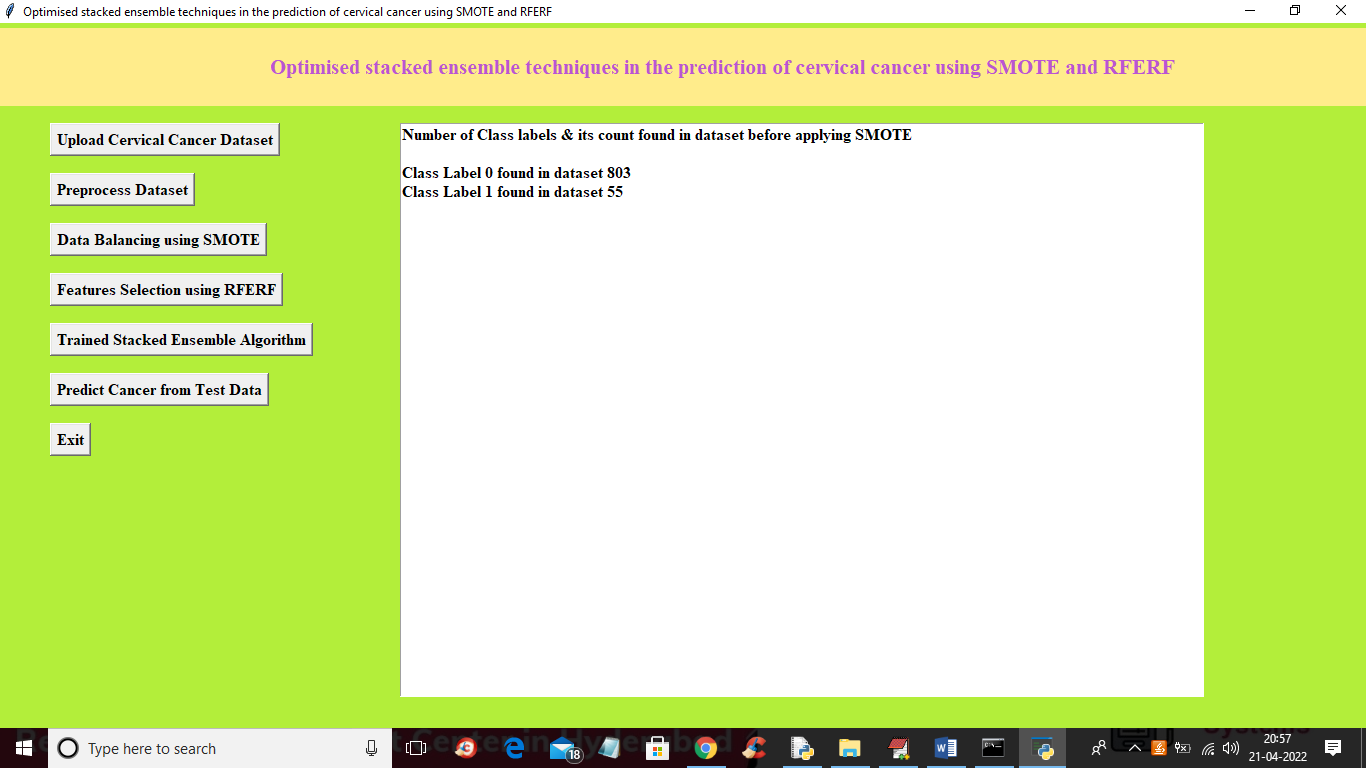
In above screen click on ‘Upload Cervical Cancer Dataset’ button to upload dataset and to get below screen



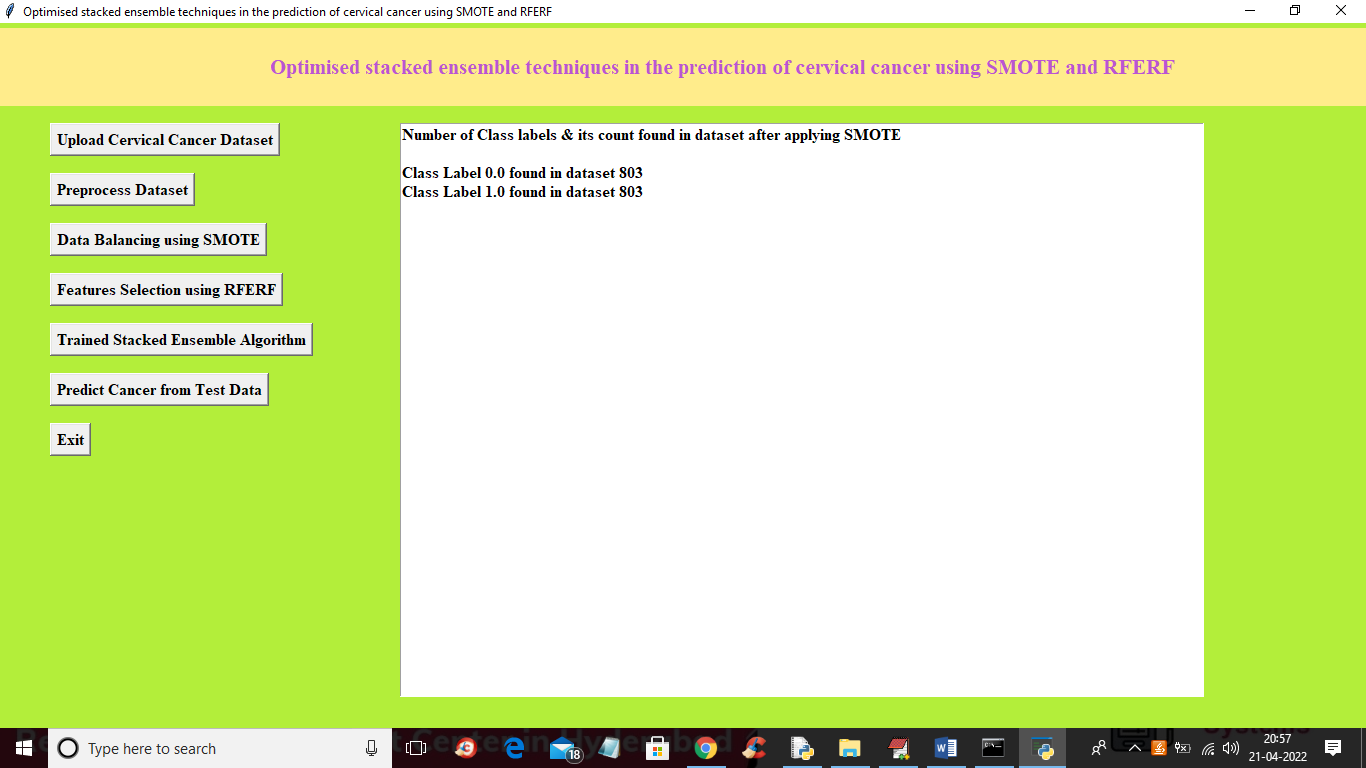
In above screen selecting and uploading dataset file and then click on ‘Open’ button to load dataset and to get below screen



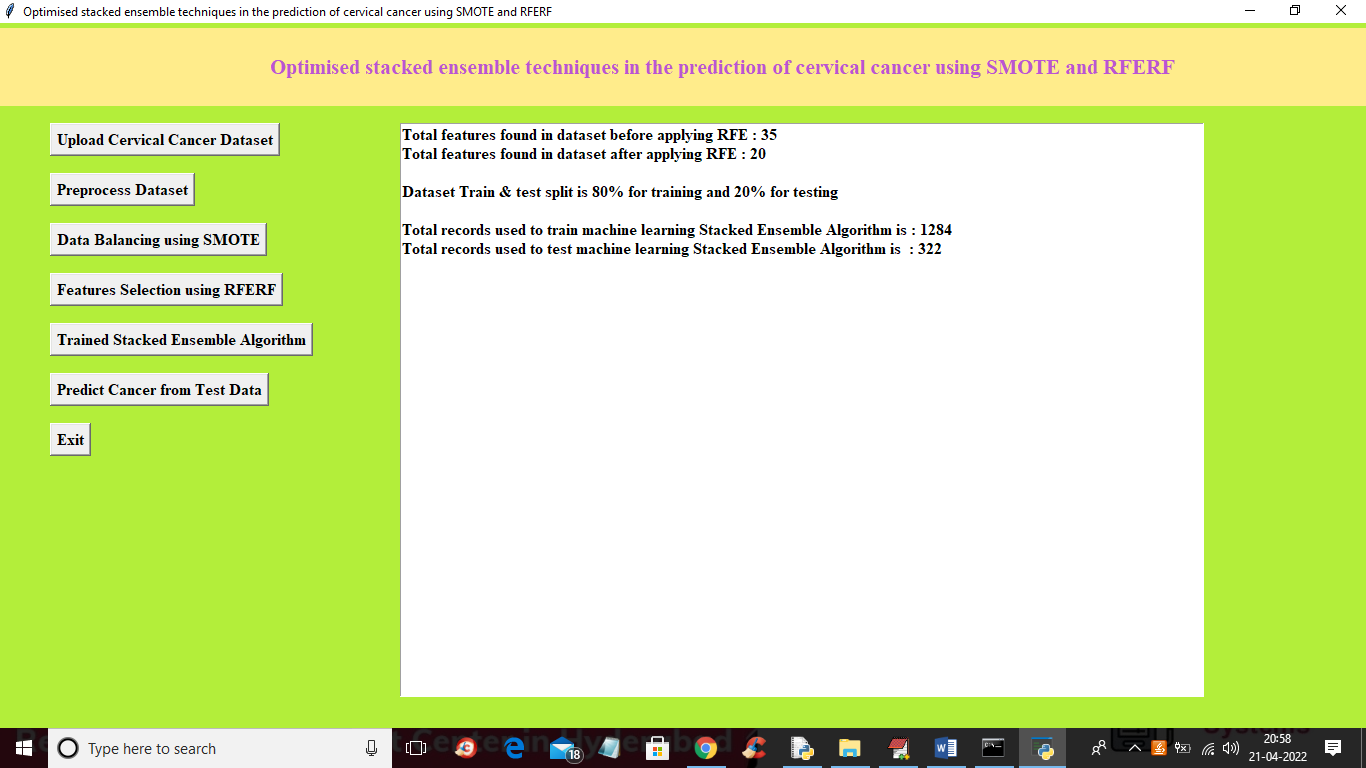
In above screen we can see dataset loaded and in graph x-axis represents class label 0 (normal) and 1 (cervical cancer) and y-axis represents number of records and we can see dataset is highly imbalanced as ‘0’ contains more than 800 records and ‘1’ class label contains only 53 records. Now close above graph and then click on ‘Preprocess Dataset" button to remove missing values and normalize dataset



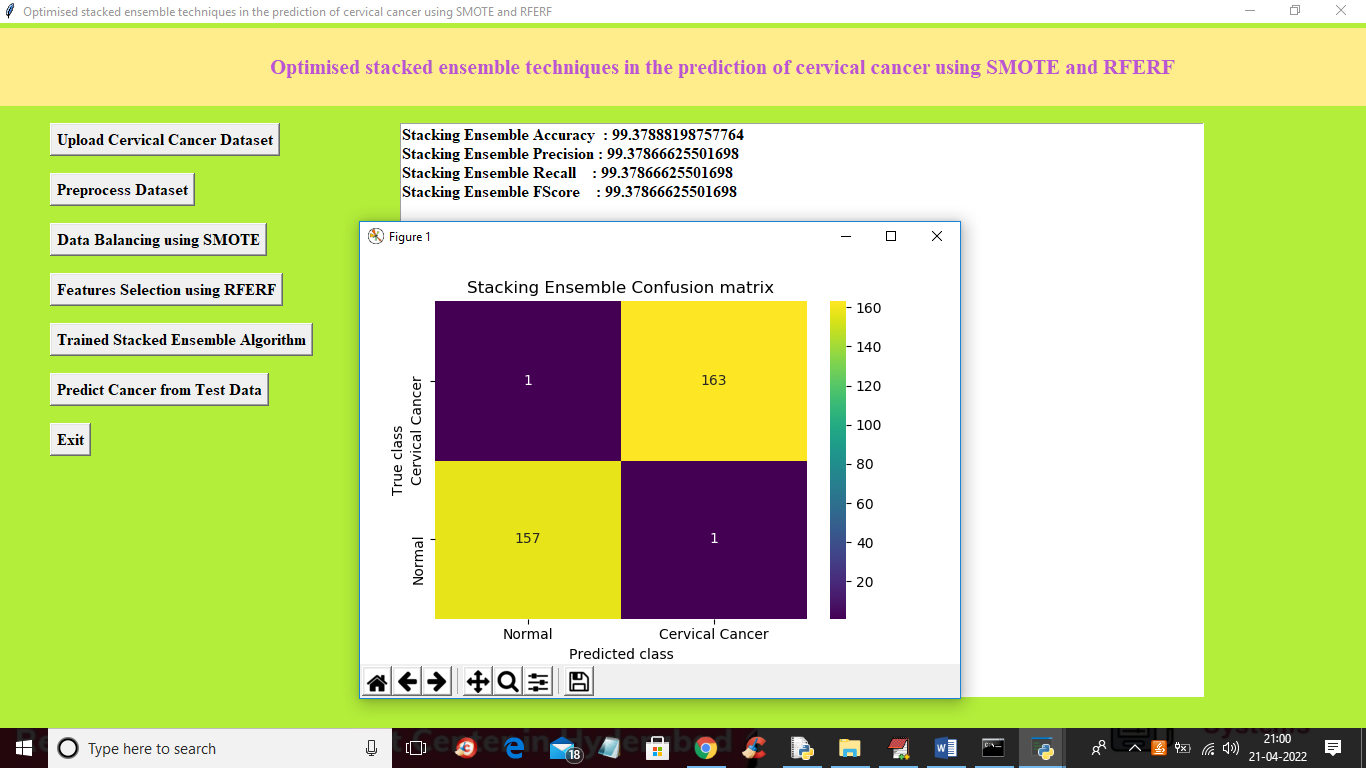
In above screen dataset is processed and class 0 contains 805 records and 1 contains 55 records so click on ‘Data Balancing using SMOTE’ button to balanced dataset and get below output



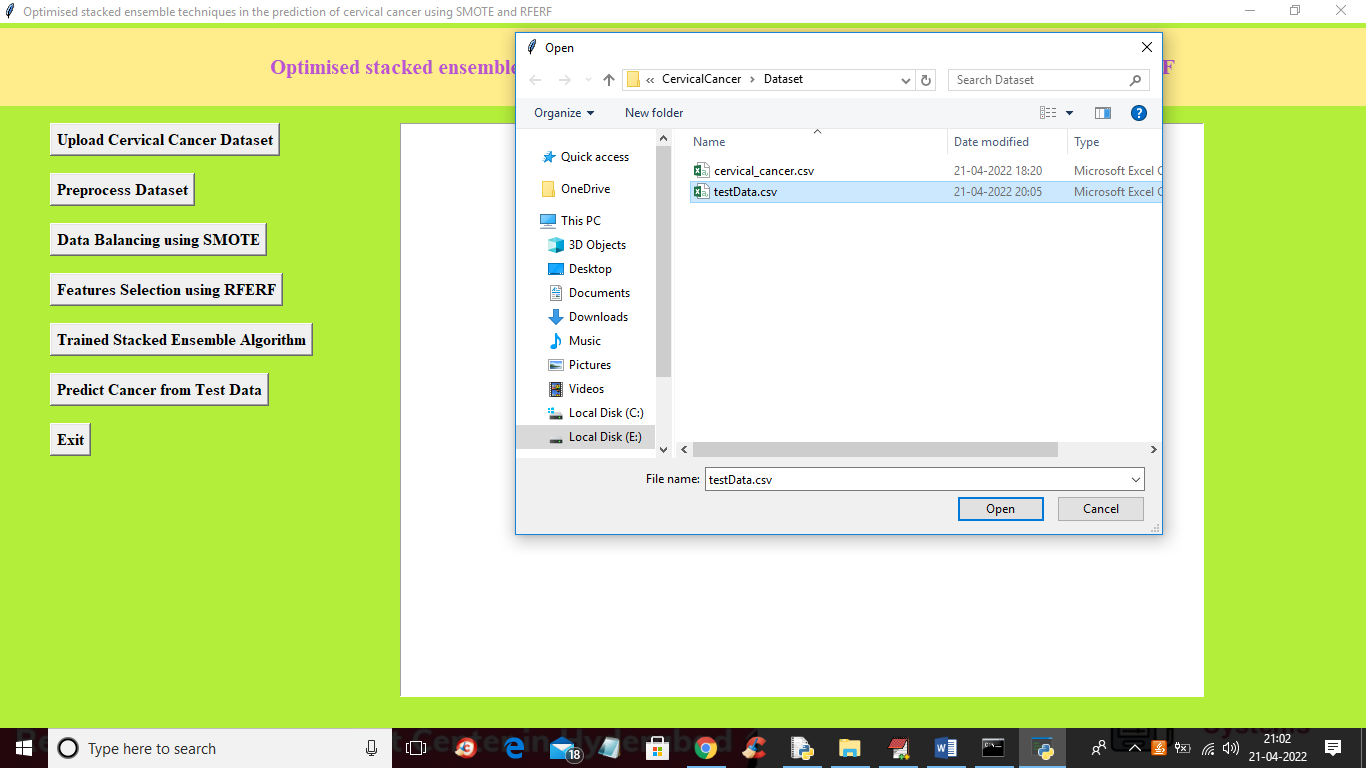
In above screen after applying smote both classes contains 803 records so dataset is balanced and now click on ‘Features Selection using RFERF’ button to get below output



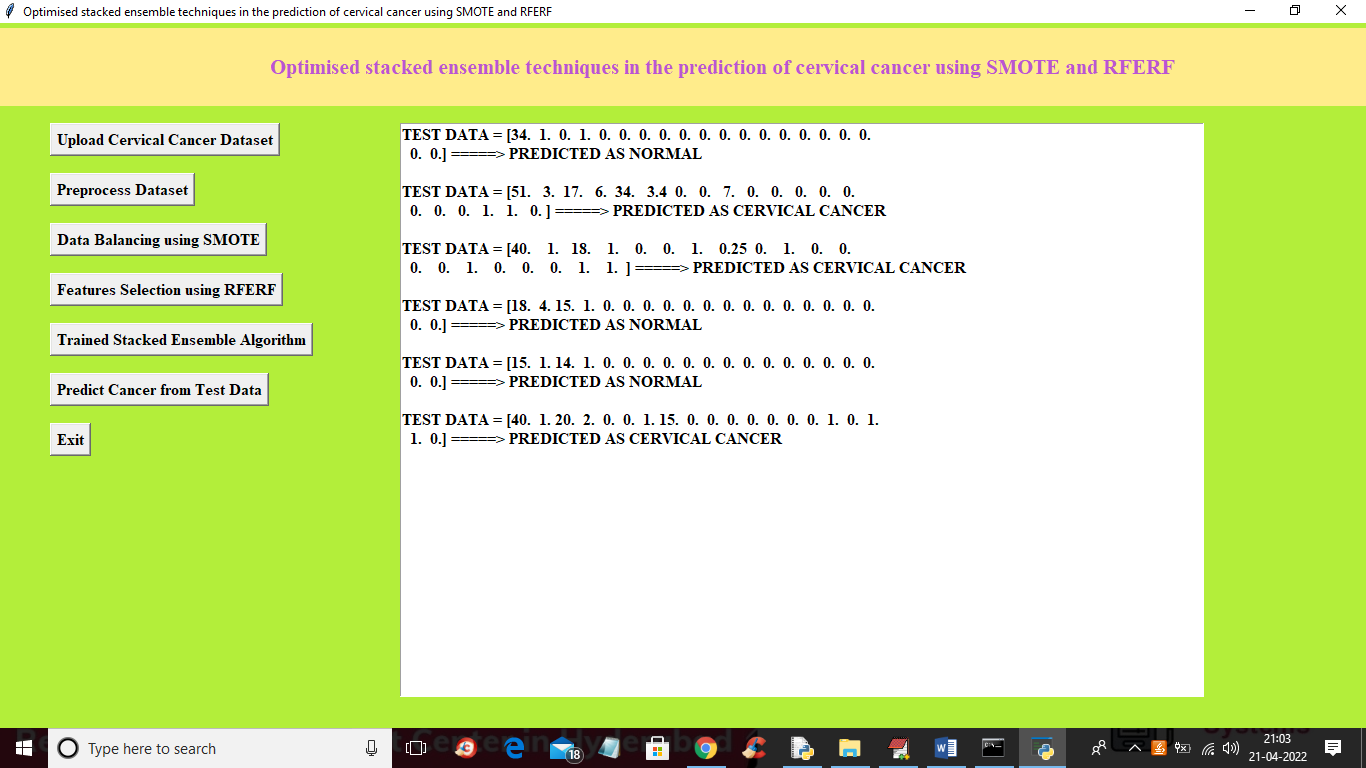
In above screen in first 2 lines we can see dataset contain 35 attributes and after applying RFE, attributes size reduced to 20 and then we can see dataset train and test split details. Now dataset is ready and now click on ‘Trained Stacked Ensemble Algorithm’ button to train algorithm and get below output



In above screen with stacked ensemble algorithm we got 99.37% accuracy and in confusion matrix graph we can see x-axis represents predicted classes and y-axis represents TRUE classes and only 1 class in wrongly predicted and remaining are the correctly predicted. Now close above graph and then click on ‘Predict Cancer from Test Data’ button to upload test file and get below output



In above screen selecting and uploading ‘testData.csv’ file and then click on “open’ button to get below output



In above screen in square bracket we can see test data and then after arrow symbol we can see predicted result as NORMAL or CERVICAL CANCER.

Conclusion :

this study presents an investigation of several ensemble techniques such as Random Forest, AdaBoost, and Extreme Gradient Boosting for diagnosing cervical cancer. the data set was obtained from the UCI machine learning repository containing 858 records, 32 features, and 4 target variables. the target variables are the diagnosis test used for cervical cancer. Experiments were conducted for each target class separately. Data preprocessing includes imputing missing values and class balancing using SMOTE. Moreover, bioinspired firefly algorithm was used to optimize the models, and to identify the key features. To compare the performance of the models, the experiments were conducted with 30 features and the selected features using SMOTED data. Extreme Gradient Boosting outperformed the other two models for all four target variables. For future work, the model will be validated on multiple data sets. Also, other models that can handle outliers and unbalanced data differently should be investigated.

References :

1] American Cancer Society, Key Statistics for Cervical Cancer, American Cancer Society, Atlanta, GA, USA, 2021, https:// www.cancer.org/cancer/cervical-cancer/about/key-statistics. html.

[2] “Cervical cancer: risk factors,” 2020, https://www.cancer.net/ cancer-types/cervical-cancer/risk-factors#:%7E:text=+e% 20most%20important%20risk%20factor,100%20different% 20types%20of%20HPV.

[3] “Cervical cancer,” 2020, https://www.nccc-online.org/ hpvcervical-cancer/cervical-cancer-overview/.

[4] W. Wu and H. Zhou, “Data-driven diagnosis of cervical cancer with support vector machine-based approaches,” IEEE Access, vol. 5, 2017.

[5] S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, “Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques,” IEEE Access, vol. 6, 2018.

[6] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, “Machine learning for assisting cervical cancer diagnosis: an ensemble approach,” Future Generation Computer Systems, vol. 106, pp. 199–205, 2020.

[7] E. Karim and N. Neehal, “An empirical study of cervical cancer diagnosis using ensemble methods,” in Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, December 2019.

[8] A. Ul-Islam, S. H. Ripon, and N. Qaisar Bhuiyan, “Cervical cancer risk factors: classification and mining associations,” APTIKOM Journal on Computer Science and Information Technologies, vol. 4, no. 1, pp. 8–18, 2019.

[9] Y. M. S. Al-Wesabi, A. Choudhury, and D. Won, “Classification of cervical cancer dataset,” in Proceedings of the 2018 IISE Annual Conference and Expo, pp. 1456–1461, Orlando, FL, USA, May 2018.

[10] K. Fernandes, D. Chicco, J. S. Cardoso, and J. Fernandes, “Supervised deep learning embeddings for the prediction of cervical cancer diagnosis,” PeerJ Computer Science, vol. 4, no. 5, pp. e154–21, 2018.

[11] K. Adem, S. Kiliçarslan, and O. Comert, “Classification and ¨ diagnosis of cervical cancer with stacked autoencoder and softmax classification,” Expert Systems with Applications, vol. 115, pp. 557–564, 2019.

[12] K. Fernandes, J. S. Cardoso, and J. Fernandes, “Transfer learning with partial observability applied to cervical cancer screening,” Pattern Recognition and Image Analysis, Springer, vol.

[13] “Cervical cancer (risk factors) data set,” 2020, https://archive. ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors% 29.

[14] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, “SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary,” Journal of Artificial Intelligence Research, vol. 61, pp. 863–905, 2018.

[15] X. Yang, Nature-Inspired Metaheuristic Algorithms, Vol. 4, Luniver Press, Cambridge, UK, 2nd edition, 2010.

[16] B. H. Nguyen, B. Xue, and M. Zhang, “A survey on swarm intelligence approaches to feature selection in data mining,” Swarm and Evolutionary Computation, vol. 54, 2020.

[17] M. Anbu and G. S. Anandha Mala, “Feature selection using firefly algorithm in software defect prediction,” Cluster Computing, vol. 22, no. s5, Article ID 10925, 2019.

[18] L. Breiman, “Random forests,” in Hands-On Machine Learning with R, pp. 203–219, CRC Press, Boca Raton, FL, USA, 2019.

[19] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?” in Lecture Notes in Computer Science, vol. 7376, pp. 154–168, Springer, New York, NY, USA, 2012.

[20] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13–17, pp. 785–794, San Francisco, CA, USA, August 2016.

[21] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” Frontiers in Neurorobotics, vol. 7, 2013.

[22] J. H. Friedman and J. J. Meulman, “Multiple additive regression trees with application in epidemiology,” Statistics in Medicine, vol. 22, no. 9, pp. 1365–1381, 2003.

[23] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.

[24] S. B. Kotsiantis, “Supervised machine learning: a review of classification techniques,” Informatica, vol. 31, pp. 249–268, 2007.