**Detection of Anomalous Behavior of Smartphone Devices using Change Point Analysis & Machine Learning**

**OBJECTIVE:**

The primary goal of this project is to determine the anomalous behaviour whether there is anomalous behaviour or not and to know this we have used Decision tree, XGBoosting and AdaBoost Classifiers and Support Vector Classifier to classify anomalous behaviour.

**ABSTRACT:**

Security, reliability, and availability have become three fundamental characteristics that smartphones and IoT (Internet of Things) devices have to possess to provide end-users a trustworthy experience. These properties can be degraded by extraneous events or anomalous behavior provoking damage in hardware, changes in software, theft of user information, and impact of device performance in terms of speed or availability. Considering these facts, this paper focuses on anomaly detection on smartphones using their power consumption signals. These signals represent the dynamic behavior of the device due to the action of different hardware components controlled by one or many applications at the same time. This behavior can be seen as a non-stationary process due to the changes in time of its statistical properties. Considering this assumption, our methodology uses a feature extraction technique that is based on changepoint detection theory. Then, it fits three machine learning classifiers to inject diversity and maximize the detection performance. The methodology was validated on a dataset of an emulated malware running in the background of a smartphone. Our results have been compared with several power signal based approaches demonstrating that the proposed technique performs better in terms of detection accuracy.

**Keywords:** Anomalous Behaviour, Decision tree, XGBoosting and AdaBoost Classifiers and Support Vector Classifier.

**INTRODUCTION**

In recent years, the Internet of Things (IoT) market is witnessing an explosion in the number of devices connected and to be connected. It is expected that by 2020 the number of devices will reach ∼50 billion. A significant percentage of these devices are smartphones. According to Statista in 2019, there are approximately 2.7 billion smartphone users around the world. A common characteristic of IoT devices and smartphones is the continuous necessity for Internet access through wireless communication systems to transmit sensitive and private information of users. For this reason, these devices have been of interest to cybercriminals who have developed applications with malicious code to steal passwords, emails, contacts, photos, recordings, health insights, or another valuable user information. Furthermore, other cybercriminals have focused their efforts to degrade and harm the performance of the infrastructure of cellular, private, and public networks converting IoT devices into botnets to provoke denial of service of these networks. To counter such malicious activities, researchers, companies, and even governments have been developing different methodologies. Most of these approaches are based on analyzing the static characteristics of the applications’ source code. A disadvantage of this strategy is that these methods are susceptible to obfuscation and modification of the code to avoid being detected. Thus, other researchers have been developing methodologies in which they analyze the dynamic characteristics of the device such as network traffic, power consumption, Central Processing Unit (CPU) activity, and temperature while applications are running. This analysis can be done in real time (on-line) or offline (analysis of measurements obtained beforehand). Other researchers have used hybrid techniques which are combinations of static and dynamic characteristics to do more effective recognition.

The present work proposes a novel methodology to decide if a smartphone is running a malicious application using the power consumption of the device. The hypothesis is that the power consumed by a device contains encoded and useful information that can be used to identify the presence of malwares. That is due to the fact that when a malware is installed in a device, it must perform some activities that depicts the combination of the energy consumed for each of the hardware components of the device such as CPU, network components, screen, Global Positioning System (GPS), accelerometers, or other components. This methodology uses offline processing technique and off-device measurement in which an external device is used to collect the power consumption to improve the resolution of the power traces assuming that important features can be embedded in very short periods of time. Furthermore, we use the theory of changepoint detection to extract features of a non-stationary time series signal. The features extracted by this theory have been used as the input to a classifier to define a binary classification problem applying different machine learning techniques.

**LITERATURE SURVEY**

**[1] D. Evans, “The internet of things: How the next evolution of the internet is changing everything,” CISCO white paper, Tech. Rep., 2011.**

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drastically

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drastically

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Mobile app distribution platforms such as Google Play Store allow

users to share their feedback about downloaded apps in the form

of a review comment and a corresponding star rating. Typically,

the star rating ranges from one to ve stars, with one star denoting

a high sense of dissatisfaction with the app and ve stars denoting

a high sense of satisfaction.

Unfortunately, due to a variety of reasons, oen the star rating

provided by a user is inconsistent with the opinion expressed in

the review. For example, consider the following review for the

Facebook App on Android; “Awesome App”. One would reasonably

expect the rating for this review to be ve stars, but the actual

rating is one star!

Such inconsistent ratings can lead to a deated (or inated)

overall average rating of an app which can aect user downloads,

as typically users look at the average star ratings while making a

decision on downloading an app. Also, the app developers receive

a biased feedback about the application that does not represent

ground reality. is is especially signicant for small apps with a

few thousand downloads as even a small number of mismatched

reviews can bring down the average rating drasticall

Nowadays, we experience an abundance of Internet of Things middleware solutions that make the sensors and the actuators are able to connect to the Internet. These solutions, referred to as platforms to gain a widespread adoption, have to meet the expectations of different players in the IoT ecosystem, including devices. Low cost devices are easily able to connect wirelessly to the Internet, from handhelds to coffee machines, also known as Internet of Things (IoT). This research describes the methodology and the development process of creating an IoT platform. This paper also presents the architecture and implementation for the IoT platform. The goal of this research is to develop an analytics engine which can gather sensor data from different devices and provide the ability to gain meaningful information from IoT data and act on it using machine learning algorithms. The proposed system is introducing the use of a messaging system to improve the overall system performance as well as provide easy scalability.

**[2] Statista, “Number of mobile phone users worldwide from 2015 to 2020 (in billions).” [Online]. Available: https://www.statista.com/statistics/274774/forecast-ofmobile-phone-users-worldwide/.**

The advancement of virtual reality has sharpened the concept of augmented reality (AR) to new dimension of perceptions of seeing, hearing and immersing in a real world. The evolution of mobile devices has pioneered AR as a state-of-the-art technology in the last decade giving rise to more and more location-based mobile AR (LBMAR) systems. However, notably there are very limited review studies that have focused on investigating factors such as: growth, types, characteristics, features, sensors, application domains and their respective challenges. This study presents a systematic review on location-based mobile augmented reality (LBMAR) system. A total of 35 studies published between the years 2013 and 2018 in top six most popular indexed databases are reviewed. The systematic review has been conducted using Kitchenham method, and the analysis of the findings was carried out using the PRISMA method. This chapter provides a major review of the current state of LBMAR system and outlines the research issues that require more investigation.

**[3]** **A. Arabo and B. Pranggono, “Mobile malware and smart device security: Trends, challenges and solutions,” in 2013 19th International Conference on Control Systems and Computer Science, May 2013, pp. 526–531.**

environments and requirements. Tracking and understanding

changes in modules and relationships in a software project is

difﬁcult, but even more so when the software goes through several

types of changes. The typical complexity and size of software

also makes it harder to grasp software evolution patterns. In

this paper, we present an interactive matrix-based visualization

technique that, combined with animation, depicts how software

designs evolve. For example, it shows which new modules

and couplings are added and removed over time. Our generic

visualization supports dynamic and weighted digraphs and is

applied in the context of software evolution. Analyzing source

code changes is important to determine the software’s structural

organization and identify quality issues over time. To demonstrate

our approach, we explore open-source repositories and discuss

some of our ﬁndings regarding these evolving software designs

This work is part of the research to study trends and challenges of cyber security to smart devices in smart homes. We have seen the development and demand for seamless interconnectivity of smart devices to provide various functionality and abilities to users. While these devices provide more features and functionality, they also introduce new risks and threats. Subsequently, current cyber security issues related to smart devices are discussed and analyzed. The paper begins with related background and motivation. We identified mobile malware as one of the main issue in the smart devices’ security. In the near future, mobile smart device users can expect to see a striking increase in malware and notable advancements in malware-related attacks, particularly on the Android platform as the user base has grown exponentially. We discuss and analyzed mobile malware in details and identified challenges and future trends in this area. Then we propose and discuss an integrated security solution for cyber security in smart devices to tackle the issue.

**[4]** **T. Kim, B. Kang, M. Rho, and et. all, “A multimodal deep learning method for android malware detection using various features,” IEEE Trans. on Info. Forensics and Security, vol. 14, no. 3, 2019.**

With the widespread use of smartphones, the number of malware has been increasing exponentially. Among smart devices, Android devices are the most targeted devices by malware because of their high popularity. This paper proposes a novel framework for Android malware detection. Our framework uses various kinds of features to reflect the properties of Android applications from various aspects, and the features are refined using our existence-based or similarity-based feature extraction method for effective feature representation on malware detection. Besides, a multimodal deep learning method is proposed to be used as a malware detection model. This paper is the first study of the multimodal deep learning to be used in the Android malware detection. With our detection model, it was possible to maximize the benefits of encompassing multiple feature types. To evaluate the performance, we carried out various experiments with a total of 41,260 samples. We compared the accuracy of our model with that of other deep neural network models. Furthermore, we evaluated our framework in various aspects including the efficiency in model updates, the usefulness of diverse features, and our feature representation method. In addition, we compared the performance of our framework with those of other existing methods including deep learning based methods.

**[5]** **Y.-S. Yen and H.-M. Sun, “An android mutation malware detection based on deep learning using visualization of importance from codes,” Microelectronics Reliability, vol. 93, pp. 109–114, 2019.**

The rapid proliferation of Android malware is challenging the classification of the Android malware family. The traditional static method for classification is easily affected by the confusion and reinforcement, while the dynamic method is expensive in computation. To solve these problems, this paper proposes an Android malware familial classification method based on Dalvik Executable (DEX) file section features. First, the DEX file is converted into RGB (Red/Green/Blue) image and plain text respectively, and then, the color and texture of image and text are extracted as features. Finally, a feature fusion algorithm based on multiple kernel learning is used for classification. In this experiment, the Android Malware Dataset (AMD) was selected as the sample set. Two different comparative experiments were set up, and the method in this paper was compared with the common visualization method and feature fusion method. The results show that our method has a better classification effect with precision, recall and F1 score reaching 0.96. Besides, the time of feature extraction in this paper is reduced by 2.999 seconds compared with the method of frequent subsequence. In conclusion, the method proposed in this paper is efficient and precise in the classification of the Android malware family.

**SYSTEM ANALYSIS & FEASIBILITY STUDY**

**Existing Method:**

In the existing system, implementation of machine learning algorithms is bit complex to build due to the lack of information about the data visualization. Mathematical calculations are used in existing system for model building this may takes the lot of time and complexity. To overcome all this, we use machine learning packages available in the scikit-learn library.

**Disadvantages:**

* High complexity.
* Time consuming.

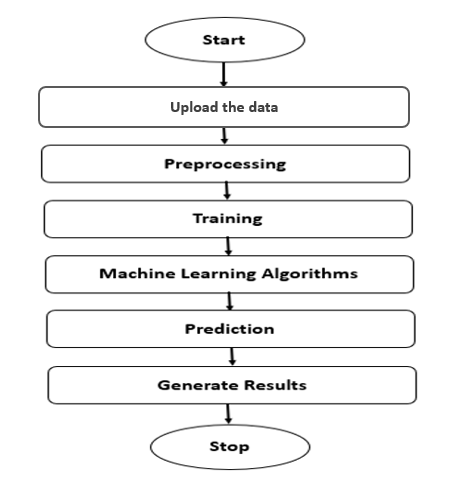
**Proposed System:**

Proposed several machine learning models to classify whether there is foodborne disease or not , but none have adequately addressed this misdiagnosis problem. Also, similar studies that have proposed models for evaluation of such performance classification mostly do not consider the heterogeneity and the size of the data Therefore, we propose a Disease, Random Forest, Decission Tree, Gradient Boosting and AdaBoost Classifier to predict.

**Advantages**:

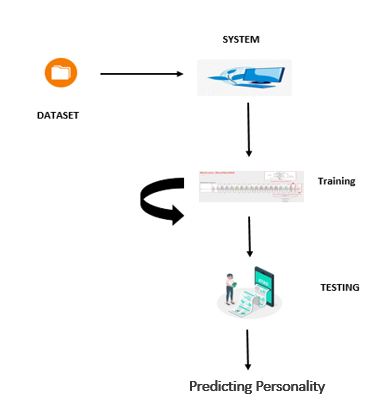
* Highest accuracy
* Reduces time complexity.
* Easy to use

**Block Diagram:**



**Fig**: Block Diagram

**Architecture:**



**METHODOLOGY AND ALGORITHMS:**

**1. AdaBoost Classifier:**

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as variance for supervised learning. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boosting with a slight difference. Let’s discuss this difference in detail.

First, let us discuss how boosting works. It makes ‘n’ number of decision trees during the data training period. As the first decision tree/model is made, the incorrectly classified record in the first model is given priority. Only these records are sent as input for the second model. The process goes on until we specify a number of base learners we want to create. Remember, repetition of records is allowed with all boosting techniques.

This figure shows how the first model is made and errors from the first model are noted by the algorithm. The record which is incorrectly classified is used as input for the next model. This process is repeated until the specified condition is met. As you can see in the figure, there are ‘n’ number of models made by taking the errors from the previous model. This is how boosting works. The models 1,2, 3,…, N are individual models that can be known as decision trees. All types of boosting models work on the same principle.

Since we now know the boosting principle, it will be easy to understand the AdaBoost algorithm. Let’s dive into AdaBoost’s working. When the random forest is used, the algorithm makes an ‘n’ number of trees. It makes proper trees that consist of a start node with several leaf nodes. Some trees might be bigger than others, but there is no fixed depth in a random forest. With AdaBoost, however, the algorithm only makes a node with two leaves, known as Stump.

The figure here represents the stump. It can be seen clearly that it has only one node with two leaves. These stumps are weak learners and boosting techniques prefer this. The order of stumps is very important in AdaBoost. The error of the first stump influences how other stumps are made. Let’s understand this with an example.

Here’s a sample dataset consisting of only three features where the output is in categorical form. The image shows the actual representation of the dataset. As the output is in binary/categorical form, it becomes a classification problem. In real life, the dataset can have any number of records and features in it. Let us consider 5 datasets for explanation purposes. The output is in categorical form, here in the form of Yes or No. All these records will be assigned a sample weight. The formula used for this is ‘W=1/N’ where N is the number of records. In this dataset, there are only 5 records, so the sample weight becomes 1/5 initially. Every record gets the same weight. In this case, it’s 1/5.

Learn AdaBoost Model from Data

Ada Boosting is best used to boost the performance of decision trees and this is based on binary classification problems.

AdaBoost was originally called AdaBoost.M1 by the author. More recently it may be referred to as discrete Ada Boost. As because it is used for classification rather than regression.

AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners.

**2. Decision Tree Classifier:**

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal.

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn’t split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

Although, a real dataset will have a lot more features and this will just be a branch in a much bigger tree, but you can’t ignore the simplicity of this algorithm. The feature importance is clear and relations can be viewed easily. This methodology is more commonly known as learning decision tree from data and above tree is called Classification tree as the target is to classify passenger as survived or died. Regression trees are represented in the same manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees.

So, what is actually going on in the background? Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop. As a tree generally grows arbitrarily, you will need to trim it down for it to look beautiful. Let’s start with a common technique used for splitting.

**3. XGBoosting Classifier:**

XGBoost stands for “Extreme Gradient Boosting”. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

Boosting

Boosting is an ensemble learning technique to build a strong classifier from several weak classifiers in series. Boosting algorithms play a crucial role in dealing with bias-variance trade-off. Unlike bagging algorithms, which only controls for high variance in a model, boosting controls both the aspects (bias & variance) and is considered to be more effective.

Below are the few types of boosting algorithms:

• AdaBoost (Adaptive Boosting)

• Gradient Boosting

• XGBoost

• CatBoost

• Light GBM

XGBoost

XGBoost stands for eXtreme Gradient Boosting. It became popular in the recent days and is dominating applied machine learning and Kaggle competitions for structured data because of its scalability.

XGBoost is an extension to gradient boosted decision trees (GBM) and specially designed to improve speed and performance.

XGBoost Features

• Regularized Learning: Regularization term helps to smooth the final learnt weights to avoid over-fitting. The regularized objective will tend to select a model employing simple and predictive functions.

• Gradient Tree Boosting: The tree ensemble model cannot be optimized using traditional optimization methods in Euclidean space. Instead, the model is trained in an additive manner.

• Shrinkage and Column Subsampling: Besides the regularized objective, two additional techniques are used to further prevent over fitting. The first technique is shrinkage introduced by Friedman. Shrinkage scales newly added weights by a factor η after each step of tree boosting. Similar to a learning rate in stochastic optimization, shrinkage reduces the influence of each tree and leaves space for future trees to improve the model.

The second technique is the column (feature) subsampling. This technique is used in Random Forest. Column sub-sampling prevents over-fitting even more so than the traditional row sub-sampling. The usage of column sub-samples also speeds up computations of the parallel algorithm.

SPLITTING ALGORITHMS

• Exact Greedy Algorithm: The main problem in tree learning is to find the best split. This algorithm enumerates over all the possible splits on all the features. It is computationally demanding to enumerate all the possible splits for continuous features.

• Approximate Algorithm: The exact greedy algorithm is very powerful since it enumerates over all possible splitting points greedily. However, it is impossible to efficiently do so when the data does not fit entirely into memory. Approximate Algorithm proposes candidate splitting points according to percentiles of feature distribution. The algorithm then maps the continuous features into buckets split by these candidate points, aggregates the statistics and finds the best solution among proposals based on the aggregated statistics.

• Weighted Quantile Sketch: One important step in the approximate algorithm is to propose candidate split points. XGBoost has a distributed weighted quantile sketch algorithm to effectively handle weighted data.

• Sparsity-aware Split Finding: In many real-world problems, it is quite common for the input x to be sparse. There are multiple possible causes for sparsity:

• Presence of missing values in the data

• Frequent zero entries in the statistics

• Artifacts of feature engineering such as one-hot encoding

It is important to make the algorithm aware of the sparsity pattern in the data. XGBoost handles all sparsity patterns in a unified way.

System Features

1. Parallelization of tree construction using all of your CPU cores during training. Collecting statistics for each column can be parallelized, giving us a parallel algorithm for split finding.

2. Cache-aware Access: XGBoost has been designed to make optimal use of hardware. This is done by allocating internal buffers in each thread, where the gradient statistics can be stored.

3. Blocks for Out-of-core Computation for very large datasets that don’t fit into memory.

4. Distributed Computing for training very large models using a cluster of machines.

5. Column Block for Parallel Learning. The most time-consuming part of tree learning is to get the data into sorted order. In order to reduce the cost of sorting, the data is stored in the column blocks in sorted order in compressed format.

Goals of XGBoost

6. Execution Speed: XGBoost was almost always faster than the other benchmarked implementations from R, Python Spark and H2O and it is really faster when compared to the other algorithms.

7. Model Performance: XGBoost dominates structured or tabular datasets on classification and regression predictive modelling problems.

Conclusion

XGBoost is a faster algorithm when compared to other algorithms because of its parallel and distributed computing. XGBoost is developed with both deep considerations in terms of systems optimization and principles in machine learning. The goal of this library is to push the extreme of the computation limits of machines to provide a scalable, portable and accurate library.

**4. Support Vector Classifier:**

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

Possible hyper planes :

To separate the two classes of data points, there are many possible Hyper planes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyper planes and Support Vectors

Hyper planes in 2D and 3D feature space

Hyper planes are decision boundaries that help classify the data points. Data points falling on either side of the hyper plane can be attributed to different classes. Also, the dimension of the hyper plane depends upon the number of features. If the number of input features is 2, then the hyper plane is just a line. If the number of input features is 3, then the hyper plane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

Support Vectors

Support vectors are data points that are closer to the hyper plane and influence the position and orientation of the hyper plane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyper plane. These are the points that help us build our SVM.

Large Margin Intuition

In logistic regression, we take the output of the linear function and squash the value within the range of [0,1] using the sigmoid function. If the squashed value is greater than a threshold value (0.5) we assign it a label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values ([-1, 1]) which acts as margin.

Cost Function and Gradient Updates

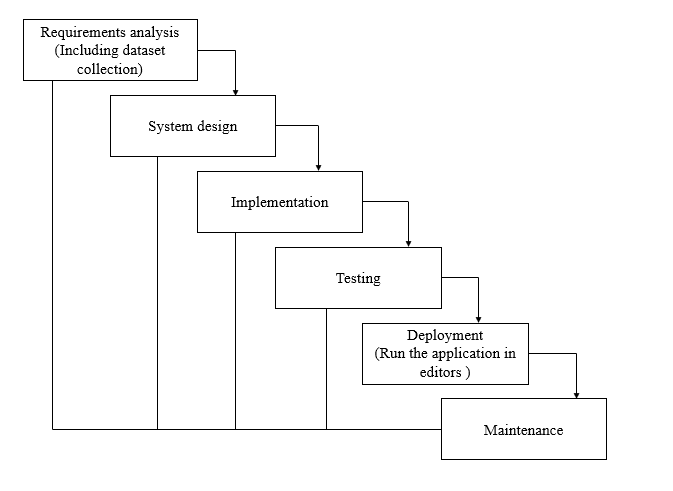
In the SVM algorithm, we are looking to maximize the margin between the data points and the hyper plane. The loss function that helps maximize the margin is hinge loss.

Hinge loss function (function on left can be represented as a function on the right)

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we then calculate the loss value. We also add a regularization parameter the cost function. The objective of the regularization parameter is to balance the margin maximization and loss. After adding the regularization parameter.

**SOFTWARE DEVELOPMENT LIFE CYCLE – SDLC:**

In our project we use waterfall model as our software development cycle because of its step-by-step procedure while implementing.



**Fig1**: Waterfall Model

* **Requirement Gathering and analysis** − All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.
* **System Design** − the requirement specifications from first phase are studied in this phase and the system design is prepared. This system design helps in specifying hardware and system requirements and helps in defining the overall system architecture.
* **Implementation** − with inputs from the system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality, which is referred to as Unit Testing.
* **Integration and Testing** − All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration the entire system is tested for any faults and failures.
* **Deployment of system** − Once the functional and non-functional testing is done; the product is deployed in the customer environment or released into the market.
* **Maintenance** − There are some issues which come up in the client environment. To fix those issues, patches are released. Also, to enhance the product some better versions are released. Maintenance is done to deliver these changes in the customer environment.

**FEASIBILITY STUDY**

The feasibility of the project is analysed 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

**Economic 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.

**SYSTEM REQUIREMENTS SPECIFICATION**

**Functional and non-functional requirements:**

Requirement’s analysis is very critical process that enables the success of a system or software project to be assessed. Requirements are generally split into two types: Functional and non-functional requirements.

**Functional Requirements**: These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

Examples of functional requirements:

1. Authentication of user whenever he/she logs into the system
2. System shutdown in case of a cyber-attack
3. A verification email is sent to user whenever he/she register for the first time on some software system.

**Non-functional requirements**: These are basically the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to other. They are also called non-behavioral requirements.  
They basically deal with issues like:

* Portability
* Security
* Maintainability
* Reliability
* Scalability
* Performance
* Reusability
* Flexibility

Examples of non-functional requirements:

1. Emails should be sent with a latency of no greater than 12 hours from such an activity.
2. The processing of each request should be done within 10 seconds
3. The site should load in 3 seconds whenever of simultaneous users are > 10000

**SOFTWARE AND HARDWARE REQUIREMENTS:**

**Hardware:**

Operating system : Windows 7 or 7+

RAM : 8 GB

Hard disc or SSD : More than 500 GB

Processor : Intel 3rd generation or high or Ryzen with 8 GB Ram

**Software:**

Software’s : Python 3.6 or high version

IDE : PyCharm.

Framework : Flask

**SYSTEM DESIGN:**

## **Input Design:**

In an information system, input is the raw data that is processed to produce output. During the input design, the developers must consider the input devices such as PC, MICR, OMR, etc.

Therefore, the quality of system input determines the quality of system output. Well-designed input forms and screens have following properties −

* It should serve specific purpose effectively such as storing, recording, and retrieving the information.
* It ensures proper completion with accuracy.
* It should be easy to fill and straightforward.
* It should focus on user’s attention, consistency, and simplicity.
* All these objectives are obtained using the knowledge of basic design principles regarding −
  + What are the inputs needed for the system?
  + How end users respond to different elements of forms and screens.

### Objectives for Input Design:

The objectives of input design are −

* To design data entry and input procedures
* To reduce input volume
* To design source documents for data capture or devise other data capture methods
* To design input data records, data entry screens, user interface screens, etc.
* To use validation checks and develop effective input controls.

**Output Design:**

The design of output is the most important task of any system. During output design, developers identify the type of outputs needed, and consider the necessary output controls and prototype report layouts.

### Objectives of Output Design:

The objectives of input design are:

* To develop output design that serves the intended purpose and eliminates the production of unwanted output.
* To develop the output design that meets the end user’s requirements.
* To deliver the appropriate quantity of output.
* To form the output in appropriate format and direct it to the right person.
* To make the output available on time for making good decisions.

**MODULES:**

1. **User**:
   1. **View Home page:**

Here user view the home page of the anomalous behaviour web application.

* 1. **View Upload page:**

In the about page, users can learn more about the anomalous behaviour prediction.

* 1. **Input Model:**

The user must provide input values for the certain fields in order to get results.

* 1. **View Results:**

User view’s the generated results from the model.

* 1. **View score:**

Here user have ability to view the score in %

1. **System**
   1. **Working on dataset:**

System checks for data whether it is available or not and load the data in csv files.

* 1. **Pre-processing:**

Data need to be pre-processed according the models it helps to increase the accuracy of the model and better information about the data.

* 1. **Training the data:**

After pre-processing the data will split into two parts as train and test data before training with the given algorithms.

* 1. **Model Building**

To create a model that predicts the personality with better accuracy, this module will help user.

* 1. **Generated Score:**
  2. Here user view the score in %
  3. **Generate Results:**

We train the machine learning algorithm and predict the anomalous behaviour.

**UML DIAGRAMS**

UML stands for Unified Modelling Language. UML is a standardized general-purpose modelling 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 Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artefacts of software system, as well as for business modelling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modelling 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.

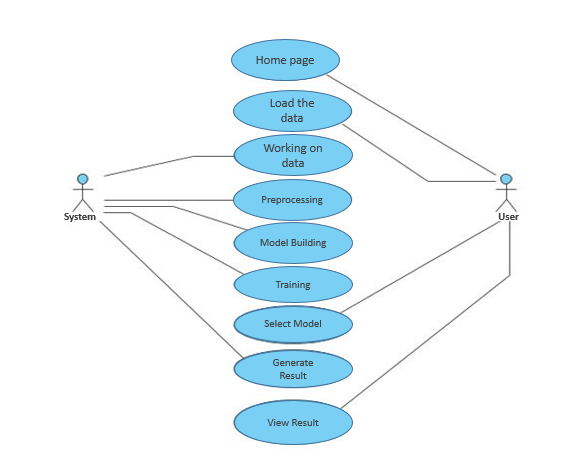
**GOALS:**

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

1. Provide users a ready-to-use, expressive visual modelling 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 modelling 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.

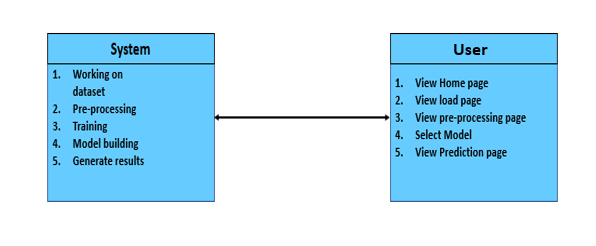
**USE CASE DIAGRAM**

* 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.

****

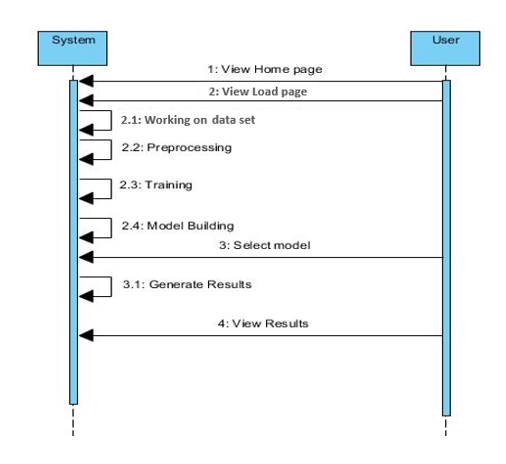
**CLASS DIAGRAM**

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



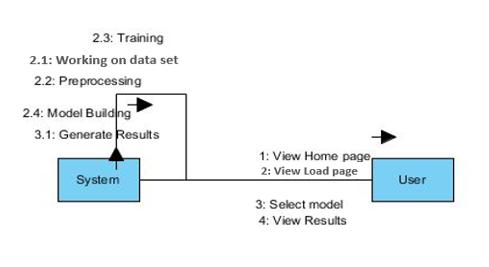
**SEQUENCE DIAGRAM**

* 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



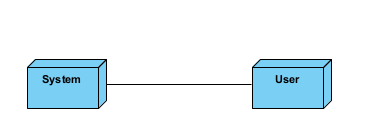
**COLLABORATION DIAGRAM:**

In collaboration diagram the method call sequence is indicated by some numbering technique as shown below. The number indicates how the methods are called one after another. We have taken the same order management system to describe the collaboration diagram. The method calls are similar to that of a sequence diagram. But the difference is that the sequence diagram does not describe the object organization whereas the collaboration diagram shows the object organization.



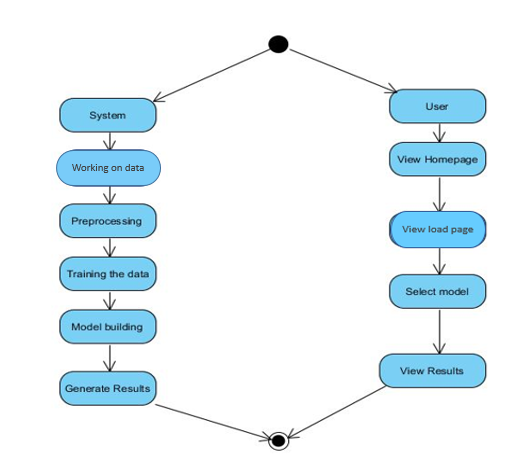
**DEPLOYMENT DIAGRAM**

Deployment diagram represents the deployment view of a system. It is related to the component diagram. Because the components are deployed using the deployment diagrams. A deployment diagram consists of nodes. Nodes are nothing but physical hardware’s used to deploy the application.



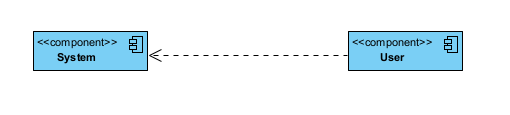
**ACTIVITY DIAGRAM:**

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling 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.



**COMPONENT DIAGRAM**:

A component diagram, also known as a UML component diagram, describes the organization and wiring of the physical **c**omponents in a system. Component diagrams are often drawn to help model implementation details and double-check that every aspect of the system's required function is covered by planned development.



**ER DIAGRAM:**

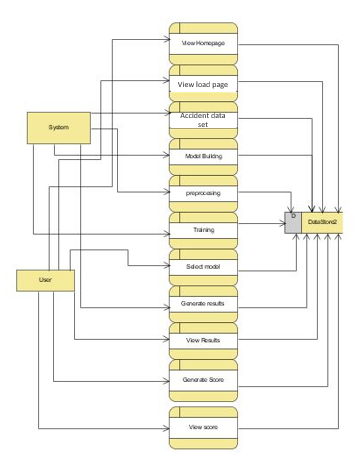
An Entity–relationship model (ER model) describes the structure of a database with the help of a diagram, which is known as Entity Relationship Diagram (ER Diagram). An ER model is a design or blueprint of a database that can later be implemented as a database. The main components of E-R model are: entity set and relationship set.

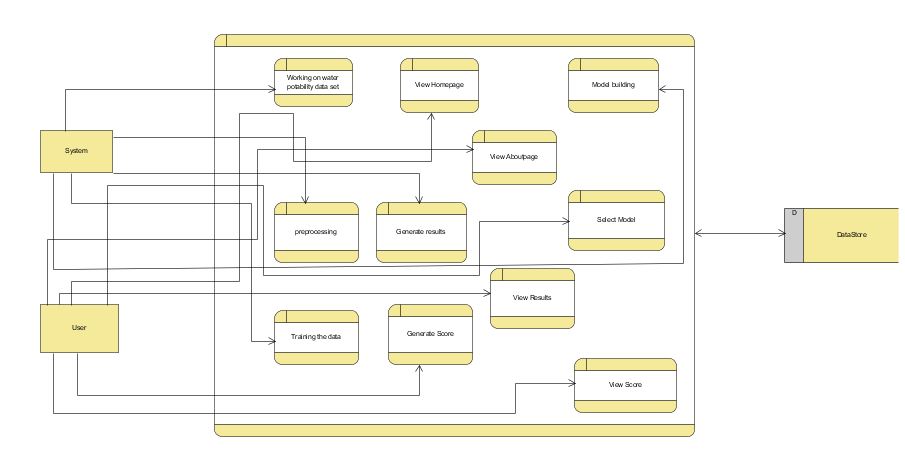
An ER diagram shows the relationship among entity sets. An entity set is a group of similar entities and these entities can have attributes. In terms of DBMS, an entity is a table or attribute of a table in database, so by showing relationship among tables and their attributes, ER diagram shows the complete logical structure of a database. Let’s have a look at a simple ER diagram to understand this concept.

# 

**DFD DIAGRAM:**

A Data Flow Diagram (DFD) is a traditional way to visualize the information flows within a system. A neat and clear DFD can depict a good amount of the system requirements graphically. It can be manual, automated, or a combination of both. It shows how information enters and leaves the system, what changes the information and where information is stored. The purpose of a DFD is to show the scope and boundaries of a system as a whole. It may be used as a communications tool between a systems analyst and any person who plays a part in the system that acts as the starting point for redesigning a system.

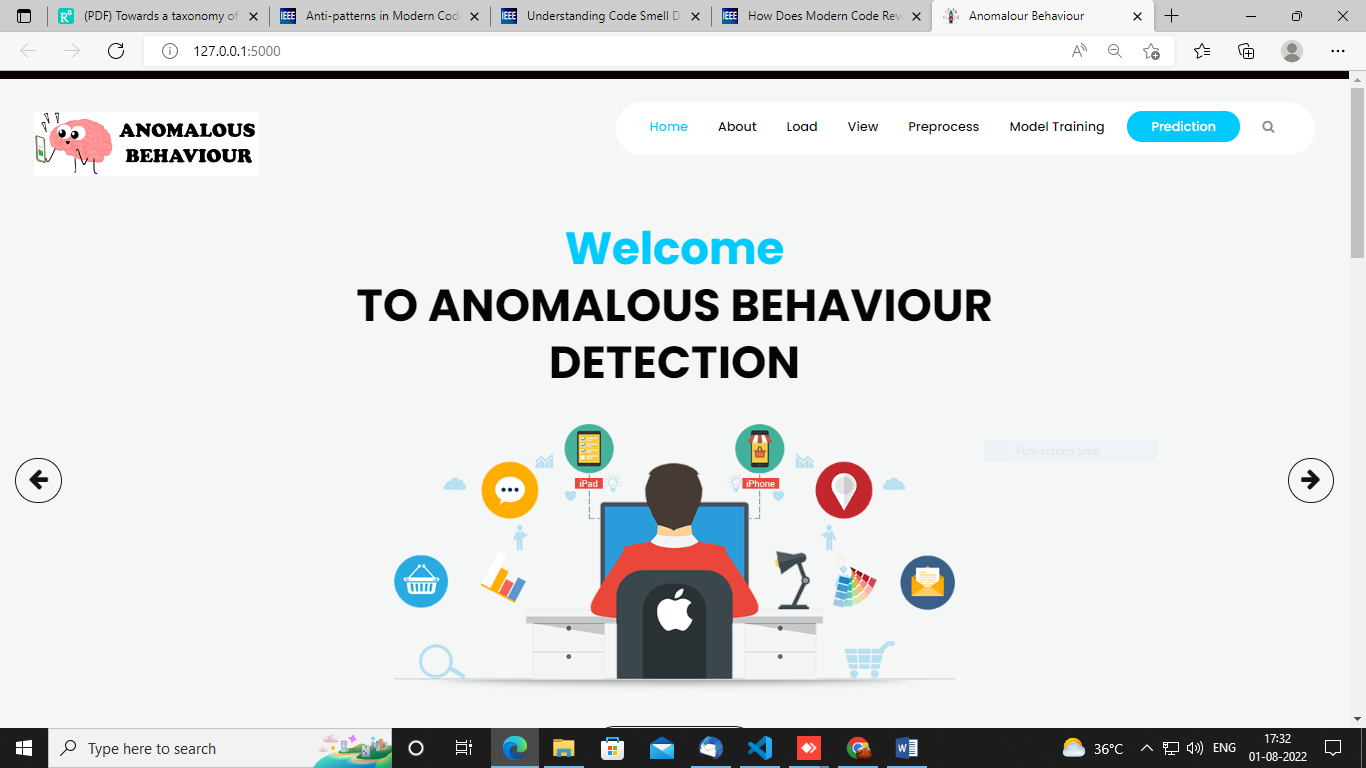
sss



**OUTPUT SCREEN SHOTS WITH DESCRIPTION.**

**Home Page:**

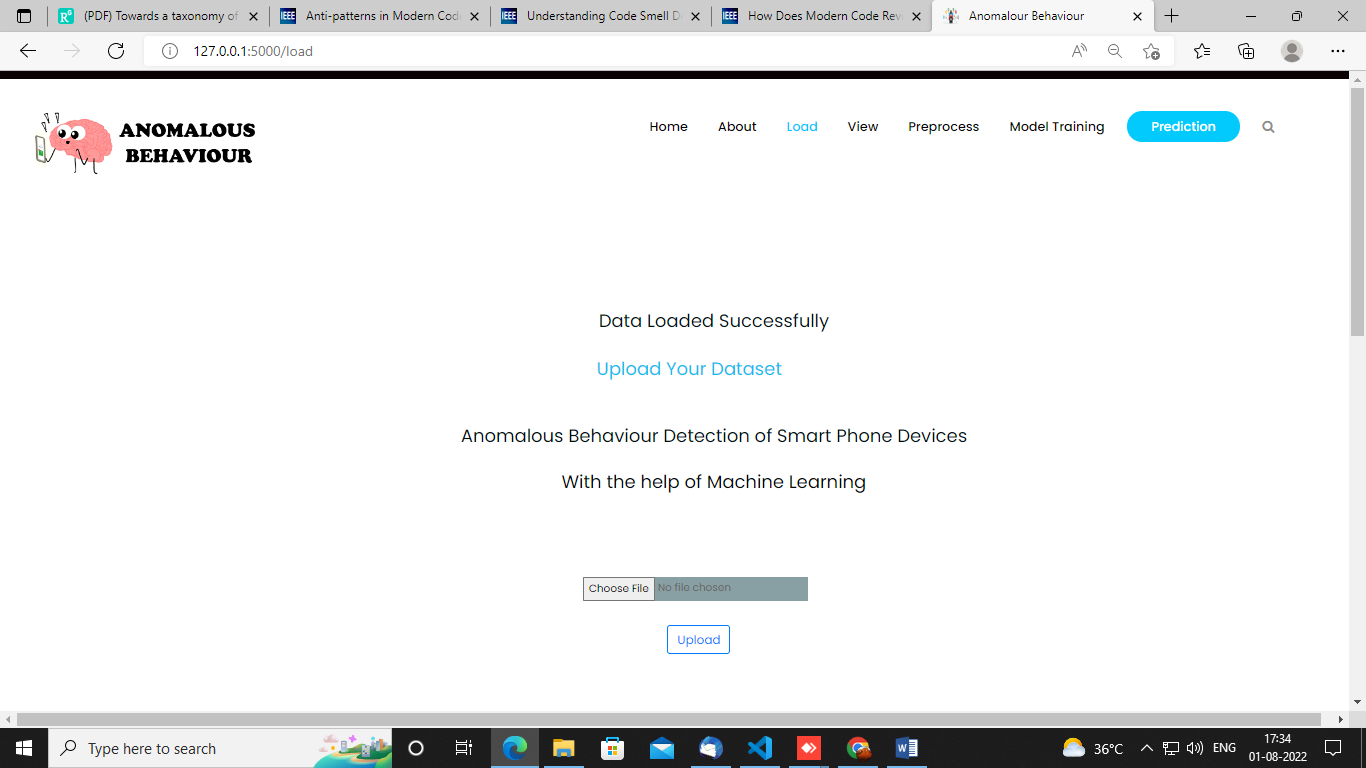
Here user view the home page of anomalous behaviour prediction web application.



**Fig1**: Home Page

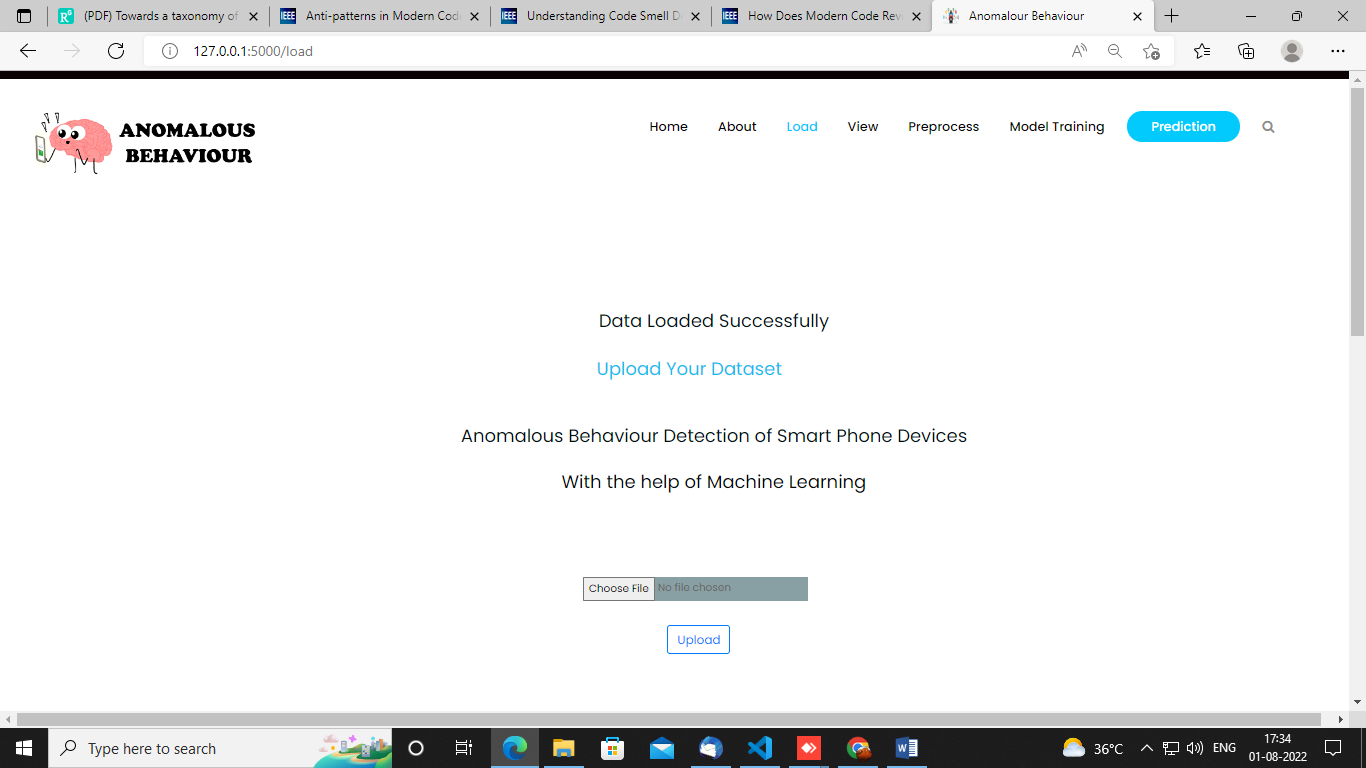
**Load:**

In the load page, users can load the anomalous behaviour dataset.

****

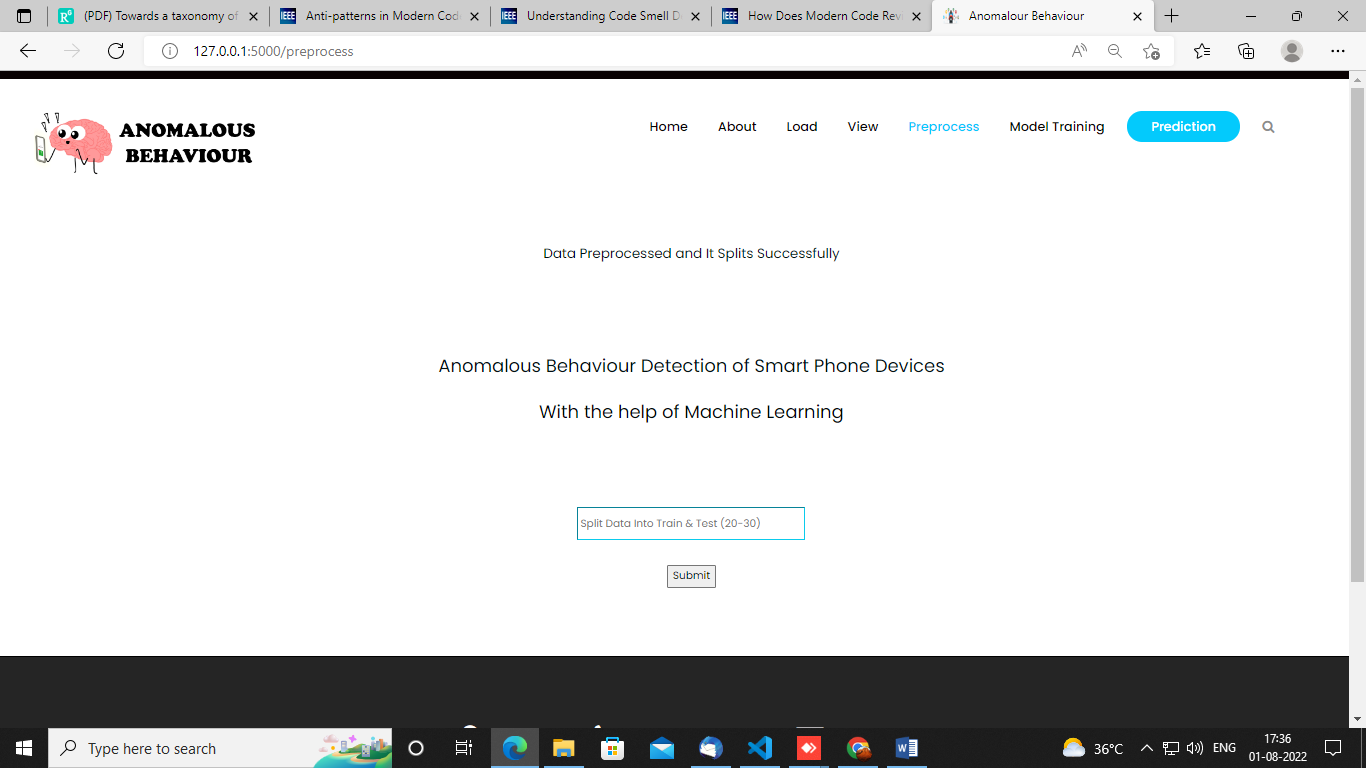
**View:**

Here we can see the uploaded data set.



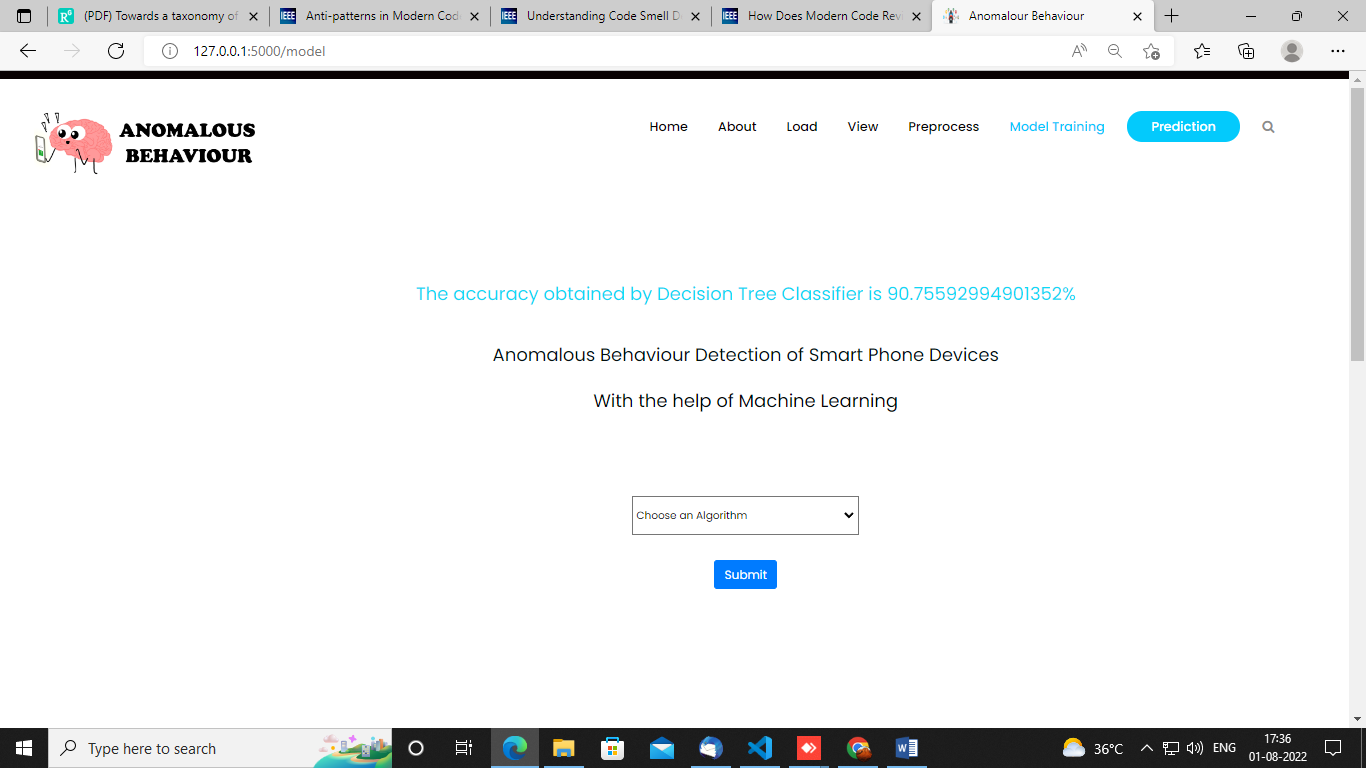
**Pre-process:**

Here we can prepare our data in such a way that our system should understand i.e, we will make our data noise free.



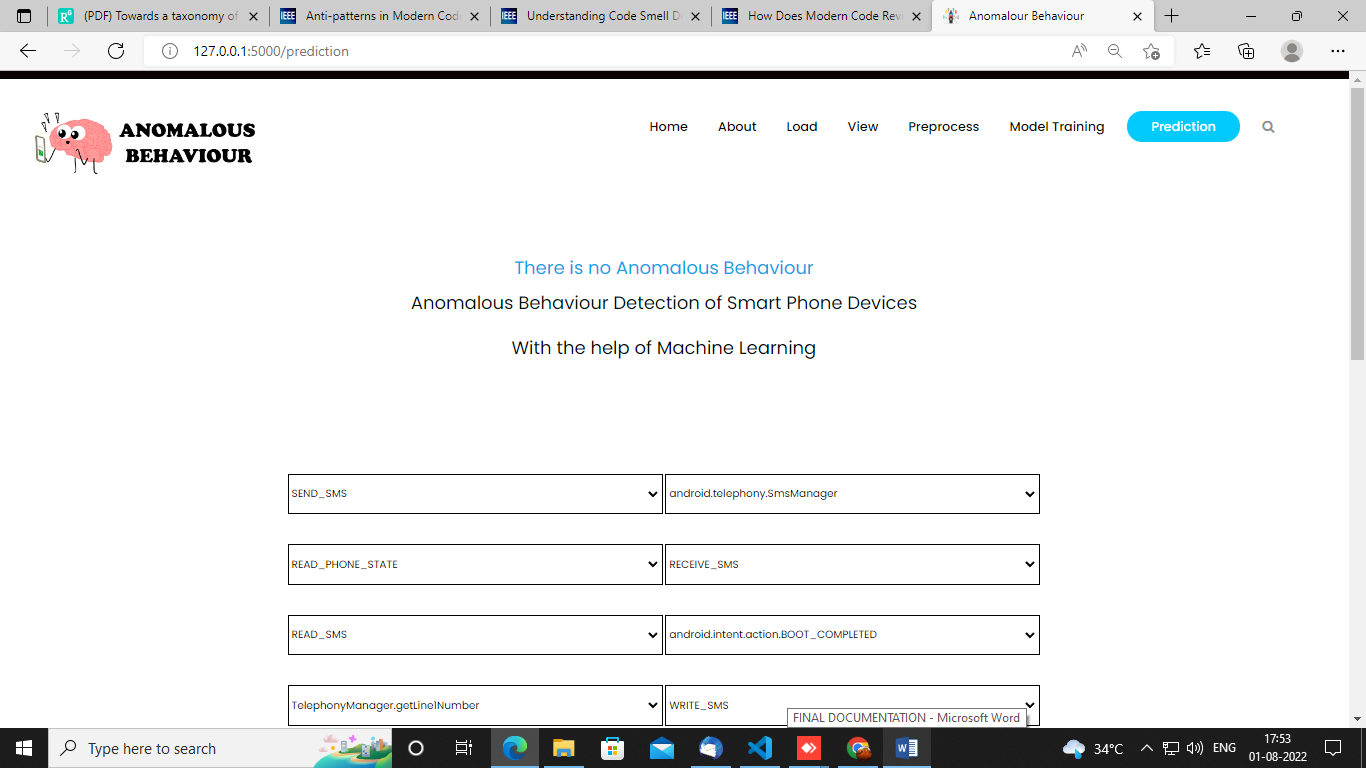
**Model:**

Here we can train our data using different algorithm.



**Prediction:**

This page show the detection result of the anomalous prediction data.



**CONCLUSION:**

In this paper, we proposed a new methodology to identify when smartphone applications behave anomalously. The work considers the use of a changepoint detection theory to extract features and three machine learning techniques to train a classifier from the power consumed by the smartphone. We can conclude that the proposed methodology performs better in terms of F1-measure accuracy comparing it with three other methodologies. We can emphasize that our methodology can recognize malware acting in short periods of time which it was a disadvantage of the other methodologies. As a future work, we aim to apply the methodology on real malwares instead of using an emulated malware.

**REFERENCES:**

[1] D. Evans, “The internet of things: How the next evolution of the internet is changing everything,” CISCO white paper, Tech. Rep., 2011.

[2] Statista, “Number of mobile phone users worldwide from 2015 to 2020 (in billions).” [Online]. Available: <https://www.statista.com/statistics/274774/forecast-ofmobile-phone-users-worldwide/>

[3] A. Arabo and B. Pranggono, “Mobile malware and smart device security: Trends, challenges and solutions,” in 2013 19th International Conference on Control Systems and Computer Science, May 2013, pp. 526–531.

[4] T. Kim, B. Kang, M. Rho, and et. all, “A multimodal deep learning method for android malware detection using various features,” IEEE Trans. on Info. Forensics and Security, vol. 14, no. 3, 2019.

[5] Y.-S. Yen and H.-M. Sun, “An android mutation malware detection based on deep learning using visualization of importance from codes,” Microelectronics Reliability, vol. 93, pp. 109–114, 2019.

[6] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. Siemens, “Drebin: Effective and explainable detection of android malware in your pocket.” in Ndss, vol. 14, 2014, pp. 23–26.

[7] P. Faruki, A. Bharmal, V. Laxmi, and et. all, “Android security: A survey of issues, malware penetration, and defenses,” IEEE Communications Surveys Tutorials, vol. 17, no. 2, pp. 998–1022, Secondquarter 2015.

[8] K. Ariyapala, H. G. Do, H. N. Anh, and et. all, “A host and network based intrusion detection for android smartphones,” in 30th Int. Conf. on Advanced Info. Net. and Apps Workshops (WAINA), March 2016.

[9] M. Curti, A. Merlo, M. Migliardi, and S. Schiappacasse, “Towards energy-aware intrusion detection systems on mobile devices,” in Int. Conf. on High Performance Computing Simulation (HPCS), July 2013.

[10] H. Kim, J. Smith, and K. G. Shin, “Detecting energy-greedy anomalies and mobile malware variants,” in Proceedings of the 6th international conference on Mobile systems, applications, and services. ACM, 2008.

**BIBLIOGRAPHY:**

**SOFTWARE INSTALLATION FOR MACHINE LEARNING PROJECTS:**

**Installing Python:**

1. To download and install Python visit the official website of Python <https://www.python.org/downloads/> and choose your version.



1. Once the download is complete, run the exe for install Python. Now click on Install Now.
2. You can see Python installing at this point.
3. When it finishes, you can see a screen that says the Setup was successful. Now click on "Close".

**Installing PyCharm:**

1. To download PyCharm visit the website <https://www.jetbrains.com/pycharm/download/> and click the "DOWNLOAD" link under the Community Section.



1. Once the download is complete, run the exe for install PyCharm. The setup wizard should have started. Click “Next”.
2. On the next screen, Change the installation path if required. Click “Next”.
3. On the next screen, you can create a desktop shortcut if you want and click on “Next”.
4. Choose the start menu folder. Keep selected JetBrains and click on “Install”.
5. Wait for the installation to finish.
6. Once installation finished, you should receive a message screen that PyCharm is installed. If you want to go ahead and run it, click the “Run PyCharm Community Edition” box first and click “Finish”.
7. After you click on "Finish," the Following screen will appear.



9. You need to install some packages to execute your project in a proper way.

10. Open the command prompt/ anaconda prompt or terminal as administrator.

11. The prompt will get open, with specified path, type “pip install package name” which you want to install (like NumPy, pandas, sea born, scikit-learn, Matplotlib, Pyplot)

Ex: Pip install NumPy



# **INTRODUCTION TO PYTHON**

* Python

### What Is a Script?

Up to this point, I have concentrated on the interactive programming capability of Python.  This is a very useful capability that allows you to type in a program and to have it executed immediately in an interactive mode

Scripts are reusable

Basically, a script is a text file containing the statements that comprise a Python program.  Once you have created the script, you can execute it over and over without having to retype it each time.

Scripts are editable

Perhaps, more importantly, you can make different versions of the script by modifying the statements from one file to the next using a text editor.  Then you can execute each of the individual versions.  In this way, it is easy to create different programs with a minimum amount of typing.

You will need a text editor

Just about any text editor will suffice for creating Python script files.

You can use *Microsoft Notepad, Microsoft WordPad, Microsoft Word,*or just about any word processor if you want to.

Difference between a script and a program

Script:

Scripts are distinct from the core code of the application, which is usually written in a different language, and are often created or at least modified by the end-user. Scripts are often interpreted from source code or byte code, whereas the applications they control are traditionally compiled to native machine code.

Program:

The program has an executable form that the computer can use directly to execute the instructions.

The same program in its human-readable source code form, from which executable programs are derived (e.g., compiled)

Python

What is Python? Chances you are asking yourself this. You may have found this book because you want to learn to program but don’t know anything about programming languages. Or you may have heard of programming languages like C, C++, C#, or Java and want to know what Python is and how it compares to “big name” languages. Hopefully I can explain it for you.

Python concepts

If you’re not interested in the how’s and whys of Python, feel free to skip to the next chapter. In this chapter I will try to explain to the reader why I think Python is one of the best languages available and why it’s a great one to start programming with.

• Open-source general-purpose language.

• Object Oriented, Procedural, Functional

• Easy to interface with C/ObjC/Java/Fortran

• Easy-is to interface with C++ (via SWIG)

• Great interactive environment

• Great interactive environment

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 interact with the interpreter directly to write your programs.
* Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
* Python is a Beginner's Language − Python is a great language for the beginner-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 Possum 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, and 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 Possum 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 clearly defined 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-maintained.
* A broad standard library − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
* Interactive Mode − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
* Portable − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
* Extendable − you can add low-level modules to the Python interpreter. These modules 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 and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
* Scalable − Python provides a better structure and support for large programs than shell scripting.

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.

Dynamic vs. Static

Types Python is a dynamic-typed language. Many other languages are static typed, such as C/C++ and Java. A static typed language requires the programmer to explicitly tell the computer what type of “thing” each data value is.

For example, in C if you had a variable that was to contain the price of something, you would have to declare the variable as a “float” type.

This tells the compiler that the only data that can be used for that variable must be a floating point number, i.e. a number with a decimal point.

If any other data value was assigned to that variable, the compiler would give an error when trying to compile the program.

Python, however, doesn’t require this. You simply give your variables names and assign values to them. The interpreter takes care of keeping track of what kinds of objects your program is using. This also means that you can change the size of the values as you develop the program. Say you have another decimal number (a.k.a. a floating point number) you need in your program.

With a static typed language, you have to decide the memory size the variable can take when you first initialize that variable. A double is a floating point value that can handle a much larger number than a normal float (the actual memory sizes depend on the operating environment).

If you declare a variable to be a float but later on assign a value that is too big to it, your program will fail; you will have to go back and change that variable to be a double.

With Python, it doesn’t matter. You simply give it whatever number you want and Python will take care of manipulating it as needed. It even works for derived values.

For example, say you are dividing two numbers. One is a floating point number and one is an integer. Python realizes that it’s more accurate to keep track of decimals so it automatically calculates the result as a floating point number

Variables

Variables are nothing but reserved memory locations to store values. This means that when you create a variable you reserve some space in memory.

Based on the data type of a variable, the interpreter allocates memory and decides what can be stored in the reserved memory. Therefore, by assigning different data types to variables, you can store integers, decimals or characters in these variables.

Standard Data Types

The data stored in memory can be of many types. For example, a person's age is stored as a numeric value and his or her address is stored as alphanumeric characters. Python has various standard data types that are used to define the operations possible on them and the storage method for each of them.

Python has five standard data types −

* Numbers
* String
* List
* Tuple
* Dictionary

## Python Numbers

Number data types store numeric values. Number objects are created when you assign a value to them

## Python Strings

Strings in Python are identified as a contiguous set of characters represented in the quotation marks. Python allows for either pairs of single or double quotes. Subsets of strings can be taken using the slice operator ([ ] and [:]) with indexes starting at 0 in the beginning of the string and working their way from -1 at the end.

## Python Lists

Lists are the most versatile of Python's compound data types. A list contains items separated by commas and enclosed within square brackets ([]). To some extent, lists are similar to arrays in C. One difference between them is that all the items belonging to a list can be of different data type.

The values stored in a list can be accessed using the slice operator ([ ] and [:]) with indexes starting at 0 in the beginning of the list and working their way to end -1. The plus (+) sign is the list concatenation operator, and the asterisk (\*) is the repetition operator.

## Python Tuples

A tuple is another sequence data type that is similar to the list. A tuple consists of a number of values separated by commas. Unlike lists, however, tuples are enclosed within parentheses.

The main differences between lists and tuples are: Lists are enclosed in brackets ([ ]) and their elements and size can be changed, while tuples are enclosed in parentheses (( )) and cannot be updated. Tuples can be thought of as read-only lists.

## Python Dictionary

Python's dictionaries are kind of hash table type. They work like associative arrays or hashes found in Perl and consist of key-value pairs. A dictionary key can be almost any Python type, but are usually numbers or strings. Values, on the other hand, can be any arbitrary Python object.

Dictionaries are enclosed by curly braces ({ }) and values can be assigned and accessed using square braces ([]).

Different modes in python

Python has two basic modes: normal and interactive.

The normal mode is the mode where the scripted and finished .pie files are run in the Python interpreter.

Interactive mode is a command line shell which gives immediate feedback for each statement, while running previously fed statements in active memory. As new lines are fed into the interpreter, the fed program is evaluated both in part and in whole

# 20 Python libraries

1. Requests. The most famous http library written by Kenneth remits. It’s a must have for every python developer.

2. Scrappy. If you are involved in web scraping then this is a must have library for you. After using this library you won’t use any other.

3. Python. A guy toolkit for python. I have primarily used it in place of tinder. You will really love it.

4. Pillow. A friendly fork of PIL (Python Imaging Library). It is more user friendly than PIL and is a must have for anyone who works with images.

5. SQL Alchemy. A database library. Many love it and many hate it. The choice is yours.

6. Beautiful Soup. I know it’s slow but this xml and html parsing library is very useful for beginners.

7. Twisted. The most important tool for any network application developer. It has a very beautiful ape and is used by a lot of famous python developers.

8. Numbly. How can we leave this very important library? It provides some advance math functionalities to python.

9. Skippy. When we talk about numbly then we have to talk about spicy. It is a library of algorithms and mathematical tools for python and has caused many scientists to switch from ruby to python.

10. Matplotlib. A numerical plotting library. It is very useful for any data scientist or any data analyser.

11. Pygmy. Which developer does not like to play games and develop them? This library will help you achieve your goal of 2d game development.

12. Piglet. A 3d animation and game creation engine. This is the engine in which the famous [python port](https://github.com/fogleman/Minecraft) of mine craft was made

13. Pit. A GUI toolkit for python. It is my second choice after python for developing GUI’s for my python scripts.

14. Pit. Another python GUI library. It is the same library in which the famous Bit torrent client is created.

15. Scaly. A packet sniffer and analyser for python made in python.

16. Pywin32. A python library which provides some useful methods and classes for interacting with windows.

17. Notch. Natural Language Toolkit – I realize most people won’t be using this one, but it’s generic enough. It is a very useful library if you want to manipulate strings. But its capacity is beyond that. Do check it out.

18. Nose. A testing framework for python. It is used by millions of python developers. It is a must have if you do test driven development.

19. Simply. Simply can-do algebraic evaluation, differentiation, expansion, complex numbers, etc. It is contained in a pure Python distribution.

20. I Python. I just can’t stress enough how useful this tool is. It is a python prompt on steroids. It has completion, history, shell capabilities, and a lot more. Make sure that you take a look at it.

NumPy

Humpy’s main object is the homogeneous multidimensional array. It is a table of elements (usually numbers), all of the same type, indexed by a tuple of positive integers. In numbly dimensions are called axes. The number of axes is rank.

• Offers Matlab-ish capabilities within Python

• Fast array operations

• 2D arrays, multi-D arrays, linear algebra etc.

Matplotlib

• High quality plotting library.

Python class and objects

These are the building blocks of OOP. Class creates a new object. This object can be anything, whether an abstract data concept or a model of a physical object, e.g. a chair. Each class has individual characteristics unique to that class, including variables and methods. Classes are very powerful and currently “the big thing” in most programming languages. Hence, there are several chapters dedicated to OOP later in the book.

The class is the most basic component of object-oriented programming. Previously, you learned how to use functions to make your program do something.

Now will move into the big, scary world of Object-Oriented Programming (OOP). To be honest, it took me several months to get a handle on objects.

When I first learned C and C++, I did great; functions just made sense for me.

Having messed around with BASIC in the early ’90s, I realized functions were just like subroutines so there wasn’t much new to learn.

However, when my C++ course started talking about objects, classes, and all the new features of OOP, my grades definitely suffered.

Once you learn OOP, you’ll realize that it’s actually a pretty powerful tool. Plus many Python libraries and APIs use classes, so you should at least be able to understand what the code is doing.

One thing to note about Python and OOP: it’s not mandatory to use objects in your code in a way that works best; maybe you don’t need to have a full-blown class with initialization code and methods to just return a calculation. With Python, you can get as technical as you want.

As you’ve already seen, Python can do just fine with functions. Unlike languages such as Java, you aren’t tied down to a single way of doing things; you can mix functions and classes as necessary in the same program. This lets you build the code

Objects are an encapsulation of variables and functions into a single entity. Objects get their variables and functions from classes. Classes are essentially a template to create your objects.

Here’s a brief list of Python OOP ideas:

• The class statement creates a class object and gives it a name. This creates a new namespace.

• Assignments within the class create class attributes. These attributes are accessed by qualifying the name using dot syntax: ClassName.Attribute.

• Class attributes export the state of an object and its associated behaviour. These attributes are shared by all instances of a class.

• Calling a class (just like a function) creates a new instance of the class.

This is where the multiple copy’s part comes in.

• Each instance gets ("inherits") the default class attributes and gets its own namespace. This prevents instance objects from overlapping and confusing the program.

• Using the term self identifies a particular instance, allowing for per-instance attributes. This allows items such as variables to be associated with a particular instance.

Inheritance

First off, classes allow you to modify a program without really making changes to it.

To elaborate, by sub classing a class, you can change the behaviour of the program by simply adding new components to it rather than rewriting the existing components.

As we’ve seen, an instance of a class inherits the attributes of that class.

However, classes can also inherit attributes from other classes. Hence, a subclass inherits from a superclass allowing you to make a generic superclass that is specialized via subclasses.

The subclasses can override the logic in a superclass, allowing you to change the behaviour of your classes without changing the superclass at all.

Operator Overloads

Operator overloading simply means that objects that you create from classes can respond to actions (operations) that are already defined within Python, such as addition, slicing, printing, etc.

Even though these actions can be implemented via class methods, using overloading ties the behavior closer to Python’s object model and the object interfaces are more consistent to Python’s built-in objects, hence overloading is easier to learn and use.

User-made classes can override nearly all of Python’s built-in operation methods

Exceptions

I’ve talked about exceptions before but now I will talk about them in depth. Essentially, exceptions are events that modify program’s flow, either intentionally or due to errors.

They are special events that can occur due to an error, e.g. trying to open a file that doesn’t exist, or when the program reaches a marker, such as the completion of a loop.

Exceptions, by definition, don’t occur very often; hence, they are the "exception to the rule" and a special class has been created for them. Exceptions are everywhere in Python.

Virtually every module in the standard Python library uses them, and Python itself will raise them in a lot of different circumstances.

Here are just a few examples:

• Accessing a non−existent dictionary key will raise a Key Error exception.

• Searching a list for a non−existent value will raise a Value Error exception

. • Calling a non−existent method will raise an Attribute Error exception.

• Referencing a non−existent variable will raise a Name Error exception.

• Mixing data types without coercion will raise a Type Error exception.

One use of exceptions is to catch a fault and allow the program to continue working; we have seen this before when we talked about files.

This is the most common way to use exceptions. When programming with the Python command line interpreter, you don’t need to worry about catching exceptions.

Your program is usually short enough to not be hurt too much if an exception occurs.

Plus, having the exception occur at the command line is a quick and easy way to tell if your code logic has a problem.

However, if the same error occurred in your real program, it will fail and stop working. Exceptions can be created manually in the code by raising an exception.

It operates exactly as a system-caused exceptions, except that the programmer is doing it on purpose. This can be for a number of reasons. One of the benefits of using exceptions is that, by their nature, they don’t put any overhead on the code processing.

Because exceptions aren’t supposed to happen very often, they aren’t processed until they occur.

Exceptions can be thought of as a special form of the if/elf statements. You can realistically do the same thing with if blocks as you can with exceptions.

However, as already mentioned, exceptions aren’t processed until they occur; if blocks are processed all the time.

Proper use of exceptions can help the performance of your program.

The more infrequent the error might occur, the better off you are to use exceptions; using if blocks requires Python to always test extra conditions before continuing.

Exceptions also make code management easier: if your programming logic is mixed in with error-handling if statements, it can be difficult to read, modify, and debug your program.

User-Defined Exceptions

I won’t spend too much time talking about this, but Python does allow for a programmer to create his own exceptions.

You probably won’t have to do this very often but it’s nice to have the option when necessary.

However, before making your own exceptions, make sure there isn’t one of the built-in exceptions that will work for you.

They have been "tested by fire" over the years and not only work effectively, they have been optimized for performance and are bug-free.

Making your own exceptions involves object-oriented programming, which will be covered in the next chapter

. To make a custom exception, the programmer determines which base exception to use as the class to inherit from, e.g. making an exception for negative numbers or one for imaginary numbers would probably fall under the Arithmetic Error exception class.

To make a custom exception, simply inherit the base exception and define what it will do.

Python modules

Python allows us to store our code in files (also called modules). This is very useful for more serious programming, where we do not want to retype a long function definition from the very beginning just to change one mistake. In doing this, we are essentially defining our own modules, just like the modules defined already in the Python library.

To support this, Python has a way to put definitions in a file and use them in a script or in an interactive instance of the interpreter. Such a file is called a module; definitions from a module can be imported into other modules or into the main module.

Testing code

As indicated above, code is usually developed in a file using an editor.

To test the code, import it into a Python session and try to run it.

Usually there is an error, so you go back to the file, make a correction, and test again.

This process is repeated until you are satisfied that the code works. T

His entire process is known as the development cycle.

There are two types of errors that you will encounter. Syntax errors occur when the form of some command is invalid.

This happens when you make typing errors such as misspellings, or call something by the wrong name, and for many other reasons. Python will always give an error message for a syntax error.

Functions in Python

It is possible, and very useful, to define our own functions in Python. Generally speaking, if you need to do a calculation only once, then use the interpreter. But when you or others have need to perform a certain type of calculation many times, then define a function.

You use functions in programming to bundle a set of instructions that you want to use repeatedly or that, because of their complexity, are better self-contained in a sub-program and called when needed. That means that a function is a piece of code written to carry out a specified task.

## To carry out that specific task, the function might or might not need multiple inputs. When the task is carved out, the function can or cannot return one or more values.

## There are three types of functions in python:

## Help (), min (), print ().

Namespaces in Python are implemented as Python dictionaries, this means it is a mapping from names (keys) to objects (values). The user doesn't have to know this to write a Python program and when using namespaces.

Some namespaces in Python:

* global names of a module
* local names in a function or method invocation
* built-in names: this namespace contains built-in functions (e.g. abs(), camp(), ...) and built-in exception names

Garbage Collection

Garbage Collector exposes the underlying memory management mechanism of Python, the automatic garbage collector. The module includes functions for controlling how the collector operates and to examine the objects known to the system, either pending collection or stuck in reference cycles and unable to be freed.

Python XML Parser

XML is a portable, open source language that allows programmers to develop applications that can be read by other applications, regardless of operating system and/or developmental language.

What is XML? The Extensible Markup Language XML is a markup language much like HTML or SGML.

This is recommended by the World Wide Web Consortium and available as an open standard.

XML is extremely useful for keeping track of small to medium amounts of data without requiring a SQL-based backbone.

XML Parser Architectures and APIs the Python standard library provides a minimal but useful set of interfaces to work with XML.

The two most basic and broadly used APIs to XML data are the SAX and DOM interfaces.

Simple API for XML SAX: Here, you register callbacks for events of interest and then let the parser proceed through the document.

This is useful when your documents are large or you have memory limitations, it parses the file as it reads it from disk and the entire file is never stored in memory.

Document Object Model DOM API : This is a World Wide Web Consortium recommendation wherein the entire file is read into memory and stored in a hierarchical tree − based form to represent all the features of an XML document.

SAX obviously cannot process information as fast as DOM can when working with large files. On the other hand, using DOM exclusively can really kill your resources, especially if used on a lot of small files.

SAX is read-only, while DOM allows changes to the XML file. Since these two different APIs literally complement each other, there is no reason why you cannot use them both for large projects.

Python Web Frameworks

A web framework is a code library that makes a developer's life easier when building reliable, scalable and maintainable web applications.

## Why are web frameworks useful?

Web frameworks encapsulate what developers have learned over the past twenty years while programming sites and applications for the web. Frameworks make it easier to reuse code for common HTTP operations and to structure projects so other developers with knowledge of the framework can quickly build and maintain the application.

Common web framework functionality

Frameworks provide functionality in their code or through extensions to perform common operations required to run web applications. These common operations include:

1. URL routing
2. HTML, XML, JSON, and other output format tinplating
3. Database manipulation
4. Security against Cross-site request forgery (CSRF) and other attacks
5. Session storage and retrieval

Not all web frameworks include code for all of the above functionality. Frameworks fall on the spectrum from executing a single use case to providing every known web framework feature to every developer. Some frameworks take the "batteries-included" approach where everything possible comes bundled with the framework while others have a minimal core package that is amenable to extensions provided by other packages.

## Comparing web frameworks

There is also a repository called [compare-python-web-frameworks](https://github.com/mattmakai/compare-python-web-frameworks) where the same web application is being coded with varying Python web frameworks, tinplating engines and object.

## Web framework resources

* When you are learning how to use one or more web frameworks it's helpful to have an idea of what the code under the covers is doing.
* Frameworks is a really well done short video that explains how to choose between web frameworks. The author has some particular opinions about what should be in a framework. For the most part I agree although I've found sessions and database ORMs to be a helpful part of a framework when done well.
* What is a web framework? Is an in-depth explanation of what web frameworks are and their relation to web servers?
* Jingo vs. Flash vs. Pyramid: Choosing a Python web framework contains background information and code comparisons for similar web applications built in these three big Python frameworks.
* This fascinating blog post takes a look at the code complexity of several Python web frameworks by providing visualizations based on their code bases.
* Python’s web frameworks benchmarks  is a test of the responsiveness of a framework with encoding an object to JSON and returning it as a response as well as retrieving data from the database and rendering it in a template. There were no conclusive results but the output is fun to read about nonetheless.
* What web frameworks do you use and why are they awesome? Is a language agnostic Reedit discussion on web frameworks? It's interesting to see what programmers in other languages like and dislike about their suite of web frameworks compared to the main Python frameworks.
* This user-voted question & answer site asked "What are the best general purpose Python web frameworks usable in production?” The votes aren't as important as the list of the many frameworks that are available to Python developers.

## Web frameworks learning checklist

1. Choose a major Python web framework (Jingo or Flask are recommended) and stick with it. When you're just starting it's best to learn one framework first instead of bouncing around trying to understand every framework.
2. Work through a detailed tutorial found within the resources links on the framework's page.
3. Study open source examples built with your framework of choice so you can take parts of those projects and reuse the code in your application.
4. Build the first simple iteration of your web application then go to the [deployment](https://www.fullstackpython.com/deployment.html) section to make it accessible on the web.

2. SYSTEM STUDY

### SYSTEM TESTING

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.

6.1 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.

# 6.2 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.

6.3 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.