

**AN EFFICIENT SPAM DETECTION TECHNIQUE FOR IOT DEVICES
USING MACHINE LEARNING**

**Project Report
at**

**Sathyabama Institute of Science and Technology
(Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

Mr.Mangali BharathKumar(38110295)
Mr.Rapol Srikanth(38110554)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI -600119

MARCH - 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of Mr.MANGALIBHARATHKUMAR(38110295),Mr.RAPOLSRIKANTH(38110554) who carried out the project entitled “ AN EFFICIENT SPAM DETECTION TECHNIQUE FOR IOT DEVICES USING MACHINE LEARNING” under my supervision from October 2021 to May 2022.

InternalGuide

(Dr. R. AROUL CANESSANE M.E., Ph.D)

Head of the Department

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

I Mr.Mangali Bharath kumar, Mr.Rapol Srikanth hereby declare that the Project Report entitled **AN EFFICIENT SPAM DETECTION TECHNIQUE FOR IOT DEVICES USING MACHINE LEARNING** done by me under the guidance of **Dr. R. AROUL CANESSANE M.E., Ph.D** (Internal) is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering / Technology degree In COMPUTER SCIENCE AND TECHNOLOGY.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D, Dean**, School of Computing **Dr. L. Lakshmanan M.E., Ph.D. ,** and **Dr.S.Vigneshwari M.E., Ph.D. Heads** of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. R. AROUL CANESSANE M.E., Ph.D** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

AN EFFICIENT SPAM DETECTION TECHNIQUE FOR IOT DEVICES USING MACHINE LEARNING

ABSTRACT

The Internet of Things (IoT) is a group of millions of devices having sensors and actuators linked over wired or wireless channel for data transmission. IoT has grown rapidly over the past decade with more than 25 billion devices are expected to be connected by 2020. The volume of data released from these devices will increase many-fold in the years to come. In addition to an increased volume, the IoT devices produces a large amount of data with a number of different modalities having varying data quality defined by its speed in terms of time and position dependency. In such an environment, machine learning algorithms can play an important role in ensuring security and authorization based on biotechnology, anomalous detection to improve the usability and security of IoT systems. On the other hand, attackers often view learning algorithms to exploit the vulnerabilities in smart IoT-based systems. Motivated from these, in this paper, we propose the security of the IoT devices by detecting spam using machine learning. To achieve this objective, Spam Detection in IoT using Machine Learning framework is proposed. In this framework, five machine learning models are evaluated using various metrics with a large collection of inputs features sets. Each model computes a spam score by considering the refined input features. This score depicts the trustworthiness of IoT device under various parameters. REFIT Smart Home dataset is used for the validation of proposed technique. The results obtained proves the effectiveness of the proposed scheme in comparison to the other existing schemes.

CONTEXT

Chapter	list of content	Page no.
	ABSTRACT	i
	LIST OF FIGURES	ii
1	INTRODUCTION	1
1.1	PROPOSED SYSTEM	1
2	LITERATURE SURVEY	19
3	SYSTEM REQUIREMENTS	24
3.1	HARDWARE REQUIREMENTS	24
3.2	SOFTWARE REQUIREMENTS	24
3.3	LANGUAGE SPECIFICATIONS	24
3.4	HISTORY OF PYTHON	25
3.5	APPLICATION OF PYTHON	25
3.6	FEATURES OF PYTHON	26
3.7	FEASIBILITY STUDY	26
3.7.1	ECONOMICAL FEASIBILITY	27
3.7.2	TECHNICAL FEASIBILITY	27
3.7.3	SOCIAL FEASIBILITY	28

4	SYSTEM ANALYSIS	29
4.1	PURPOSE	29
4.2	SCOPE	29
4.3	EXISTING SYSTEM	29
4.4	PROPOSED SYETEM	30
5	SYSTEM DESIGN	32
5.1	INPUT DESIGN	32
5.2	OUTPUT DESIGN	32
5.3	DATA FLOW DIAGRAM	33
6	MODULES	39
6.1	LOGIN MODULE	39
6.2	DATA COLLECTION MODULE	39
6.3	PRE-PROCESSING MODULE	40
6.4	TRAIN AND TEST MODULE	40
6.5	DETECTION OF SPAM	41
7	SYSTEM IMPLEMENTATION	42
7.1	SYSTEM ARCHITECTURE	42
8	SYSTEM TESTING	43
8.1	TEST PLAN	43
8.2	VERIFICATION	43

8.3	VALIDATION	43
8.4	BASIC OF SOFTWARE TESTING	43
8.5	BLACK BOX TESTING	44
8.6	WHITE BOX TESTING	44
8.7	TYPES OF TESTING	44
8.7.1	UNIT TESTING	45
8.7.2	INTEGRATION TESTING	45
8.7.3	FUNCTIONAL TESTING	45
8.7.4	SYSTEM TESTING	45
8.7.5	STRESS TESTING	46
8.7.6	PERFORMANCE TESTING	46
8.7.7	USABILITY TESTING	46
8.7.8	ACCEPTANCE TESTING	46
8.7.9	REGRESSION TESTING	47
9	RESULT AND DISCUSSION	51
10	CONCLUSION	61
11	REFERENCE	62

LIST OF FIGURES:

FIGURE NO. NO.	FIGURE NAME	PAGE
1.1	EXPLAINING THE WORKING ALGORITHM OF THE RANDOM FOREST ALGORITHM	3
1.2	EXPLAINING THE RANDOM FOREST CLASSIFIER	5
1.3	EXPLAINING THE RANDOM FOREST CLASSIFIER CLASSIFIERALGORITHM WITH EXAMPLE	6
7.1	SYSTEM ARCHITECTURE	42
9.1	HOME PAGE	51
9.2	LOGIN PAGE	52
9.3	UPLOAD PAGE	54
9.4	PREVIEW PAGE	55
9.5	TRAINED DATA	56
9.6	PREDICTION NO SPAM	57
9.7	PREDICTION SPAM	58
9.8	PERFORMANCE ANALYSIS	59
9.9	CONFUSION MATRIX	59
9.10	CHART ANALYSIS	60

CHAPTER 1

INTRODUCTION

The safety measures of IoT devices depends upon the size and type of organization in which it is imposed. The behavior of users forces the security gateways to cooperate. In other words, we can say that the location, nature, application of IoT devices decides the security measures. For instance, the smart IoT security cameras in the smart organization can capture the different parameters for analysis and intelligent decision making. The maximum care to be taken is with web based devices as maximum number of IoT devices are web dependent. It is common at the workplace that the IoT devices installed in an organization can be used to implement security and privacy features efficiently. For example, wearable devices collect and send user's health data to a connected smartphone should prevent leakage of information to ensure privacy. It has been found in the market that 25-30% of working employees connect their personal IoT devices with the organizational network. The expanding nature of IoT attracts both the audience, i.e., the users and the attackers. However, with the emergence of ML in various attacks scenarios, IoT devices choose a defensive strategy and decide the key parameters in the security protocols for trade-off between security, privacy and computation. This job is challenging as it is usually difficult for an IoT system with limited resources to estimate the current network and timely attack status.

1.1 PROPOSED ALGORITHM

RANDOM FOREST ALGORITHM

Random forest algorithm can use both for classification and the regression kind of problems. In this you are going to learn, how the **random forest algorithm** works in machine learning for the classification task.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning**, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

A random forest algorithm consists of many decision trees. The ‘forest’ generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The below diagram explains the working of the Random Forest algorithm:

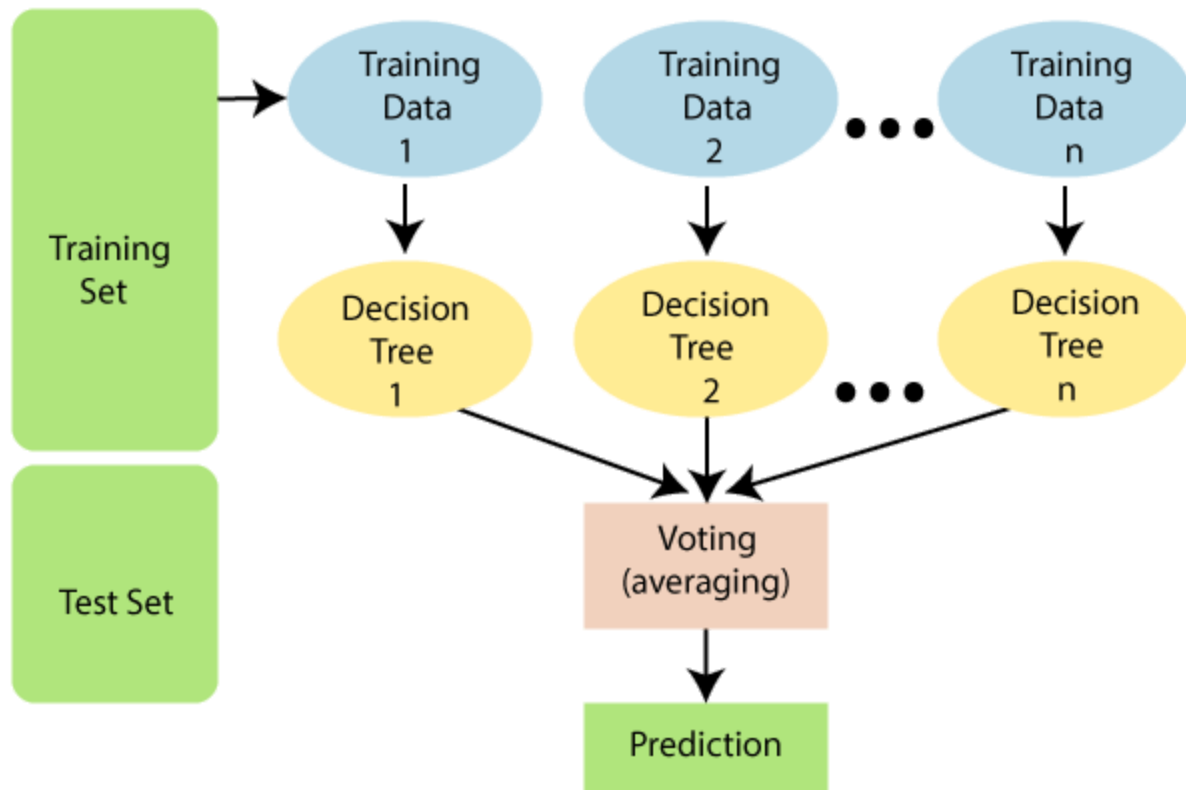


Fig 1.1: Explaining the working algorithm of the Random Forest algorithm

Below are some points that explain why we should use the Random Forest algorithm:

- o It takes less training time as compared to other algorithms.
- o It predicts output with high accuracy, even for the large dataset it runs efficiently.
- o It can also maintain accuracy when a large proportion of data is missing.

Features of a Random Forest Algorithm

- It's more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.

- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of overfitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point.

Classification in random forests

Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes.

A rain forest system relies on various decision trees. Every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system. In this case, the output chosen by the majority of the decision trees becomes the final output of the rain forest system. The diagram below shows a simple random forest classifier.

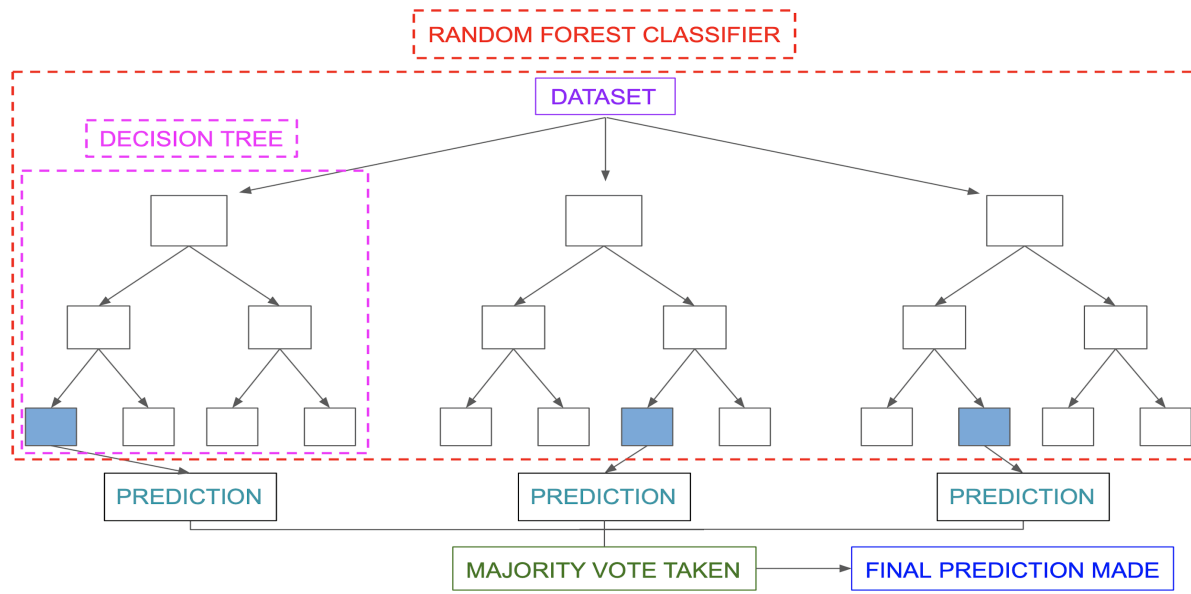


Fig 1.2: Explaining the Random Forest Classifier

Random Forest Steps

1. Randomly select “k” features from total “m” features.
 1. Where $k \ll m$
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, you can observe that we are randomly taking features and observations.

The working of the algorithm can be better understood by the below example:

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:

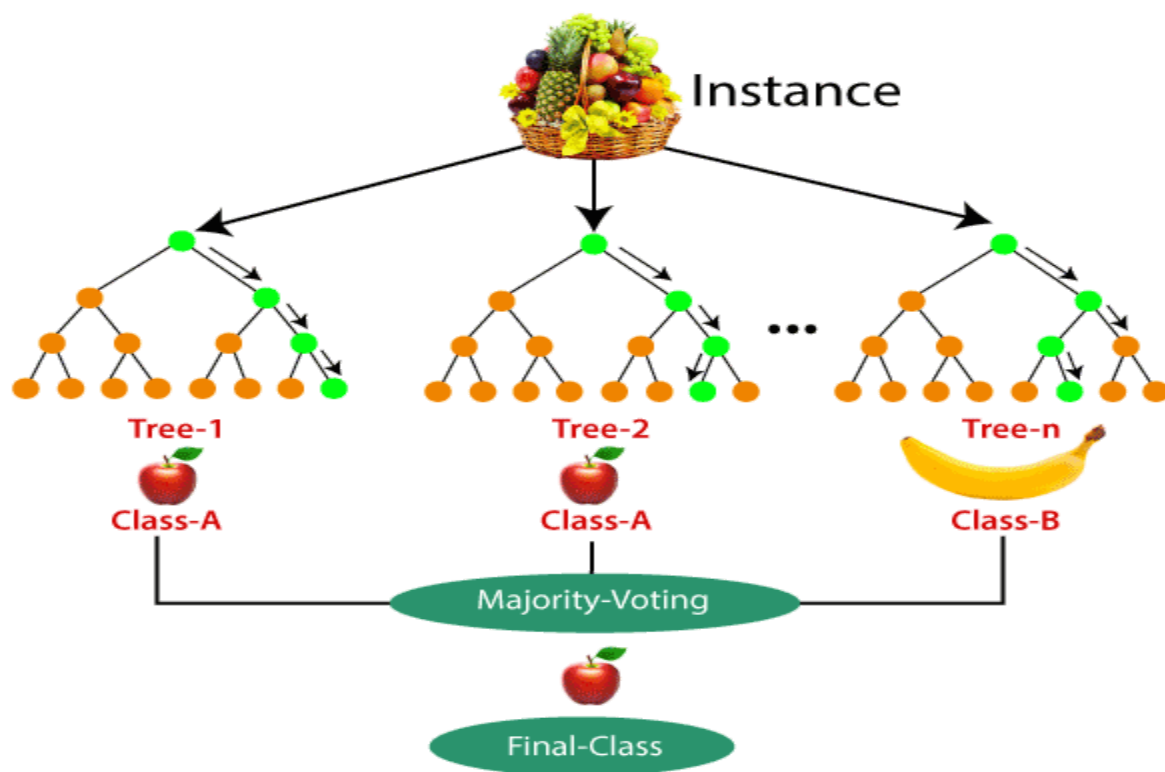


Fig 1.3: Explaining the Random Forest Classifier algorithm with example

APPLICATIONS OF RANDOM FOREST

There are mainly four sectors where Random forest mostly used:

1. **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
2. **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
3. **Land Use:** We can identify the areas of similar land use by this algorithm.
4. **Marketing:** Marketing trends can be identified using this algorithm.

Advantages of Random Forest

Random Forest is capable of performing both Classification and Regression tasks.

- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

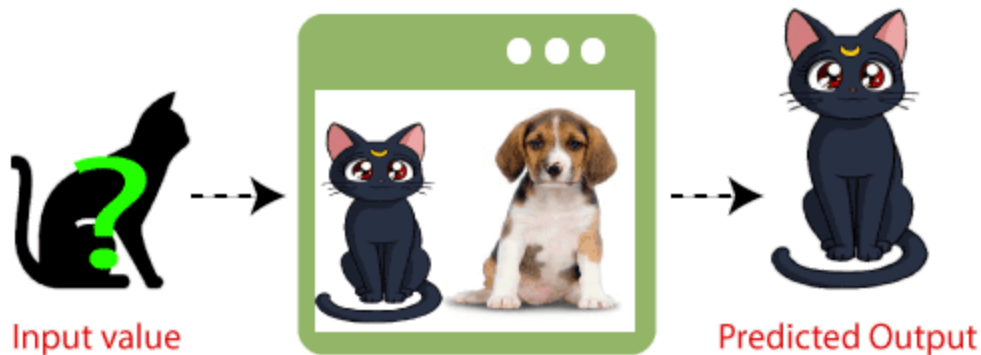
- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

KNN ALGORITHMS

- o K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- o K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

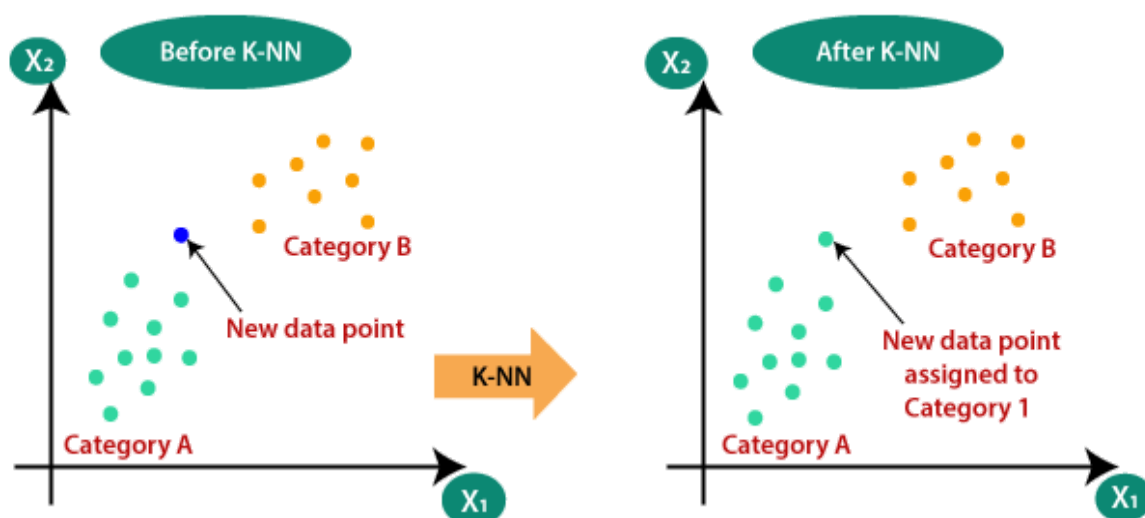
- o K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- o K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- o K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- o It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- o KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- o **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

KNN Classifier



Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x_1 , so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:

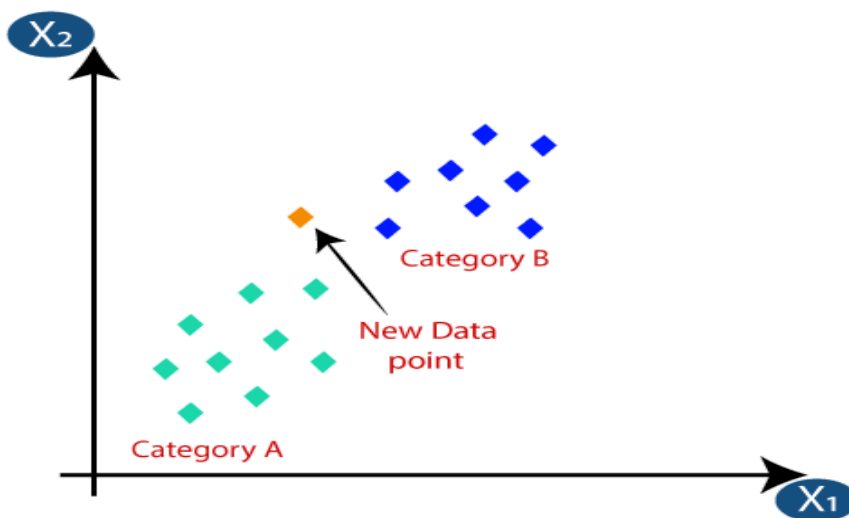


How does K-NN work?

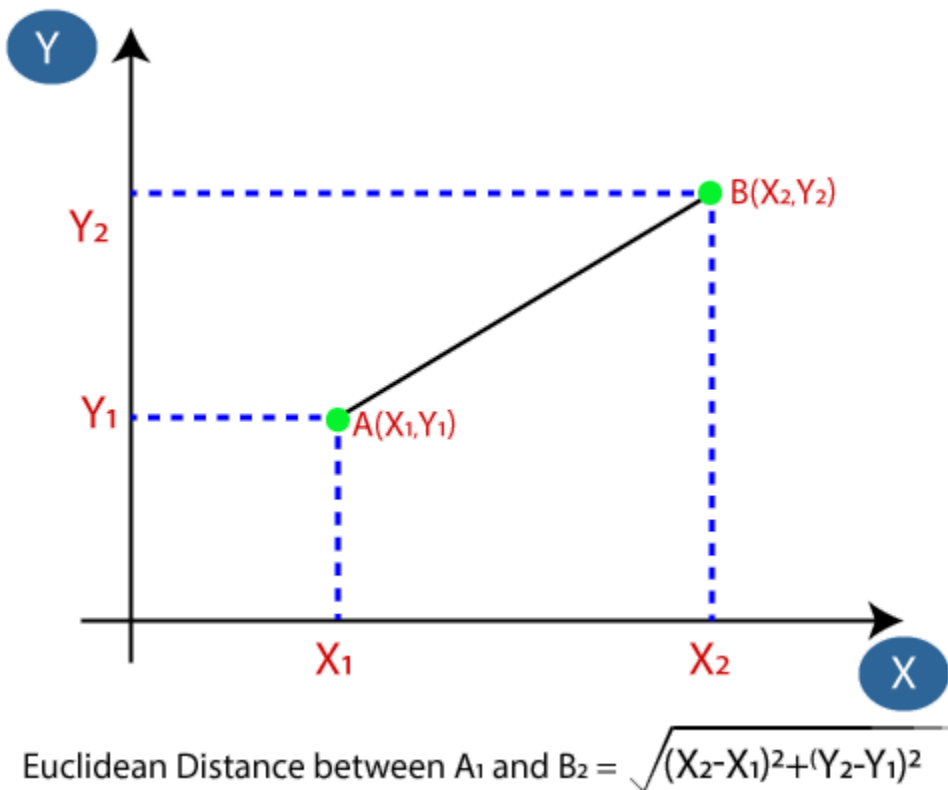
The K-NN working can be explained on the basis of the below algorithm:

- o **Step-1:** Select the number K of the neighbors
- o **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- o **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- o **Step-4:** Among these k neighbors, count the number of the data points in each category.
- o **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- o **Step-6:** Our model is ready.

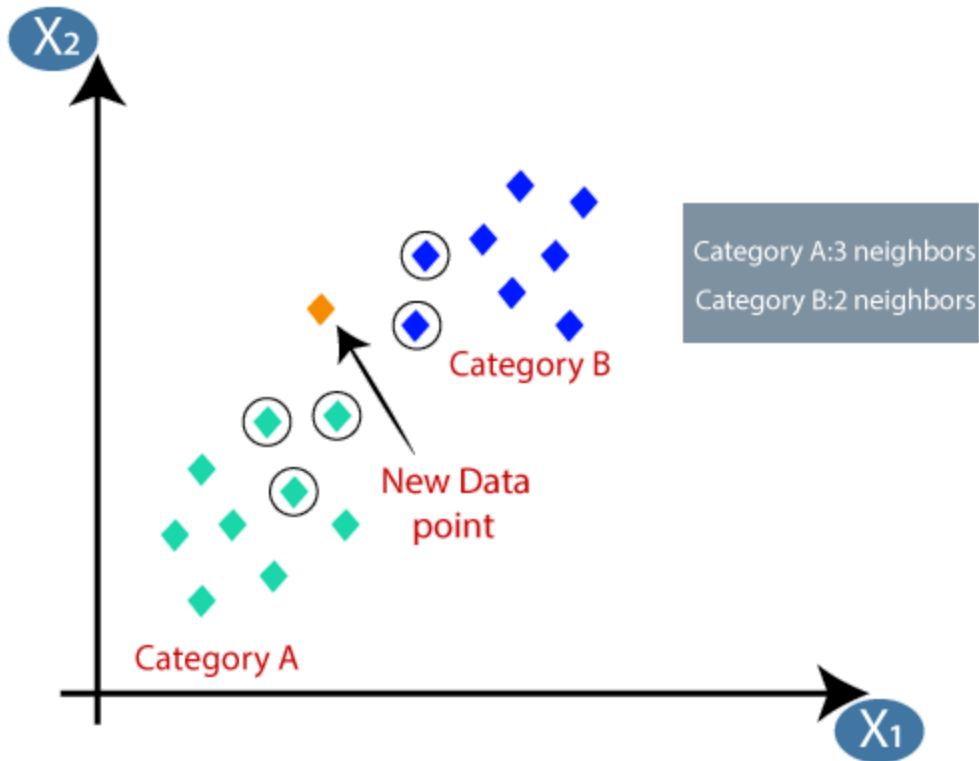
Suppose we have a new data point and we need to put it in the required category.
Consider the below image:



- o Firstly, we will choose the number of neighbors, so we will choose the $k=5$.
- o 10
- o Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



- o By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



- o As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

How to select the value of K in the K-NN Algorithm?

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- o There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- o A very low value for K such as $K=1$ or $K=2$, can be noisy and lead to the effects of outliers in the model.
- o Large values for K are good, but it may find some difficulties.

Advantages of KNN Algorithm:

- o It is simple to implement.
- o It is robust to the noisy training data
- o It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- o Always needs to determine the value of K which may be complex some time.
- o The computation cost is high because of calculating the distance between the data points for all the training samples.

Support Vector Machine Algorithm

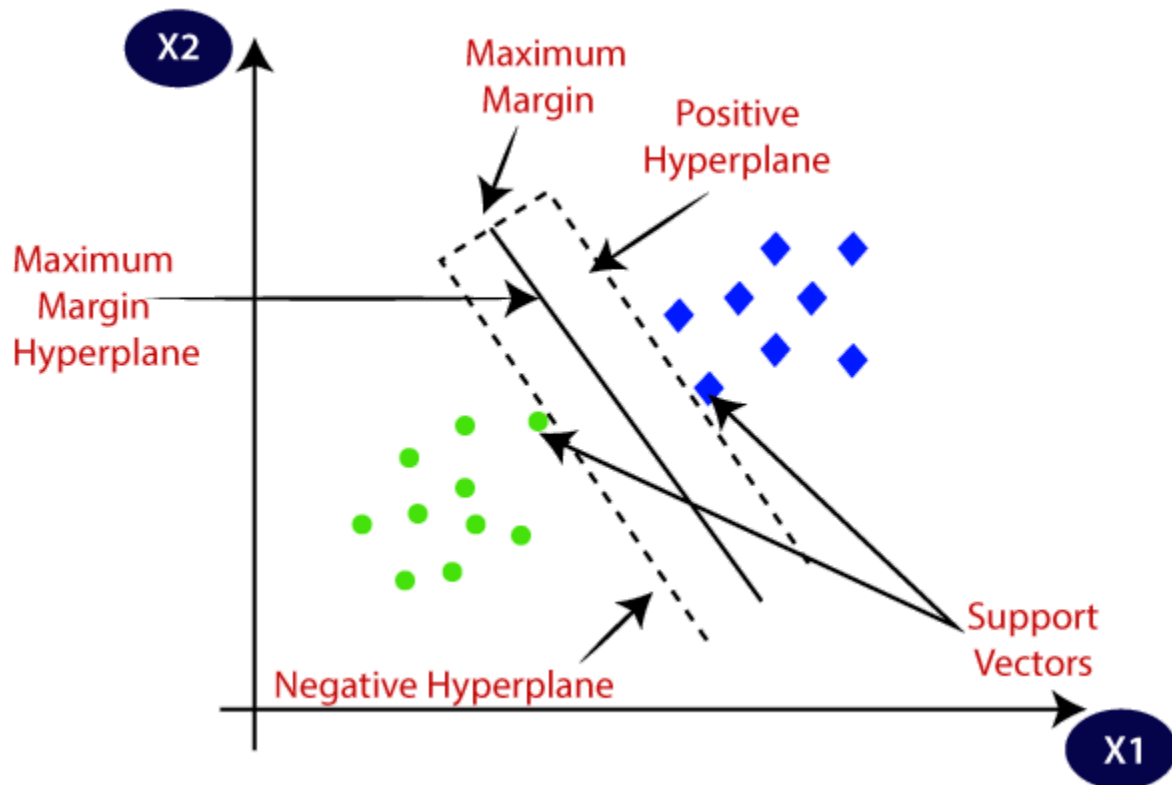
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed

as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

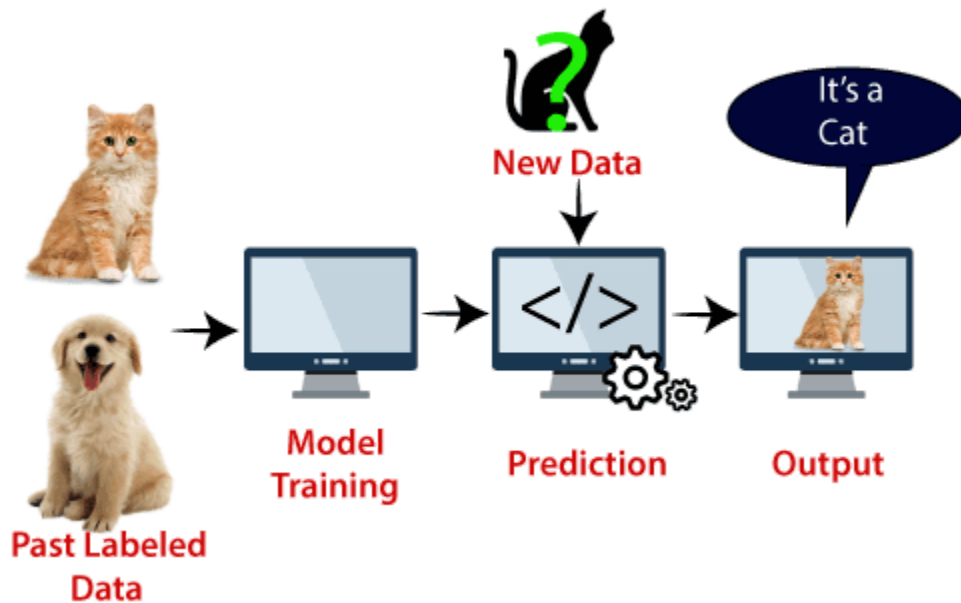
13



Example: SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and

dog. On the basis of the support vectors, it will classify it as a cat. Consider the below diagram:

14



SVM algorithm can be used for **Face detection, image classification, text categorization**, etc.

Types of SVM

SVM can be of two types:

- o **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- o **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

NAÏVE BAYES CLASSIFIER

The Naïve Bayes algorithm is compNaïve Bayes Classifier Algorithm

- o Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- o It is mainly used in *text classification* that includes a high-dimensional training dataset.
- o Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- o **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.**
- o Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles.**

Why is it called Naïve Bayes?

Rised of two words Naïve and Bayes, Which can be described as:

- o **Naïve:** It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

16

- o **Bayes:** It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem :

- o Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- o The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Working of Naïve Bayes' Classifier:

Working of Naïve Bayes' Classifier can be understood with the help of the below example:

Suppose we have a dataset of **weather conditions** and corresponding target

17

variable “**Play**”. So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:

1. Convert the given dataset into frequency tables.
2. Generate Likelihood table by finding the probabilities of given features.
3. Now, use Bayes theorem to calculate the posteri

CHAPTER 2

LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system. The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the

associated designing it is necessary to determine and survey the time factor, Once these things are satisfied and fully surveyed, then the next step is to determine

about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

An Enhanced Efficient Approach For Spam Detection In IOT Devices Using Machine Learning

19

The number of Internet of Things (IoT) devices is growing at a quick pace in smart homes, producing large amounts of knowledge, which are mostly transferred over wireless communication channels. The volume of data released from these devices also increased. In addition to an increased volume, the IoT device produces a large amount of data with several different modalities having varying data quality defined by its speed in terms of time and position dependency. However, various IoT devices are susceptible to different threats, like cyber-attacks, fluctuating network connections, leakage of data, etc. However, the unique characteristics of IoT nodes render the prevailing solutions insufficient to encompass the whole security spectrum of the IoT networks. In such an environment, machine learning algorithms can play an important role in detecting anomalies in the data, which enhances the security of IoT systems. Our methods target the data anomalies present in general smart Internet of Things (IoT) devices, allowing for easy detection of anomalous events based on stored data. The proposed algorithm is employed to detect the spamicity score of the connected IoT devices within the

network. The obtained results illustrate the efficiency of the proposed algorithm to analyze the time-series data from the IoT devices for spam detection.

Ensemble-Based Spam Detection in Smart Home IoT Devices Time Series Data Using Machine Learning Techniques

The number of Internet of Things (IoT) devices is growing at a fast pace in smart homes, producing large amounts of data, which are mostly transferred over wireless communication channels. However, various IoT devices are vulnerable to different threats, such as cyber-attacks, fluctuating network connections, leakage of

20

information, etc. Statistical analysis and machine learning can play a vital role in detecting the anomalies in the data, which enhances the security level of the smart home IoT system which is the goal of this paper. This paper investigates the trustworthiness of the IoT devices sending house appliances' readings, with the help of various parameters such as feature importance, root mean square error, hyper-parameter tuning, etc. A spamicity score was awarded to each of the IoT devices by the algorithm, based on the feature importance and the root mean square error score of the machine learning models to determine the trustworthiness of the device in the home network. A dataset publicly available for a smart home, along with weather conditions, is used for the methodology validation. The proposed algorithm is used to detect the spamicity score of the connected IoT devices in the network. The obtained results illustrate the efficacy of the proposed algorithm to analyze the time series data from the IoT devices for spam detection.

Using Machine Learning Unsolicited Information Detection Technique For Iot Devices

The unsolicited information detection technique is to forestall the phony or unapproved access into the framework. A bit of the current plans are utilized to see information in the messages, web pages, emails and some more. Be that as it may, the proposed plot is for IOT contraptions like sensors, actuators, clever house-hold machines, insightful vehicles, Augmented Reality i.e., the most recent version of Google Glasses which permits customers to transfer clear "perspective" record of different stream using wifi and other programming innovation which are associated over web or intranet for information transmission. The creation of a problematic

IOT will produce a large volume of information in different forms. A quality of these data will fluctuate depending on the time and location, which is represented by their speed. One can't depict IOT without Machine Learning(ML) considering the way that it has the greater part of the significant highlights like security, simple to utilize, reliable, as well as fit in making and utilizing a Smart gadget.

A model-based approach for identifying spammers in social networks

In this paper, we view the task of identifying spammers in social networks from a mixture modeling perspective, based on which we devise a principled unsupervised approach to detect spammers. In our approach, we first represent each user of the social network with a feature vector that reflects its behaviour and interactions with other participants. Next, based on the estimated users feature vectors, we propose a statistical framework that uses the Dirichlet distribution in order to identify spammers. The proposed approach is able to automatically discriminate between spammers and legitimate users, while existing unsupervised approaches require

human intervention in order to set informal threshold parameters to detect spammers. Furthermore, our approach is general in the sense that it can be applied to different online social sites. To demonstrate the suitability of the proposed method, we conducted experiments on real data extracted from Instagram and Twitter.

Spam detection of Twitter traffic: A framework based on random forests and non-uniform feature sampling

22

Law Enforcement Agencies cover a crucial role in the analysis of open data and need effective techniques to filter troublesome information. In a real scenario, Law Enforcement Agencies analyze Social Networks, i.e. Twitter, monitoring events and profiling accounts. Unfortunately, between the huge amount of internet users, there are people that use microblogs for harassing other people or spreading malicious contents. Users' classification and spammers' identification is a useful technique for relieving Twitter traffic from uninformative content. This work proposes a framework that exploits a non-uniform feature sampling inside a gray box Machine Learning System, using a variant of the Random Forests Algorithm to identify spammers inside Twitter traffic. Experiments are made on a popular Twitter dataset and on a new dataset of Twitter users. The new provided Twitter dataset is made up of users labeled as spammers or legitimate users, described by 54 features. Experimental results demonstrate the effectiveness of enriched feature sampling method.

CHAPTER 3

SYSTEM REQUIREMENTS

3.1 HARDWARE REQUIREMENTS

System : Pentium i3 Processor

Hard Disk : 500 GB.

Monitor : 15” LED

Input Devices : Keyboard, Mouse

Ram : 2 GB

3.2 SOFTWARE REQUIREMENTS

Operating system : Windows 10

Coding Language : Python

3.3 LANGUAGE SPECIFICATION

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). This **tutorial** gives enough understanding on **Python programming** language.

24

3.4. HISTORY OF PYTHON

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

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, 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 Rossum still holds a vital role in directing its progress.

3.5. APPLICATION OF PYTHON

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-maintain.

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.

25

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.

3.6 FEATURES OF PYTHON

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.

3.7 FEASIBILITY STUDY

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

26

company. For feasibility analysis, some understanding of the major requirements for the system is essential.

The feasibility study investigates the problem and the information needs of the stakeholders. It seeks to determine the resources required to provide an information systems solution, the cost and benefits of such a solution, and the feasibility of such a solution.

The goal of the feasibility study is to consider alternative information systems solutions, evaluate their feasibility, and propose the alternative most suitable to the organization. The feasibility of a proposed solution is evaluated in terms of its components.

3.7.1 ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system is 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.

3.7.2 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

27

on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

3.7.3 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.

CHAPTER 4

SYSTEM ANALYSIS

4.1 PURPOSE

The purpose of this document is an efficient spam detection technique for Iot devices using machine learning algorithms. In detail, this document will provide a general description of our project, including user requirements, product perspective, and overview of requirements, general constraints. In addition, it will also provide the specific requirements and functionality needed for this project - such as interface, functional requirements and performance requirements.

4.2 SCOPE

The scope of this SRSdocument persists for the entire life cycle of the project. This document defines the final state of the software requirements agreed upon by the customers and designers. Finally at the end of the project execution all the functionalities may be traceable from the SRSto the product. The document describes the functionality, performance, constraints, interface and reliability for the entire life cycle of the project.

4.3 EXISTING SYSTEM

To encompass the existing state-ofthe-art, a few surveys have also been carried out on fake user identification from Twitter. Tingmin et al provide a survey of new met methods and techniques to identify Twitter spam detection. The above surge

presents a comparative study of the current approaches. On the other hand, the authors conducted a survey on different behaviors exhibited by spammers on Twitter social network. The study also provides a literature review that recognizes the existence of spammers on Twitter social network. Despite all the existing studies, there is still a gap in the existing literature. Therefore, to bridge the gap, we review state-of-the-art in the spammer detection and fake user identification on Twitter.

DISADVANTAGES EXISTING SYSTEM

- ❖ No efficient methods used.
- ❖ No real time data's used.
- ❖ More complex

4.4 PROPOSED SYSTEM

The proposed approach detects the spam parameters causing the IoT devices to be effected. To get the best results, the IoT dataset is used for the validation of proposed approach as described in the next Section. The proposed framework, detects the spam parameters using machine learning models. The IoT dataset used for experiments, is pre-processed by using feature engineering procedure. By experimenting the framework with machine learning models, each appliance is

30

awarded with a spam score. This refines the conditions to be taken for successful working of devices in a smart home.

ADVANTAGES OF PROPOSED SYSTEM

- ❖ This study includes machine learning methodology proposed using real time datasets and with different characteristics and accomplishments.
- ❖ The proposed system is more effective and accurate than other existing systems.
- ❖ Tested with real time data's.

CHAPTER 5

SYSTEM DESIGN

5.1 INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system.

The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

5.2 OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the

32

user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

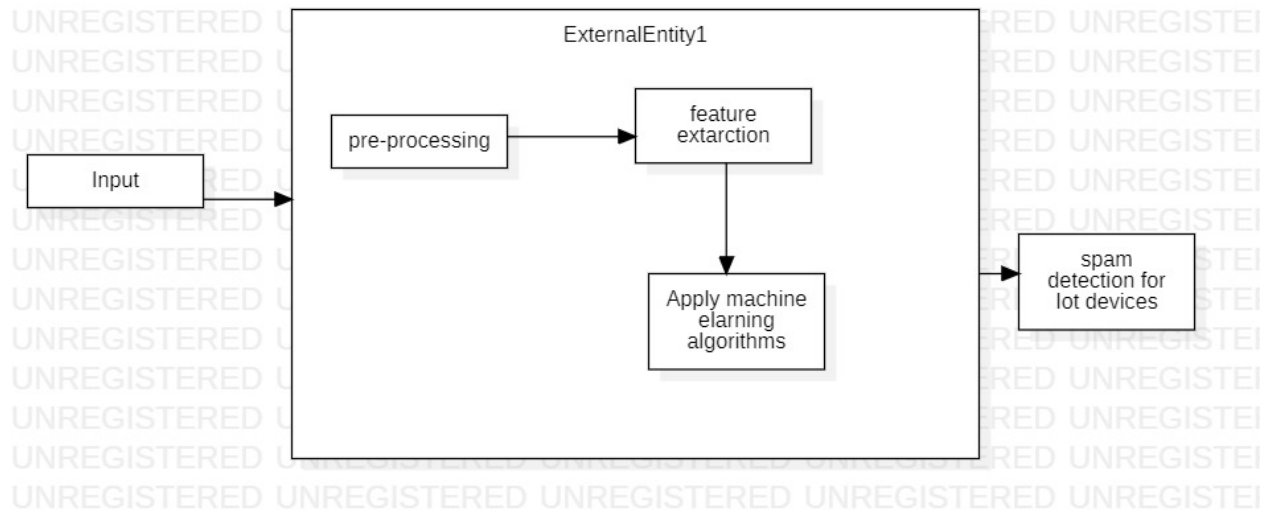
The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action

5.3 DATA FLOW DIAGRAM

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



UML DIAGRAMS

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

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

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

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

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

GOALS:

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

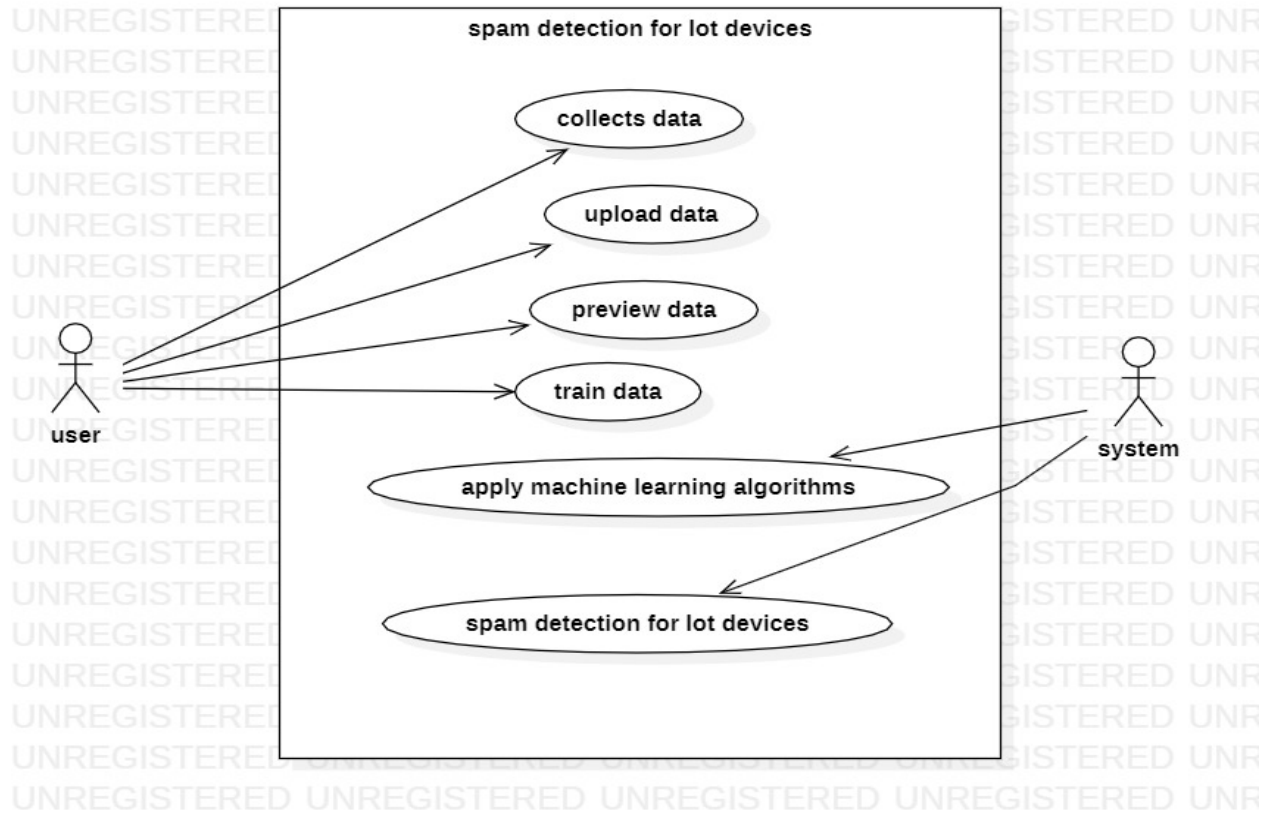
1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a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

USE CASE DIAGRAM:

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

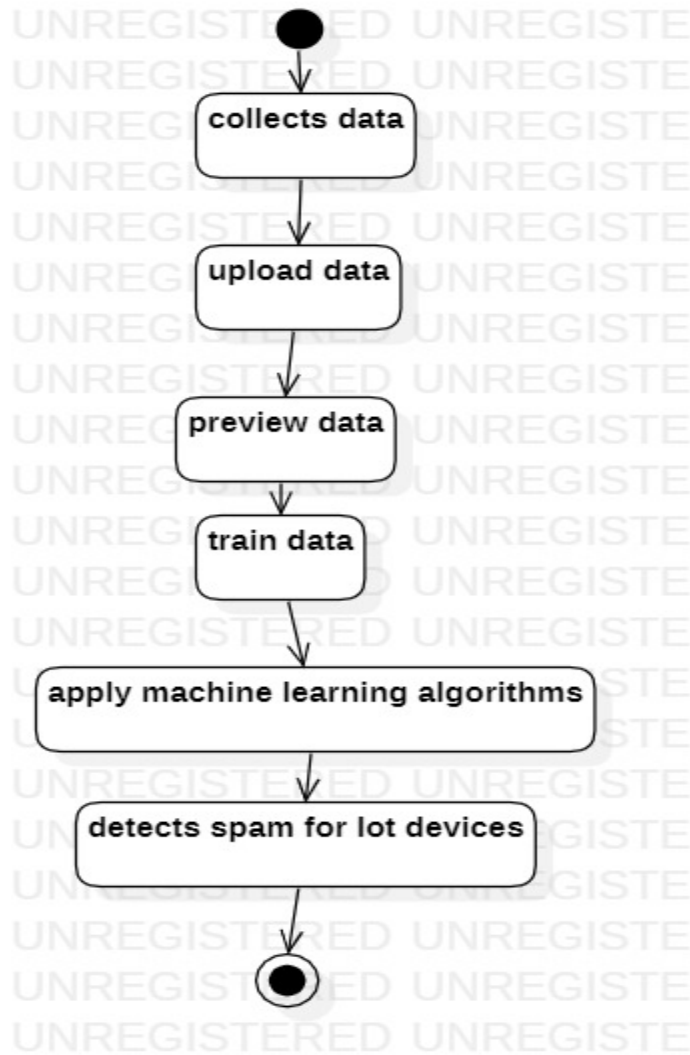
35

functions are performed for which actor. Roles of the actors in the system can be depicted.



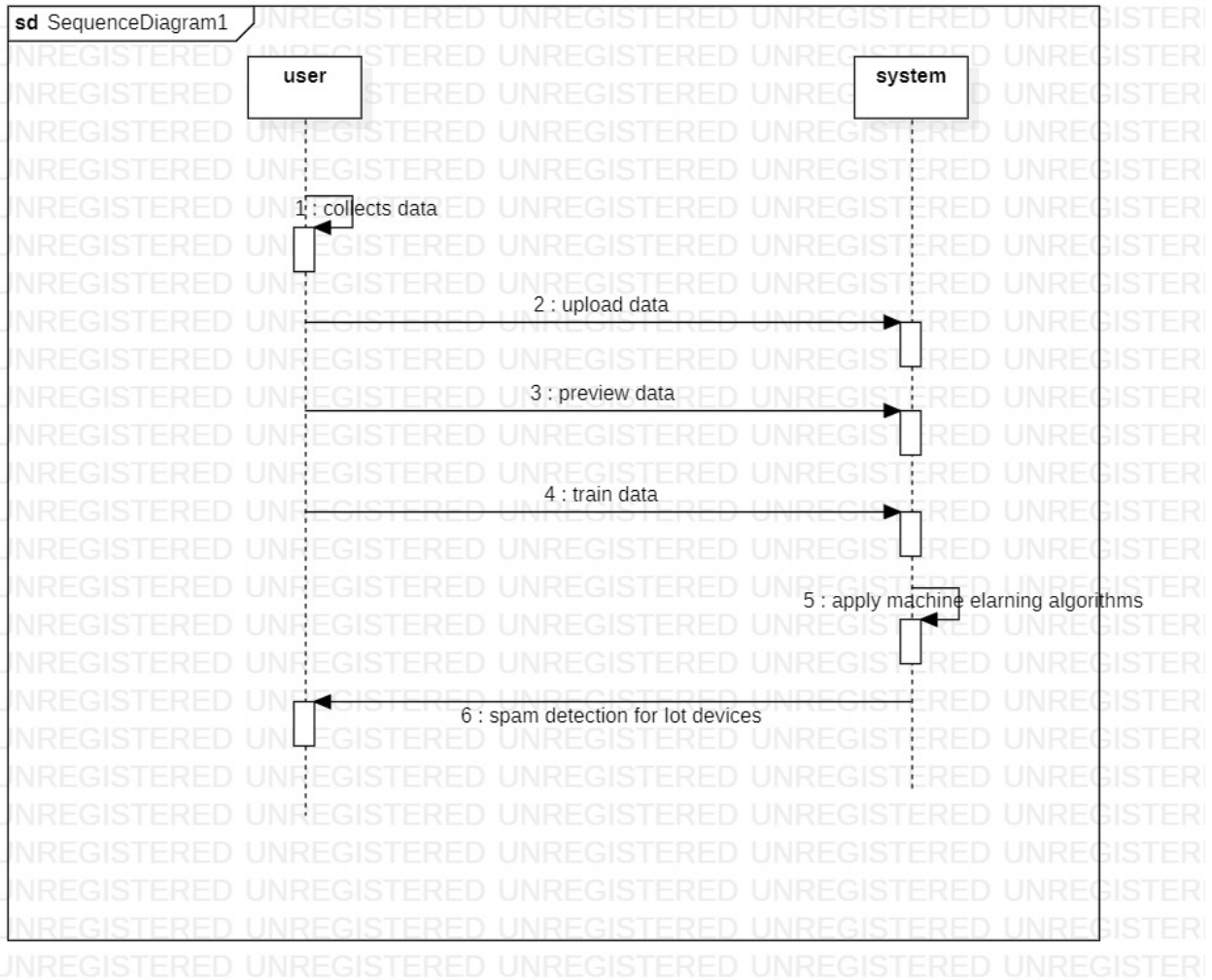
ACTIVITY DIAGRAM:

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



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.



CHAPTER 6

MODULES

MODULES

Login Module

Data Collection Module

Pre-Processing Module

Train and Test

Detection of Spam

MODULE DESCRIPTION

6.1 Login Module

In the first module, we develop the spam detecting technique for smart home system module. We build up the system with the feature of spam detecting technique for smart home System. Where, this module is used for admin login with their authentication.

6.2 Data Collection Module

We have collected the smart home dataset by REFIT. A total of twenty homes were used and advised to deploy the smart home technologies. The complete survey was conducted by the team of researchers. The experiments are varied from room to room, depending upon climate changes, floor plans, Internet supply and

other attributes. The internal environmental conditions were captured using different sensors. There were more than 100,000 data points in each home for sensor monitoring. The survey was continued for almost 18 months.

6.3 Pre-Processing Module

The preprocessing involves the selection of appliances being considered for the detection of spam parameters. The main idea is to find the various spam causing factors. Firstly, the feature reduction is done. The method used for feature reduction is the Principal Component Analysis (PCA), which reduces the dimensions of data. It results in series of Principal components (PC) which corresponds to each row with each column. In the IoT dataset used in this proposal, we have 22 features, so 22 PCs are generated such as Generation resource, Dishwasher, Home Office, Wine Cellar, Kitchen, Well, Living Room, Temperature, Visibility, Pressure, WindBearing, House Overall, Furnace, Fridge, Garage Door, Barn, Microwave, Solar, Humidity, Apparent Temperature, Wind speed, Precipintensity. The `pca()` works in such a way that it reduces the variance among the features.

6.4 Train and Test Module

We present the proposed framework for metadata features are extracted from available additional information regarding the home appliances, whereas content-based features aim to observe the components of a smart home and the quality of

the home appliances

6.5 Detection of Spam

The proposed framework, detects the spam parameters of IoT devices using machine learning models. The IoT dataset used for experiments, is pre-processed by using feature engineering procedure. By experimenting the framework with machine learning models, each IoT appliance is awarded with a spam score. This refines the conditions to be taken for successful working of IoT devices in a smart home.

CHAPTER 7

SYSTEM IMPLEMENTATION

7.1 SYSTEM ARCHITECTURE

Describing the overall features of the software is concerned with defining the requirements and establishing the high level of the system. During architectural design, the various web pages and their interconnections are identified and designed. The major software components are identified and decomposed into processing modules and conceptual data structures and the interconnections among the modules are identified. The following modules are identified in the proposed system.

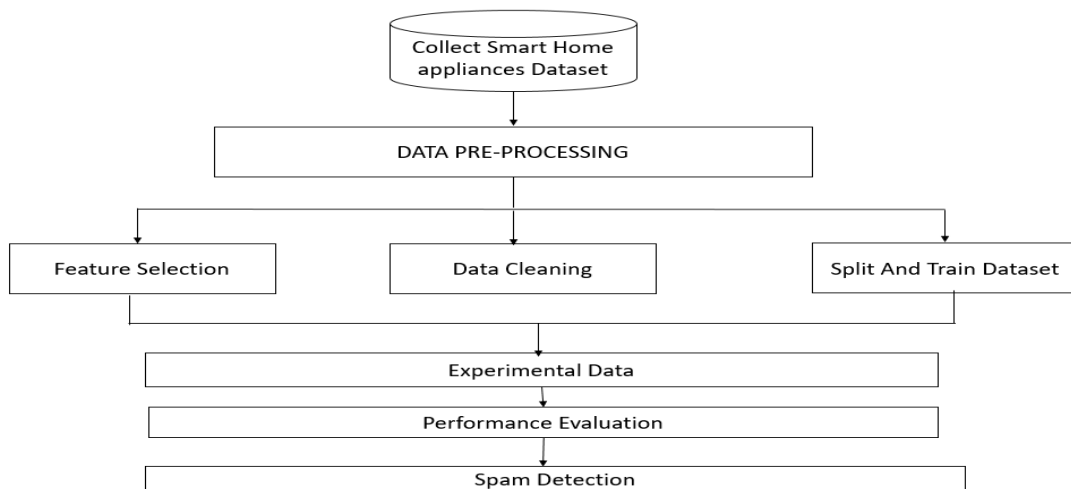


FIG:7.1 SYSTEM ARCHITECTURE

CHAPTER 8

SYSTEM TESTING

8.1 Test plan

Software testing is the process of evaluation a software item to detect differences between given input and expected output. Also to assess the feature of a software item. Testing assesses the quality of the product. Software testing is a process that should be done during the development process. In other words software testing is a verification and validation process.

8.2 Verification

Verification is the process to make sure the product satisfies the conditions imposed at the start of the development phase. In other words, to make sure the product behaves the way we want it to.

8.3 Validation

Validation is the process to make sure the product satisfies the specified requirements at the end of the development phase. In other words, to make sure the product is built as per customer requirements.

8.4 Basics of software testing

There are two basics of software testing: black box testing and white box testing.

8.5 Black box Testing

Black box testing is a testing technique that ignores the internal mechanism of the system and focuses on the output generated against any input and execution of the system. It is also called functional testing.

8.6 White box Testing

White box testing is a testing technique that takes into account the internal mechanism of a system. It is also called structural testing and glass box testing. Black box testing is often used for validation and white box testing is often used for verification.

8.7 Types of testing

There are many types of testing like

- Unit Testing
- Integration Testing
- Functional Testing
- System Testing
- Stress Testing
- Performance Testing
- Usability Testing
- Acceptance Testing
- Regression Testing
- Beta Testing

8.7.1 Unit Testing

Unit testing is the testing of an individual unit or group of related units. It falls under the class of white box testing. It is often done by the programmer to test that the unit he/she has implemented is producing expected output against given input.

8.7.2 Integration Testing

Integration testing is testing in which a group of components are combined to produce output. Also, the interaction between software and hardware is tested in integration testing if software and hardware components have any relation. It may fall under both white box testing and black box testing.

8.7.3 Functional Testing

Functional testing is the testing to ensure that the specified functionality required in the system requirements works. It falls under the class of black box testing.

8.7.4 System Testing

System testing is the testing to ensure that by putting the software in different environments (e.g., Operating Systems) it still works. System testing is done with full system implementation and environment. It falls under the class of black box testing.

8.7.5 Stress Testing

Stress testing is the testing to evaluate how system behaves under unfavorable conditions. Testing is conducted at beyond limits of the specifications. It falls under the class of black box testing.

8.7.6 Performance Testing

Performance testing is the testing to assess the speed and effectiveness of the system and to make sure it is generating results within a specified time as in performance requirements. It falls under the class of black box testing.

8.7.7 Usability Testing

Usability testing is performed to the perspective of the client, to evaluate how the GUI is user-friendly? How easily can the client learn? After learning how to use, how proficiently can the client perform? How pleasing is it to use its design? This falls under the class of black box testing.

8.7.8 Acceptance Testing

Acceptance testing is often done by the customer to ensure that the delivered product meets the requirements and works as the customer expected. It falls under

the class of black box testing.

8.7.9 Regression Testing

Regression testing is the testing after modification of a system, component, or a group of related units to ensure that the modification is working correctly and is not damaging or imposing other modules to produce unexpected results. It falls under the class of black box testing

REQUIREMENT ANALYSIS

Requirement analysis, also called requirement engineering, is the process of determining user expectations for a new modified product. It encompasses the tasks that determine the need for analyzing, documenting, validating and managing software or system requirements. The requirements should be documentable, actionable, measurable, testable and traceable related to identified business needs or opportunities and define to a level of detail, sufficient for system design.

FUNCTIONAL REQUIREMENTS

It is a technical specification requirement for the software products. It is the first step in the requirement analysis process which lists the requirements of particular software systems including functional, performance and security requirements. The function of the system depends mainly on the quality hardware used to run the software with given functionality.

Usability

It specifies how easy the system must be use. It is easy to ask queries in any format which is short or long, porter stemming algorithm stimulates the desired response for user.

Robustness

It refers to a program that performs well not only under ordinary conditions but also under unusual conditions. It is the ability of the user to cope with errors for irrelevant queries during execution.

Security

The state of providing protected access to resource is security. The system provides good security and unauthorized users cannot access the system there by providing high security.

Reliability

It is the probability of how often the software fails. The measurement is often expressed in MTBF (Mean Time Between Failures). The requirement is needed in order to ensure that the processes work correctly and completely without being aborted. It can handle any load and survive and survive and even capable of working around any failure.

Compatibility

It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience.

Flexibility

The flexibility of the project is provided in such a way that it has the ability to run on different environments being executed by different users.

Safety

Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one's personal information.

NON- FUNCTIONAL REQUIREMENTS**Portability**

It is the usability of the same software in different environments. The project can be run in any operating system.

Performance

These requirements determine the resources required, time interval, throughput and everything that deals with the performance of the system.

Accuracy

The result of the requesting query is very accurate and high speed of retrieving information. The degree of security provided by the system is high and effective.

Maintainability

Project is simple as further updates can be easily done without affecting its stability. Maintainability basically defines that how easy it is to maintain the system. It means that how easy it is to maintain the system, analyse, change and test the application. Maintainability of this project is simple as further updates can be easily done without affecting its stability.

CHAPTER 9

SCREENSHOTS

Home Page

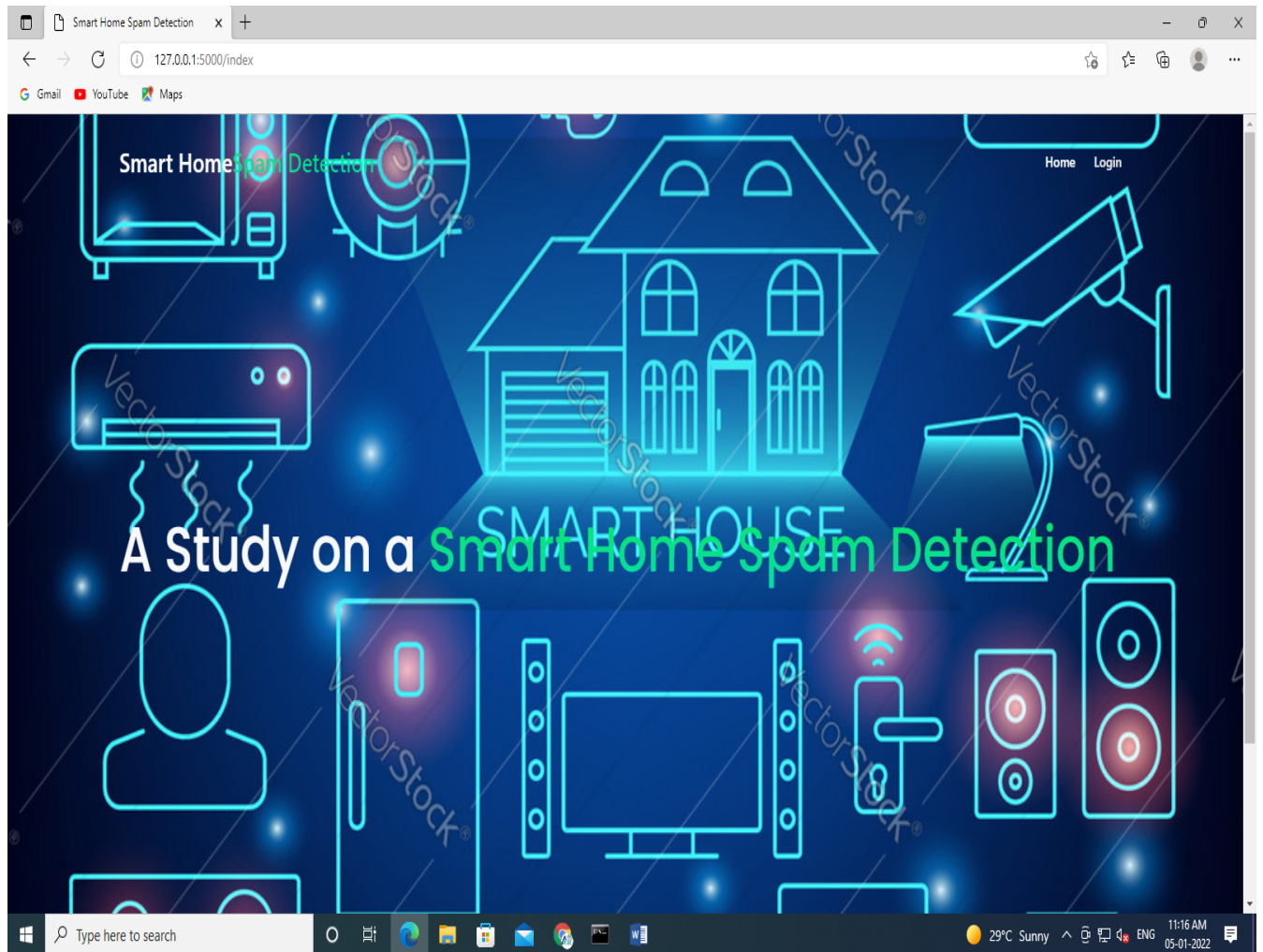


FIG: 9.1 HOME PAGE

This is the home page of our system in order to detect techniques for Iot devices using machine learning algorithms. It contains Login page and home page. Login page is to get into the system.

Login Page

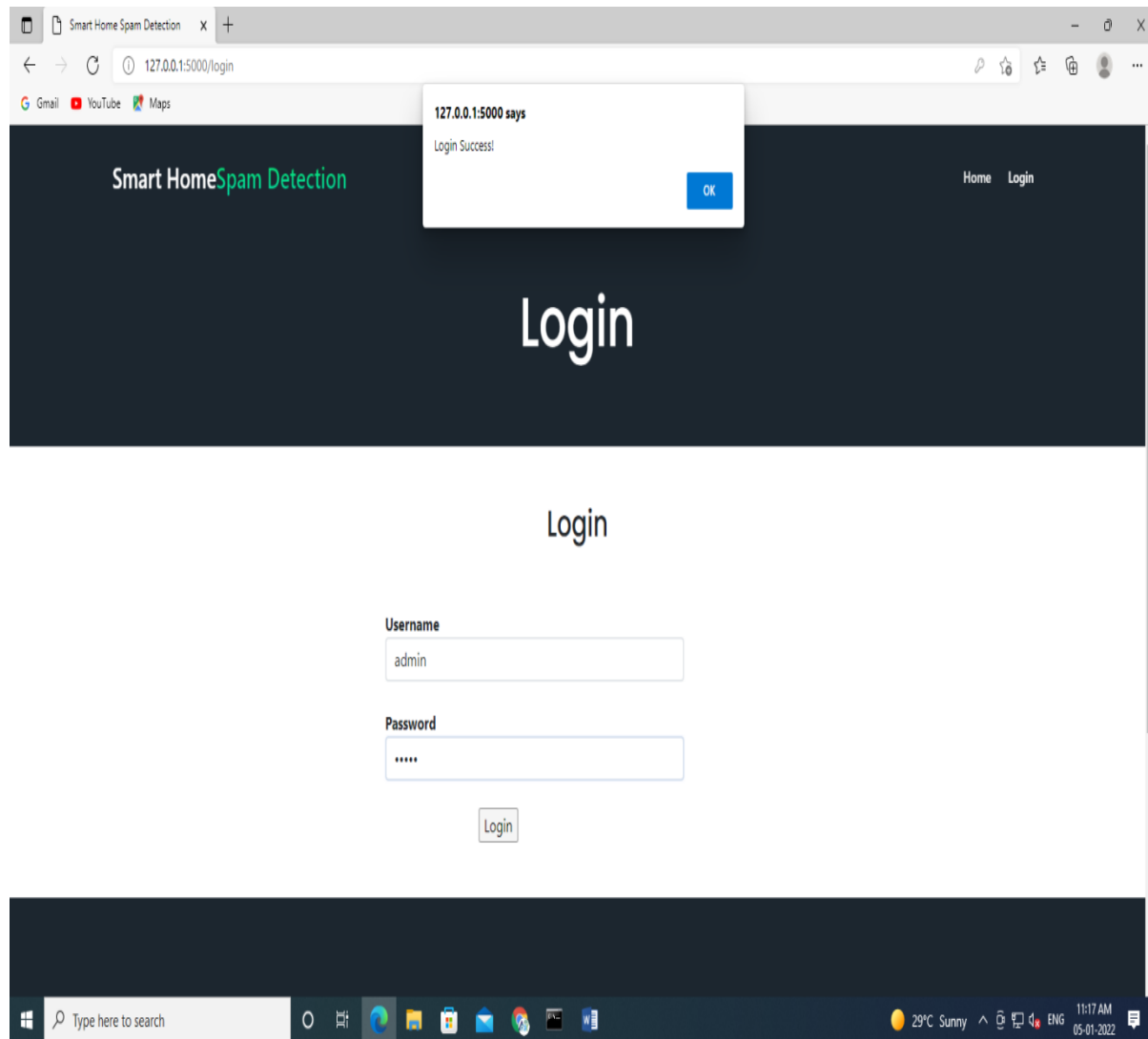
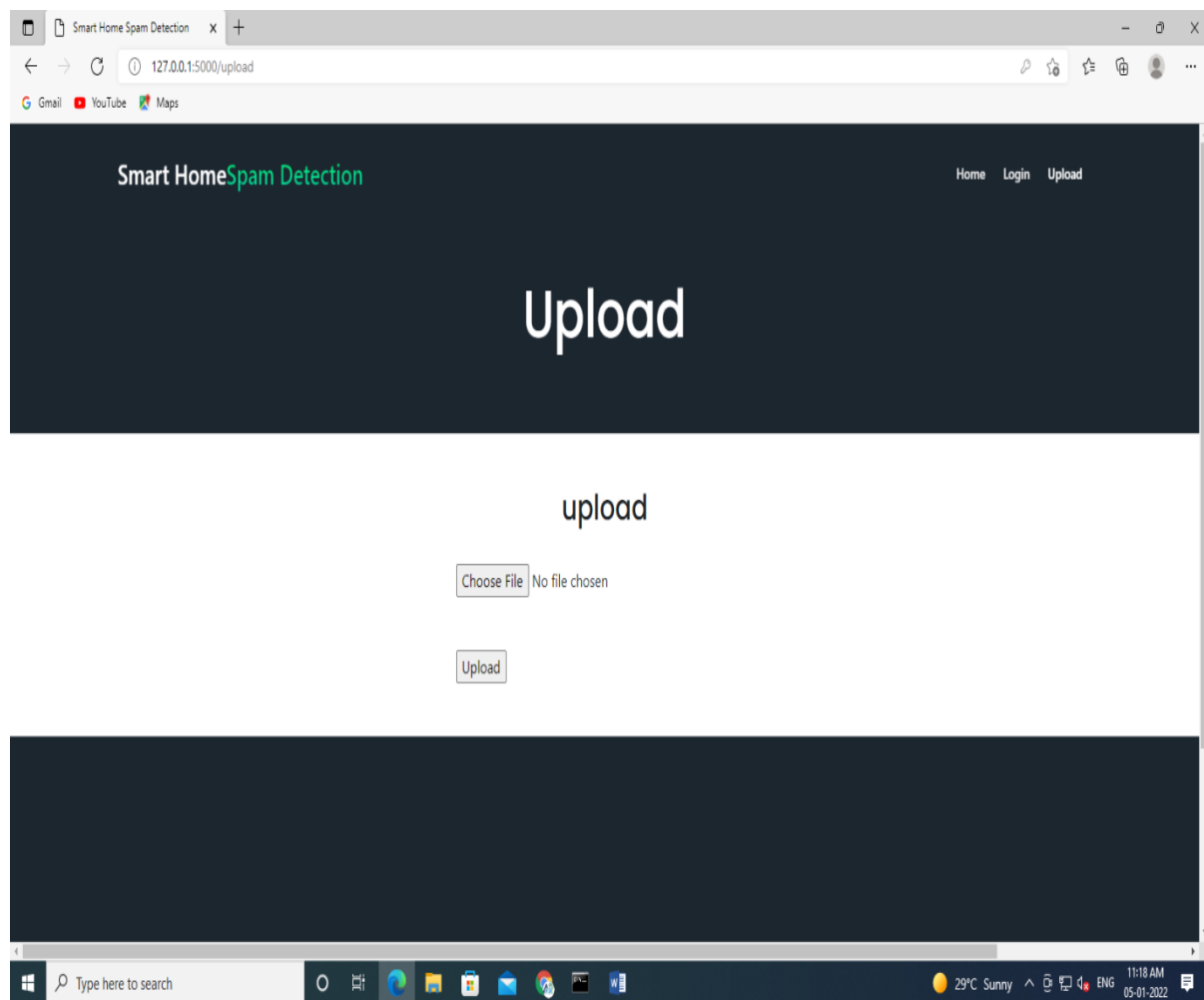


FIG: 9.2 LOGIN PAGE

This is the login page in which we will be able to login to the system by providing the required details to get into the system. The username is admin and password is

the required details to get into the system. The username is admin and password admin then if we click on login we will be able to enter to the system. If we provide invalid credentials then it will show the error that to invalid credentials.

Dataset Upload Page



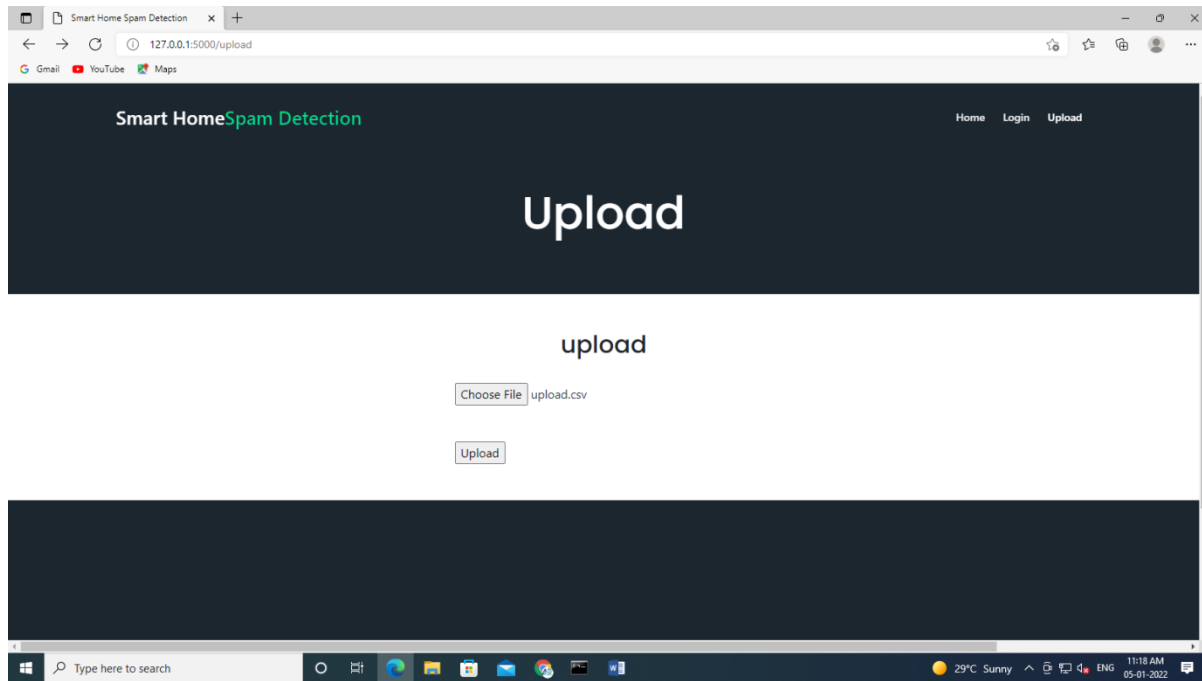


FIG: 9.3 UPLOAD PAGE

The Upload page contains of some details to upload the dataset. It contains a button called choose file if we click on then we will be able to open the file from the folder where dataset is present and then we will be able to upload the dataset to the system.

Preview Page

Smart Home Spam Detection

localhost:5000/preview

Home

Login

Upload

preview

★

	time	use [kW]	gen [kW]	House overall [kW]	Dishwasher [kW]	Furnace 1 [kW]	Furnace 2 [kW]	Home office [kW]	Fridge [kW]	Wine cellar [kW]	Garage door [kW]	Kitchen 12 [kW]	Kitchen 14 [kW]	Kitchen 38 [kW]	Barn [kW]	Well [kW]	Microwave [kW]	Living room [kW]	Sc [kW]
1	1451624400	0.932833	0.003483	0.932833	0.000033	0.020700	0.061917	0.442633	0.124150	0.006983	0.013083	0.000417	0.000150	0.000000	0.031350	0.001017	0.004067	0.001517	0.003...
2	1451624401	0.934333	0.003467	0.934333	0.000000	0.020717	0.063817	0.444067	0.124000	0.006983	0.013117	0.000417	0.000150	0.000000	0.031500	0.001017	0.004067	0.001650	0.003...
3	1451624402	0.931817	0.003467	0.931817	0.000017	0.020700	0.062317	0.446067	0.123533	0.006983	0.013083	0.000433	0.000167	0.000017	0.031517	0.001000	0.004067	0.001650	0.003...
4	1451624403	1.022050	0.003483	1.022050	0.000017	0.106900	0.068517	0.446583	0.123133	0.006983	0.013000	0.000433	0.000217	0.000000	0.031500	0.001017	0.004067	0.001617	0.003...
5	1451624404	1.139400	0.003467	1.139400	0.000133	0.236933	0.063983	0.446533	0.122850	0.006850	0.012783	0.000450	0.000333	0.000000	0.031500	0.001017	0.004067	0.001583	0.003...
6	1451624405	1.391867	0.003433	1.391867	0.000283	0.503250	0.063667	0.447033	0.122300	0.006717	0.012433	0.000483	0.000567	0.000000	0.031450	0.001017	0.004067	0.001583	0.003...

Type here to search

<

FIG: 9.4 PREVIEW PAGE

In preview page we will be able to see the dataset which we uploaded the dataset in upload page. In preview we can check the dataset that how much data and the dataset which we uploaded is correct and the fields also we can check. At the end

we can see the button called train.

Trained data

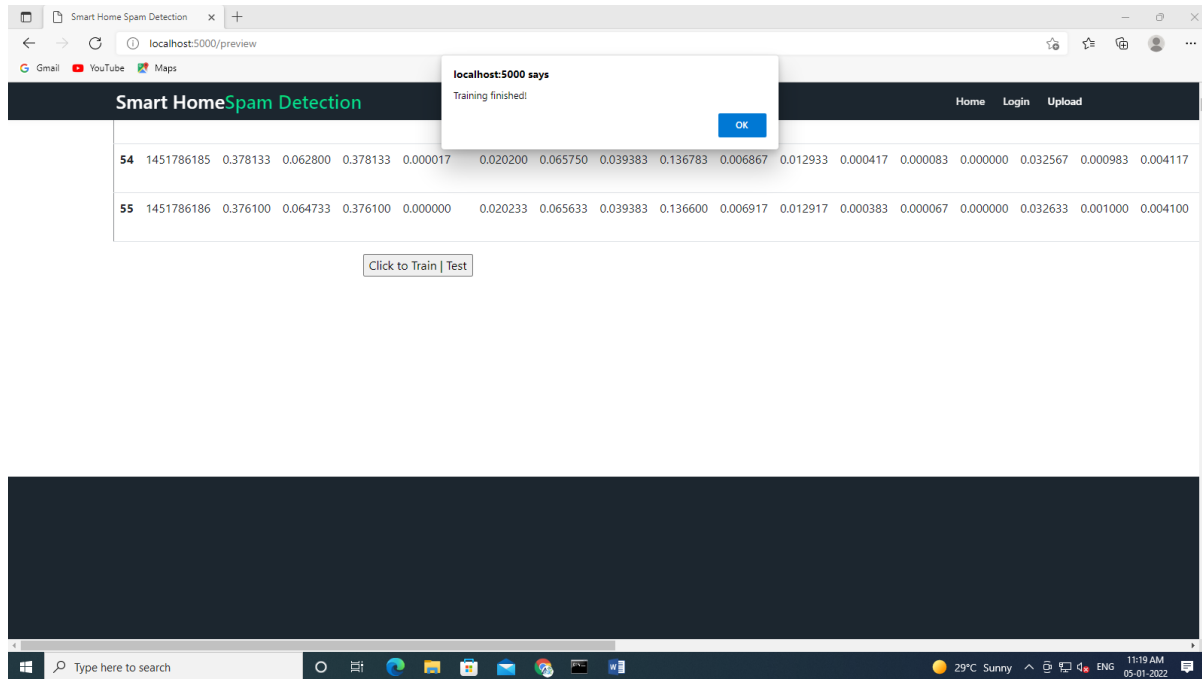


FIG: 9.5 TRAINED PAGE

In this page we will be able to train the dataset which is uploaded. Here we click on train/test then the dataset which is uploaded is trained. After completion of training the data we will get a pop window showing training finished.

Prediction Page for No Spam

Smart Home Spam Detection

Enter the details

Generation resorces[KW]: 0.003483333

Dishwasher[KW]: 3.33E-05

Home office[KW]: 0.0394

Kitchen[KW]: 0.000416667

temperature: 36.14

House overall[kW]: 0.571883333

Microwave[kW]: 0.004083333

precipintensity: 0

Predict

Prediction is :

Smart Home Spam Detection

Enter the details

Generation resorces[KW]: Generation

Dishwasher[KW]: Dishwasher

Home office[KW]: Home office

Kitchen[KW]: Kitchen

temperature: temperature

House overall[kW]: House overall

Microwave[kW]: Microwave

precipintensity: precipintensity

Predict

Prediction is : No Spam

FIG: 9.6 PREDICTION NO SPAM

In this prediction page we have to give sufficient details like generation resource, dishwasher, Home office, kitchen, temperature, microwave, intensity in order to predict spam for Iot devices is present or not.

Prediction for Spam

Smart HomeSpam Detection

Home Login Upload Prediction Performance

Smart Home Spam Detection

Enter the details

Generation resources[KW]: 0.057633333

Dishwasher[KW]: 1.67E-05

Home office[KW]: 0.932833333

Kitchen[KW]: 0.00048

temperature: 47.34

House overall[kW]: 0.381466667

Microwave[kW]: 0.004083333

precipintensity: 0.00028

Predict

Prediction is :

Smart HomeSpam Detection

Home Login Upload Prediction Performance

Smart Home Spam Detection

Enter the details

Generation resources[KW]: Generation

Dishwasher[KW]: Dishwasher

Home office[KW]: Home office

Kitchen[KW]: kitchen

temperature: temperature

House overall[kW]: House overall

Microwave[kW]: Microwave

precipintensity: precipintensity

Predict

Prediction is : Spam

FIG: 9.7 PREDICTION SPAM

Performance Analysis

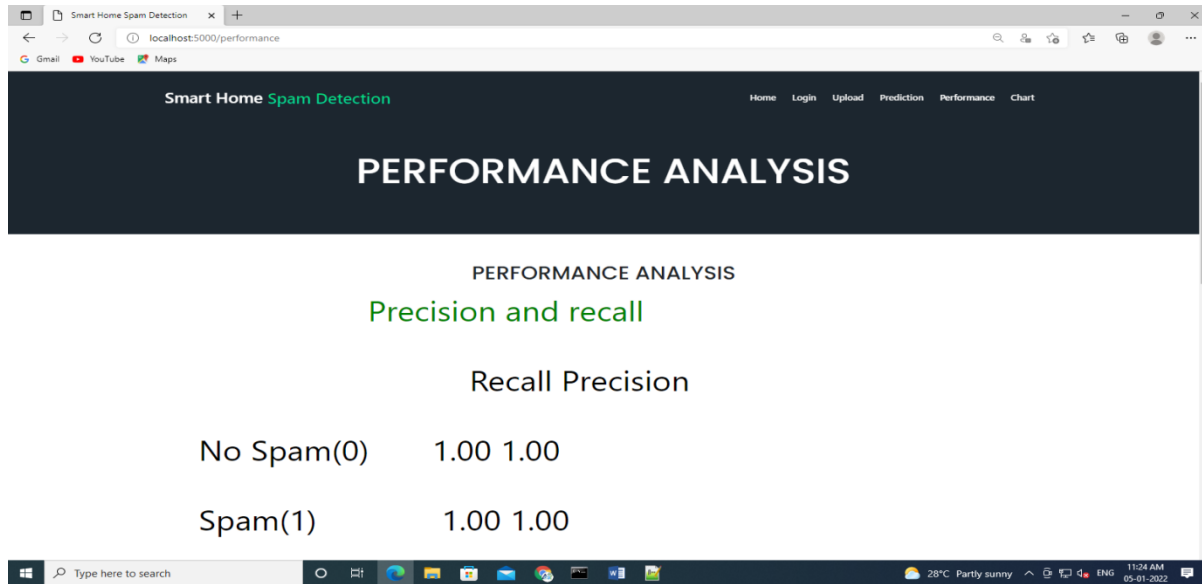


FIG: 9.8 PERFORMANCE ANALYSIS

In performance analysis here we will be giving analysis for spam and no spam along with recall value and precision value.

Confusion Matrix

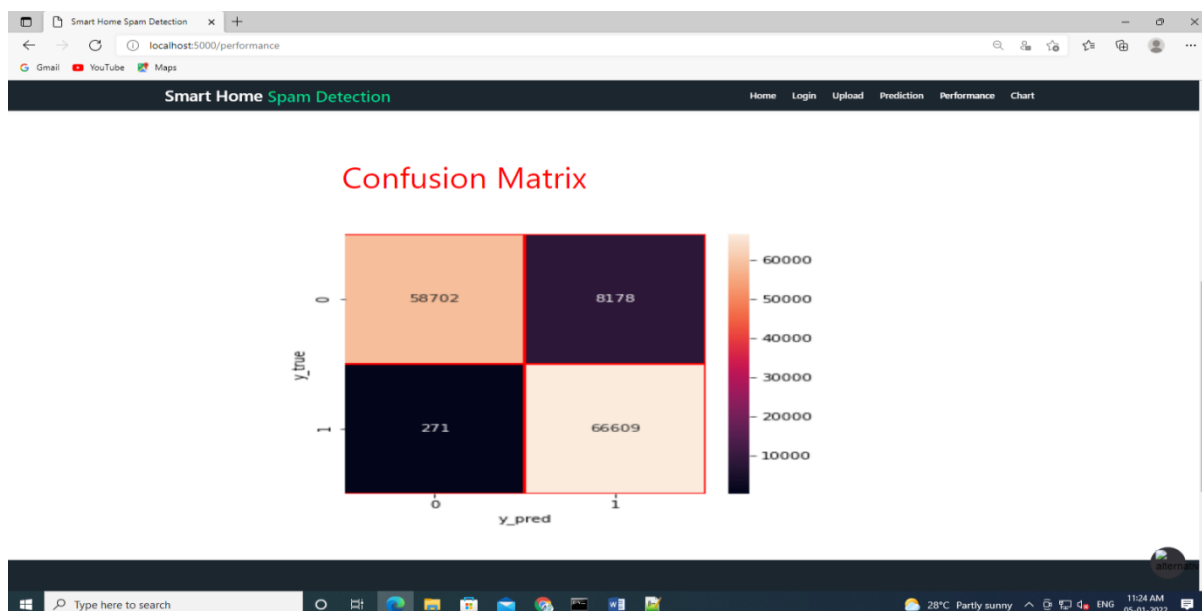


FIG: 9.9 CONFUSION MATRIX

59

In this page we will show the confusion matrix for the spam detection system for Iot devices using machine learning algorithms.

Chart Analysis

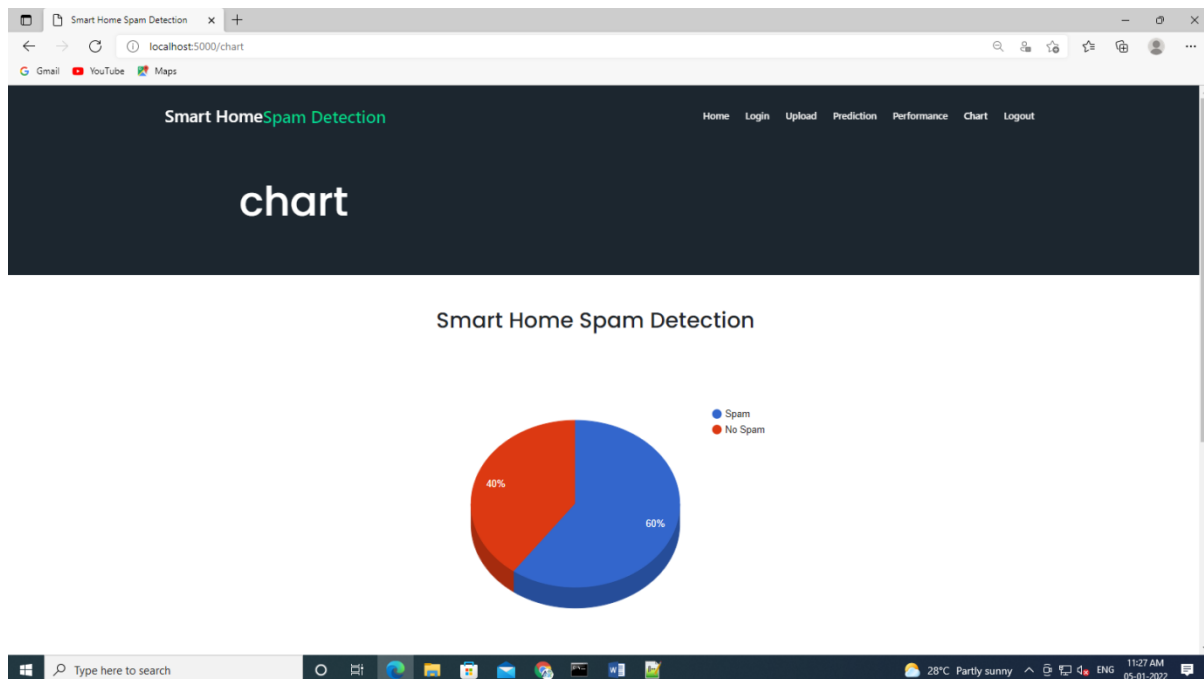


FIG: 9.10 CHART ANALYSIS

In this we will show the graph analysis by using chart. In this pie-chart we will be able to show the percentage of spam detection for the Iot devices using machine learning algorithms.

CHAPTER 10

CONCLUSION

In this paper, we have discussed that how our system detects techniques for Iot devices using machine learning algorithms. The proposed system is also scalable for detecting techniques for Iot devices by using techniques after collecting data. The system is not having complex process to detect techniques for Iot devices that the data like the existing system. Proposed system gives genuine and fast result than existing system. Here in this system we use machine learning algorithms to detects techniques for Iot devices.

CHAPTER 11

REFERENCES

- [1] C. Chen, S. Wen, J. Zhang, Y. Xiang, J. Oliver, A. Alelaiwi, and M. M. Hassan, “Investigating the deceptive information in Twitter spam,” *Future Gener. Comput. Syst.*, vol. 72, pp. 319–326, Jul. 2017.
- [2] I. David, O. S. Siordia, and D. Moctezuma, “Features combination for the detection of malicious Twitter accounts,” in *Proc. IEEE Int. Autumn Meeting Power, Electron. Comput. (ROPEC)*, Nov. 2016, pp. 1–6.
- [3] M. Babcock, R. A. V. Cox, and S. Kumar, “Diffusion of pro- and anti-false information tweets: The black panther movie case,” *Comput. Math. Org. Theory*, vol. 25, no. 1, pp. 72–84, Mar. 2019.
- [4] S. Keretna, A. Hossny, and D. Creighton, “Recognising user identity in Twitter social networks via text mining,” in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 3079–3082.
- [5] C. Meda, F. Bisio, P. Gastaldo, and R. Zunino, “A machine learning approach for Twitter spammers detection,” in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2014, pp. 1–6.
- [6] W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, “Real-time Twitter content polluter detection based on direct features,” in *Proc. 2nd Int. Conf. Inf. Sci. Secur. (ICISS)*, Dec. 2015, pp. 1–4.
- [7] H. Shen and X. Liu, “Detecting spammers on Twitter based on content and social interaction,” in *Proc. Int. Conf. Netw. Inf. Syst. Comput.*, pp. 413–417, Jan. 2015.
- [8] G. Jain, M. Sharma, and B. Agarwal, “Spam detection in social media using convolutional and long short term memory neural network,” *Ann. Math. Artif. Intell.*, vol. 85, no. 1, pp. 21–44, Jan. 2019.

[9] M. Washha, A. Qaroush, M. Mezghani, and F. Sedes, “A topic-based hidden Markov model for real-time spam tweets filtering,” *Procedia Comput. Sci.*, vol. 112, pp. 833–843, Jan. 2017.

62

[10] F. Pierri and S. Ceri, “False news on social media: A data-driven survey,” 2019, arXiv:1902.07539. [Online]. Available: <https://arxiv.org/abs/1902.07539>

[11] S. Sadiq, Y. Yan, A. Taylor, M.-L. Shyu, S.-C. Chen, and D. Feaster, “AAFA: Associative affinity factor analysis for bot detection and stance classification in Twitter,” in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2017, pp. 356–365.

[12] M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya, “Segregating spammers and unsolicited bloggers from genuine experts on Twitter,” *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 551–560, Jul./Aug. 2018.

