**Fake Job Post Prediction Using Machine Learning Algorithms**

*A Mini Project Report submitted to JNTU Hyderabad in partial fulfilment of the requirements for the award of the degree*

# BACHELOR OF TECHNOLOGY

In

# COMPUTER SCIENCE AND ENGINEERING (CYBER SECURITY)

***Submitted by***

**M. ABHINAI 21S11A6201**

**T. SRIKANTH 21S11A6256**

**J. SANTOSH GOUD 21S11A6250**

**R. LAHARI 21S11A6217**

*Under the Guidance of*

***M.BHARATH KUMAR***

*B. Tech.*

## Asst. Professor of CSE

***DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING (Cyber Security)***

# MALLA REDDY INSTITUTE OF TECHNOLOGY & SCIENCE

*(Approved by AICTE New Delhi and Affiliated to JNTUH) (Accredited by NBA & NAAC with “A” Grade)*

*An ISO 9001: 2015Certified Institution Maisammaguda, Medchal (M), Hyderabad TS- 500100*

## JUNE 2024

*DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING (Cyber Security)*

# MALLA REDDY INSTITUTE OF TECHNOLOGY & SCIENCE

*(Approved by AICTE New Delhi and Affiliated to JNTUH) (Accredited by NBA & NAAC with “A” Grade)*

*An ISO 9001: 2015 Certified Institution Maisammaguda, Medchal (M), Hyderabad-500100, T. S.*

## JUNE 2024

***CERTIFICATE***

This is to certify that the mini project entitled **“Fake Job Post Prediction Using Machine Learning Algorithms”** has been submitted by **M. ABHINAI (21S11A6201), T. SRIKANTH (21S11A6256), J. SANTOSH GOUD (21S11A6250) and R. LAHARI (21S11A6217)** in partial fulfillment of the requirements for the award of **BACHELOR OF TECHNOLOGY**  **IN COMPUTER SCIENCE IN ENGINEERING & (CYBER SECURITY).** This record of bonafide work carried out by them under my guidance and supervision. **The result embodied in this mini project report has not been submitted to any other University or Institute for the award of any degree.**

# Mrs. T. Sai Kumari M.BHARATH KUMAR

**HOD-CSE(CS)****Project Guide**

**External Examiner**

**ACKNOWLEDGEMENT**

The Mini Project work carried out by our team in the Department of Computer Science and Engineering (CS), Malla Reddy Institute of Technology and Science, Hyderabad. ***This work is original and has not been submitted in part or full for any degree or diploma of any other university.***

We wish to acknowledge our sincere thanks to our project guide **Mr. M. BHARATH KUMAR** Asst.Professor of Computer Science & Engineering (CS) for formulation of the problem, analysis, guidance and her continuous supervision during the course of work.

We acknowledge our sincere thanks to **Dr. Vaka Murali Mohan,** Principal and **T.SAI KUMARI**, Head of the Department and Coordinator, faculty members of CSE(CS) Department for their kind cooperation in making this Mini Project work a success.

We extend our gratitude to **Sri. Ch. Malla Reddy**, Founder Chairman MRGI and **Sri. Ch. Mahender Reddy,** Secretary MRGI, **Dr.Ch. Bhadra Reddy**, President MRGI, **Sri. Ch. Shalini Reddy**, Director MRGI, **Sri. P. Praveen Reddy**, Director MRGI, for their kind cooperation in providing the infrastructure for completion of our Mini Project.

We acknowledge our special thanks to the entire teaching faculty and non-teaching staff members of the Computer Science & Engineering (CS)Department for their support in making this project work a success.

|  |  |  |
| --- | --- | --- |
| **M.ABHINAI** | **21S11A6201** |  |
| **T. SRIKANTH** | **21S11A6256** |  |
| **J. SANTOSH GOUD** | **21S11A6250** |  |
| **R. LAHARI** | **21S11A6217** |  |

|  |  |
| --- | --- |
| **INDEX**  **Chapter** | **Page no.** |
| **ABSTRACT** | **1** |
| **1. INTRODUCTION** | **2** |
| **2. LITERATURE SURVEY** | **4** |
| **3. EXISTING SYSTEM** | **6** |
| 3.1 Multinominal Naïve Bayes model | 7 |
| 3.1.1 Decision Tree Algorithm | 7 |
| 3.1.2 K-Nearest Neighbor (KNN) Algorithm | 9 |
| 3.2 Support Vector Machine Algorithm (SVM) | 9 |
| 3.2.1 Random Forest Algorithm | 10 |
| **4. PROPOSED SYSTEM** | **14** |
| 4.1 Emscad Dataset  4.2 Data Preprocessing in Machine learning  4.3 TF-IDF Feature extracton  4.4 Multilayer perception (MLP) | 15  15  17  19 |
| **5. UML DIAGRAMS** | **21** |
| **6. MACHINE LEARNING** | **25** |
| * 1. What is Machine Learning   2. Categories of Machine Learning   3. Need for Machine Learning   4. Applications of Machine Learning   5. How to Start Learning Machine Learning?   6. Advantages Of Machine Learning | 26  26  27  27  28  29 |
| **7. SOFTWARE ENVIRONMENT** | **31** |
| 7.1 What is python? | 32 |
| 7.2 Advantages of python | 32 |
| 7.3 Advantages of python over other languages | 33 |
| 7.4 History of python | 34 |
| 7.5 Python Development Steps | 34 |
| 7.6 Modules used In python  7.6.1 TensorFlow  7.6.2 Numpy  7.6.3 Pandas  7.6.4 Matplotlib  7.6.5 Scikit-learn  7. 7 Python Installnation | 35  35  35  35  36  36  37 |
| **8.SYSTEM REQUIREMENT SPECIFICATIONS**  8.1 Software Requirements  8.2 Hardware Requirements | **43**  44  44 |
| **9. FUNCTIONAL REQUIREMENTS** | **45** |
| **10. RESULT AND DISCUSSION** | **49** |
| **11. CONCLUSION** | **51** |

**PREDICTION OF FAKE JOB AD USING NLP-BASED MULTILAYER PERCEPTRON**

**ABSTRACT**

In modern time, the development in the field of industry and technology has opened a huge opportunity for new and diverse jobs for the job seekers. With the help of the advertisements of these job offers, job seekers find out their options depending on their time, qualification, experience, suitability etc. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous. Social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job ads has increased the percentage of fraud job postings which causes harassment to the job seekers. So, people lack in showing interest to new job postings due to preserve security and consistency of their personal, academic and professional information.

Therefore, this project proposed to use different data mining techniques and classification algorithm like K-nearest neighbour, decision tree, support vector machine, naive bayes classifier, random forest classifier, and multi-layer perceptron to predict a job Advertisement if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier, performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job ad.

**CHAPTER 1**

**INTRODUCTION**

* 1. **INTRODUCTION:**

In modern time, the development in the field of industry and technology has opened a huge opportunity for new and diverse jobs for the job seekers. With the help of the advertisements of these job offers, job seekers find out their options depending on their time, qualification, experience, suitability etc. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous. Social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this. Fake job posts create inconsistency for the job seeker to find their preferable jobs causing a huge waste of their time. An automated system to predict false job post opens a new window to face difficulties in the field of Human Resource Management.

**1.2 Fake Job Posting:**

Job Scam Online job advertisements which are fake and mostly willing to steal personal and professional information of job seekers instead of giving right jobs to them is known as job scam. Sometimes fraudulent people try to gather money illegally from job seekers. A recent survey by ActionFraud from UK has shown that more than 67% people are at great risk who look for jobs through online advertisements but unaware of fake job posts or job scam. In UK, almost 700000 job seekers complained to lose over $500000 being a victim of job scam. The report showed almost 300% increase over the last two years in UK. Students, fresh graduates are being mostly targeted by the frauds as they usually try to get a secured job for which they are willing to pay extra money.

**1.3 Common types of Job Scam:**

Fraudsters who want to gain other people’s personal information like insurance details, bank details, income tax details, date of birth, national id create fake job advertisements. Advance fee scams occur when fraudsAsk for money showing reasons like admin charges, information security checking cost, management cost etc. Sometimes fraudsters act them- selves as employers and ask people about passport details, bank statements, driving license etc. as pre-employment check. Illegal money mulling scams occur when they convince students to pay

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 LITERATURE SURVEY**

Habiba et. al proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naïve bayes classifier, random forest classifier, multilayer perceptron and deep neural network to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier, performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post.

Amaar et. al used six machine learning models to analyze whether these job ads are fraudulent or legitimate. Then, we compared all models with both BoW and TF-IDF features to analyze the classifier’s overall performance. One of the challenges in this study is our used dataset. The ratio of real and fake job posts samples is unequal, which caused the model over-fitting on majority class data. To overcome this limitation, we used the adaptive synthetic sampling approach (ADASYN), which help to balance the ratio between target classes by generating the number of samples for minority class artificially. We performed two experiments, one with the balanced dataset and the other with the imbalanced data. Through experimental analysis, ETC achieved 99.9% accuracy by using ADASYN as over-sampling ad TF-IDF as feature extraction. Further, this study also performs an in-depth comparative analysis of our proposed approach with state-of-the-art deep learning models and other re-sampling techniques.

Mehboob et. al handles the recruitment fraud/scam detection problem. Several important features of organization, job description and type of compensation are proposed and an effective recruitment fraud detection model is constructed using extreme gradient boosting method. It develops an algorithm that extracts required features from job ads and is tested using three examples. The features are further considered for two-step feature selection strategy. The findings show that features of the type of organization are most effective as a stand-alone model. The hybrid composition of selected 13 features demonstrated 97.94% accuracy and outperformed three state-of-the-art baselines. Moreover, the study finds that the most effective indicators are “salary range,” “company profile,” “organization type,” “required education” and “has multiple

**CHAPTER 3**

**EXISTING SYSTEM**

**3.1 Multinominal Naive Bayes Model**

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.

**3.1.1 How Multinomial Naive Bayes works?**

Naive Bayes is a powerful algorithm that is used for text data analysis and with problems with multiple classes. To understand Naive Bayes theorem’s working, it is important to understand the Bayes theorem concept first as it is based on the latter.Bayes theorem, formulated by Thomas Bayes, calculates the probability of an event occurring based on the prior knowledge of conditions related to an event. It is based on the following formula:

P(A|B) = P(A) \* P(B|A)/P(B)

P(B) = prior probability of B

P(A) = prior probability of class A

P(B|A) = occurrence of predictor B given class A probability

**3.2 Decision Tree Algorithm**

* Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome.
* In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed based on features of the given dataset.
* It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
* It is called a decision tree because, like a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.:

Diagram

Description automatically generated

FIG : **Decision Tree Representation**

**3.2.1 Features of Decision Tree Learning**

* Method for approximating discrete-valued functions (including boolean)
* Learned functions are represented as decision trees (or if-then-else rules)
* Expressive hypotheses space, including disjunction
* The decision tree is robust to noisy data

**3.3 K-Nearest Neighbor (KNN)** **Algorithm**

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* 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.
* 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.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

***Steps***

1. Get labeled data: The labeled data consists of features and labels. Features are the characteristics or the property of the object whereas labels are the class of the object with those features.
2. Convert labeled data to encoded data: Usually computations are based on numerical form so we convert the data to numeric form by encoding them.
3. Create feature set: Creating a set of features by packing the features.
4. Split the data for train and test: The data are split training and testing. Usually, 80% for training and 20% for testing but can select based on need.
5. Train the classifier: Training the classifier with the training data by specifying the value of k. Use k =3 for binary classification, i.e., two labels classification. If used k =1 then it is simply a nearest neighbor classifier.

**3.4 Support Vector Machine Algorithm (SVM)**

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.

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:

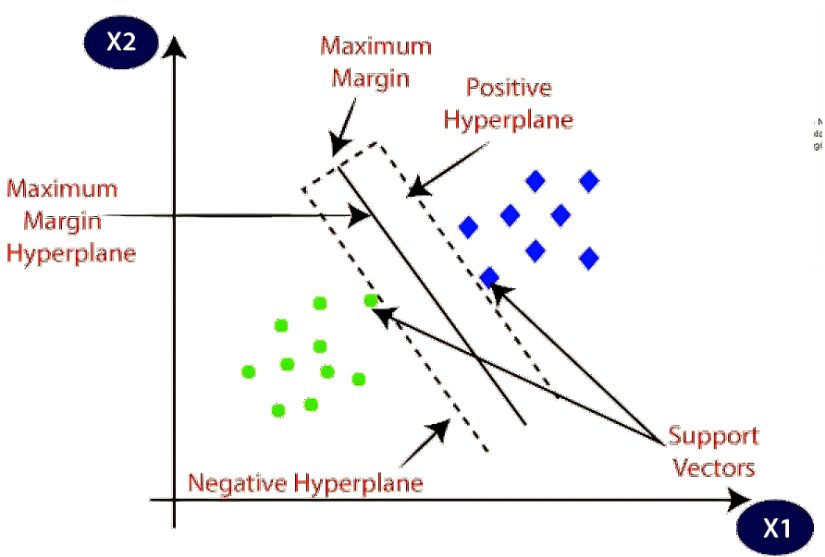


FIG: **SVM ALOGRITHM**

***What is text classification in SVM?***

Text Classification is the process of labeling or organizing text data into groups – it forms a fundamental part of Natural Language Processing. In the digital age that we live in, we are surrounded by text on our social media accounts, commercials, websites, eBooks, etc. The majority of this text data is unstructured, so classifying this data can be extremely useful.

***Columns***

* UserName
* ScreenName
* Location
* TweetAt

**3.5 Random Forest Algorithm**

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. 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 greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Diagram

Description automatically generated

Fig: Random Forest algorithm.

**3.5.1 Random Forest algorithm**

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression

**3.5.2 Important Features of Random Forest**

* ***Diversity****-* Not all attributes/variables/features are considered while making an individual tree, each tree is different.
* ***Immune******to******the******curse******of******dimensionality***- Since each tree does not consider all the features, the feature space is reduced.
* ***Parallelization****-*Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
* ***Train-Test******split****-* In a random forest we don’t have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
* ***Stability****-* Stability arises because the result is based on majority voting/ averaging.

**3.5.4 Types of Ensembles**

Ensemble uses two types of methods:

***Bagging****–* It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap.

**Diagram

Description automatically generated**

Fig.: RF Classifier analysis.

***Boosting****–* It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

Diagram

Description automatically generated

Fig. Boosting RF Classifier.

**3.5.5 Disadvantages of existing system**

* The prediction accuracy of this algorithm is lower than the other probability algorithms.
* It is not suitable for regression. Naive Bayes algorithm is only used for textual data classification and cannot be used to predict numeric values.

**CHAPTER 4**

**PROPOSED SYSTEM**

**4.1 EMSCAD Dataset**

The Employment Scam Aegean Dataset (EMSCAD) is a publicly available dataset containing 17,880 real-life job ads that aims at providing a clear picture of the Employment Scam problem to the research community and can act as a valuable testbed for scientists working on the field. To train the system, this project used EMSCAD dataset, where first row represents dataset column names and remaining rows contains dataset values such as Company profile.

***Columns***

* Title
* Location
* Department
* salary\_range
* compamy\_profile
* description
* requirements
* benefits
* telecommuting
* has\_company\_logo

**4.2 Data Preprocessing in Machine learning**

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

**4.2.1 Why do we need Data Pre-processing?**

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

* Getting the dataset
* Importing libraries
* Importing datasets
* Finding Missing Data

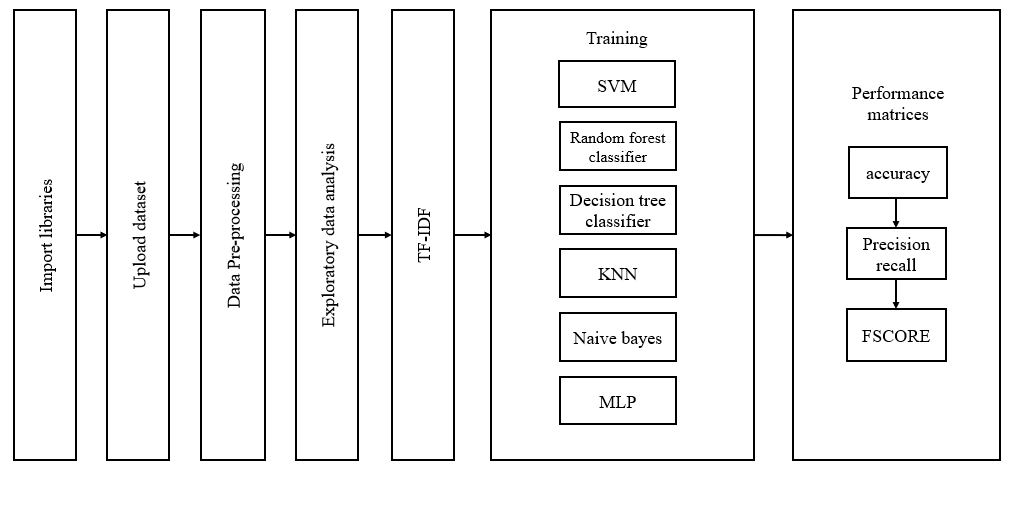
****

Fig.: Block diagram of proposed system.

**4.3 Splitting the Dataset into the Training set and Test set**

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So, we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as

A picture containing shape

Description automatically generated

Fig: Splitting of dataset.

**Training** **Set**: A subset of dataset to train the machine learning model, and we already know the output.

**Test** **set**: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

**4.3 TF-IDF Feature extraction**

SATF-IDF which stands for Term Frequency – Inverse Document Frequency. It is one of the most important techniques used for information retrieval to represent how important a specific word or phrase is to a given document. Let’s take an example, we have a string or Bag of Words (BOW) and we have to extract information from it, then we can use this approach.The tf-idf value increases in proportion to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps to adjust with respect to the fact that some words appear more frequently in general. TF-IDF use two statistical methods, first is Term Frequency and the other is Inverse Document Frequency. Term frequency refers to the total number of times a given term t appears in the document doc against (per) the total number of all words in the document and the inverse document frequency measure of how much information the word provides.

Diagram

Description automatically generated

Fig: TF-IDF block diagram.

TF-IDF do not convert directly raw data into useful features. Firstly, it converts raw strings or dataset into vectors and each word has its own vector. Then we’ll use a particular technique for retrieving the feature like Cosine Similarity which works on vectors, etc.

***Terminology***

t — term (word)

d — document (set of words)

N — count of corpus

corpus — the total document set

**4.4.1 Term Frequency (TF):** Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, “Data Science is awesome!” A simple way to start out is by eliminating documents that do not contain all three words “Data” is”, “Science”, and “awesome”, but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency

**4.4.2 Document Frequency:** This measures the importance of document in whole set of corpora, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N. In other words, DF is the number of documents in which the word is present. We consider one occurrence if the term consists in the document at least once, we do not need to know the number of times the term is present.

**4.4.3 Inverse Document Frequency (IDF)*:*** While computing TF, all terms are considered equally important. However, it is known that certain terms, such as “is”, “of”, and “that”, may appear a lot of times but have little importance. Thus, we need to weigh down the frequent terms while scale up the rare ones, by computing IDF, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

The TF-IDF now is at the right measure to evaluate how important a word is to a document in a collection or corpus. Here are many different variations of TF-IDF but for now let us concentrate on this basic version.

**4.4.4 Implementing TF-IDF:** To make TF-IDF from scratch in python, let’s imagine those two sentences from different document:

first\_sentence: “Data Science is the sexiest job of the 21st century”.

second\_sentence: “machine learning is the key for data science”.

First step we have to create the TF function to calculate total word frequency for all documents.

**4.5 Multilayer perceptron (MLP)**

MLP is one of the most frequently used neural network architectures in MDSS, and it belongs to the class of supervised neural networks. The multilayer perceptron consists of a network of nodes (processing elements) arranged in layers. A typical MLP network consists of three or more layers of processing nodes: an input layer that receives external inputs, one or more hidden layers, and an output layer which produces the classification results. Note that unlike other layers, no computation is involved in the input layer. The principle of the network is that when data are presented at the input layer, the network nodes perform calculations in the successive layers until an output value is obtained at each of the output nodes. This output signal should be able to indicate the appropriate class for the input data. That is, one can expect to have a high output value on the correct class node and low output values on all the rest. A node in MLP can be modelled as an artificial neuron

A picture containing text

Description automatically generated

where vj is the linear combination of inputs x1; x2; xp, qj is the bias, wji is the connection weight between the input xi

Diagram

Description automatically generated

Fig: Architecture of a multilayer perceptron network.

Diagram

Description automatically generated

Fig: One node of MLP: an artificial neuron.



The bias term qj contributes to the left or right shift of the sigmoid activation function, depending on whether qj takes a positive or negative value. Once the architecture of MLP has been determined, the connection weights of the network have to be computed through a training procedure based on the training patterns and the desired output. BP is one of the simplest and most general methods for the supervised training of MLP.

**CHAPTER 5**

**UML DIAGRAMS**

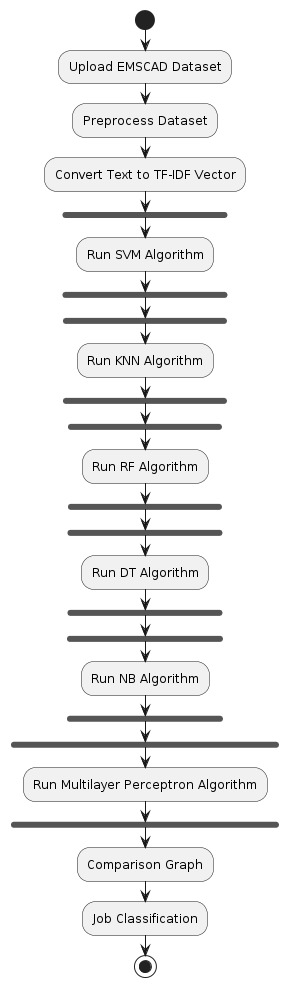


Fig.: UML DIAGRAM

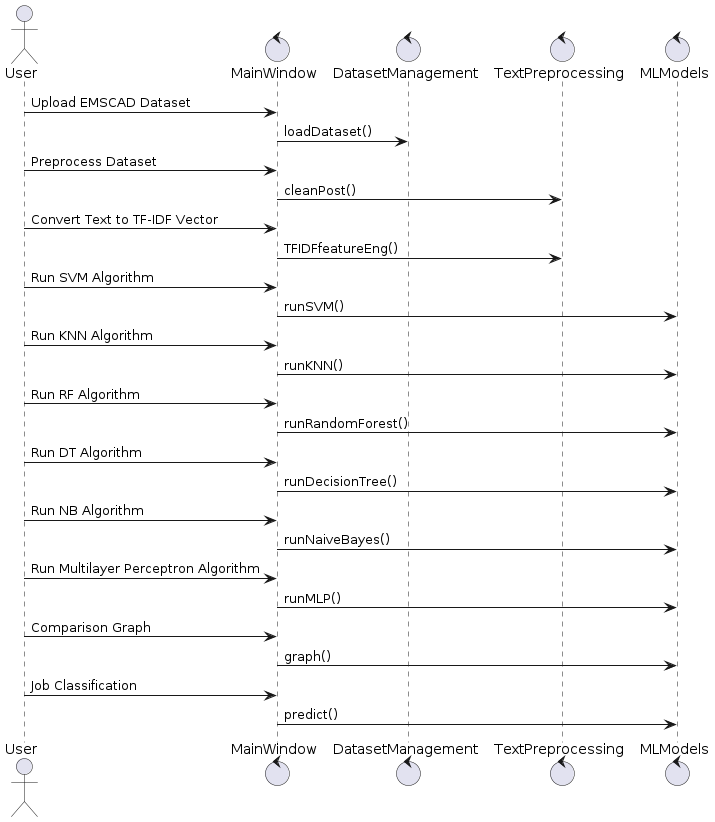


Fig.: DATA SET

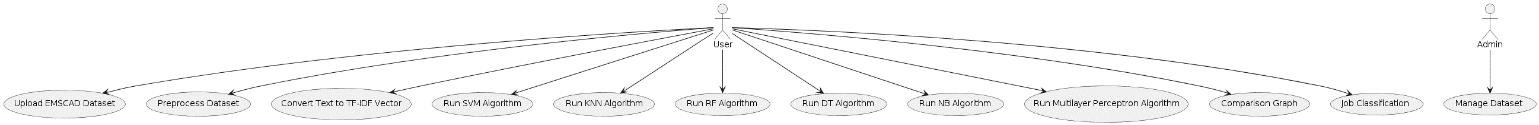


Fig.: ML ALGORITHMS

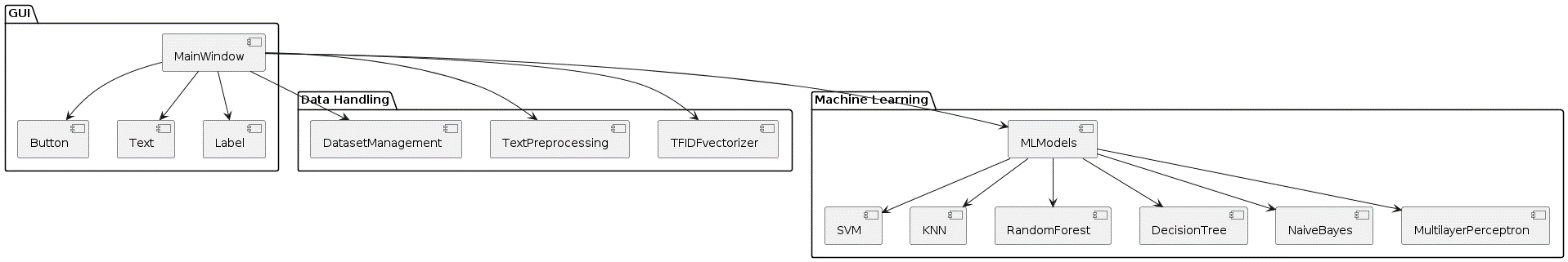


Fig.: ALGORITHMS

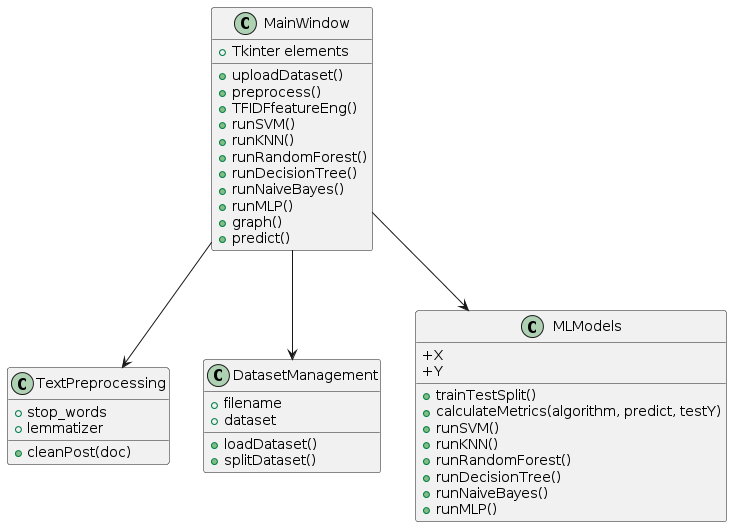


Fig.: TKINTER LIBRARIES

**CHAPTER 6**

**MACHINE LEARNING**

**6.1 What is Machine Learning**

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

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

**6.2 Categories of Machine Learning**

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

Unsupervised learning involves modeling the features of a dataset without reference to any label and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data,

**6.3 Need for Machine Learning**

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

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

**6.4.1 Applications of Machines Learning**

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

* Emotion analysis
* Sentiment analysis
* Error detection and prevention
* Weather forecasting and prediction
* Stock market analysis and forecasting
* Speech synthesis
* Speech recognition
* Customer segmentation

**6.5 How to Start Learning Machine Learning?**

Arthur Samuel coined the term “Machine Learning” in 1959 and defined it as a “Field of study that gives computers the capability to learn without being explicitly programmed”.And that was the beginning of Machine Learning! In modern times, Machine Learning is one of the most popular (if not the most!) career choices. According to Indeed, Machine Learning Engineer Is the Best Job of 2019 with a 344% growth and an average base salary of $146,085 per year.

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

**6.4.3 How to start learning ML?**

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

Step 1 – Understand the Prerequisites.

(a) Learn Linear Algebra and Multivariate Calculus

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

(b) Learn Statistics

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

(c) Learn Python

Some people prefer to skip Linear Algebra, Multivariate Calculus and Statistics and learn them as they go along with trial and error. But the one thing that you absolutely cannot skip is Python! While there are other languages you can use for Machine Learning like R, Scala, etc. Python is currently the most popular language for ML. In fact, there are many Python libraries that are specifically useful for Artificial Intelligence and Machine Learning such as Keras, TensorFlow, Scikit-learn, etc.

So, if you want to learn ML, it’s best if you learn Python! You can do that using various online resources and courses such as Fork Python available Free on GeeksforGeeks.

Step 2 – Learn Various ML Concepts

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

**6.6 Advantages of Machine learning**

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

2*.* **No human intervention needed (automation):**With ML, you don’t need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus software’s; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3**. Continuous Improvement***:* As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data, you have keeps growing, your algorithms learn to make more accurate predictions faster.

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

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

**CHAPTER 7**

**SOFTWARE ENVIRONMENT**

**7.1 What is Python?**

Below are some facts about Python.

* Python is currently the most widely used multi-purpose, high-level programming language.
* Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.
* Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.
* Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

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

* + Machine Learning
  + GUI Applications (like Kivy, Tkinter, PyQt etc.)
  + Web frameworks like Django (used by YouTube, Instagram, Dropbox)
  + Image processing (like Opencv, Pillow)
  + Web scraping (like Scrapy, BeautifulSoup, Selenium)
  + Test frameworks
  + Multimedia

**7.2 Advantages of Python**

Let’s see how Python dominates over other languages.

1. Extensive Libraries

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

2. Extensible

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

3. Embeddable

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

4. Improved Productivity

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

5. IOT Opportunities

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

**7.3 Advantages of Python Over Other Languages**

1. Less Coding

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

2. Affordable

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

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

3. Python is for Everyone

Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and machine learning, automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

**7.4 History of Python**

What do the alphabet and the programming language Python have in common? Right, both start with ABC. If we are talking about ABC in the Python context, it's clear that the programming language ABC is meant. ABC is a general-purpose programming language and programming environment, which had been developed in the Netherlands, Amsterdam, at the CWI (Centrum Wiskunde &Informatica). The greatest achievement of ABC was to influence the design of Python. Python was conceptualized in the late 1980s. Guido van Rossum worked that time in a project at the CWI, called Amoeba, a distributed operating system. In an interview with Bill Venners1, Guido van Rossum said: "In the early 1980s, I worked as an implementer on a team building a language called ABC at Centrum voor Wiskunde Informatica (CWI "Later on in the same Interview, Guido van Rossum continued: "I remembered all my experience and some of my frustration with ABC. I decided to try to design a simple scripting language that possessed some of ABC's better properties, but without its problems.

**7.5 Python Development Steps**

Guido Van Rossum published the first version of Python code (version 0.9.0) at alt sources in February 1991. This release included already exception handling, functions, and the core data types of lists, dict, str and others. It was also object oriented and had a module system.  
Python version 1.0 was released in January 1994Print is now a function. Views and iterators instead of lists

* The rules for ordering comparisons have been simplified. E.g., a heterogeneous list cannot be sorted, because all the elements of a list must be comparable to each other.
* There is only one integer type left, i.e., int. long is int as well.

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

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

* Python 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

**7.6 Modules Used in Project**

**7.6.1 TensorFlow**

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library and is also used for machine learning applications such as neural networks. It is used for both research and production at Google.‍

TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.

**7.6.2 NumPy**

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

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

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

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

**7.6.3 Pandas**

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

**7.6.4 Matplotlib**

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

**7.6.5 Scikit – learn**

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

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

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

* Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors.

**7.7 How to Install Python on Windows and Mac**

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

**Download the Correct version into the system**

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

A screenshot of a computer

Description automatically generated with medium confidence

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

Step 2: Click on the Download Tab.

Graphical user interface, application

Description automatically generated

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

Graphical user interface, application

Description automatically generated

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

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

* To download Windows 32-bit python, you can select any one from the three options: Windows x86 embeddable zip file, Windows x86 executable installer or Windows x86 web-based installer.
* To download Windows 64-bit python, you can select any one from the three options: Windows x86-64 embeddable zip file, Windows x86-64 executable installer or Windows x86-64 web-based installer.

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

**Installation of Python**

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

Graphical user interface, text, application

Description automatically generated

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

Graphical user interface, text, application, chat or text message

Description automatically generated

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

Graphical user interface, text, application, chat or text message

Description automatically generated

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

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

Verify the Python Installation

Step 1: Click on Start

Graphical user interface, application

Description automatically generated

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

Step 3: Open the Command prompt option.

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

A screenshot of a computer

Description automatically generated with medium confidence

Step 5: You will get the answer as 3.7.4

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

Check how the Python IDLE works

Step 1: Click on Start

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

Application

Description automatically generated with low confidence

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

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

Graphical user interface, text, application, email

Description automatically generated

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

Step 6: Now for e.g., enter print (“Hey World”) and Press Enter.

**CHAPTER 8**

**SYSTEM REQUIREMENTS SPECIFICATIONS**

**8.1 Software Requirements**

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation.

The appropriation of requirements and implementation constraints gives the general overview of the project in regard to what the areas of strength and deficit are and how to tackle them.

* Python IDLE 3.7 version (or)
* Anaconda 3.7 (or)
* Jupiter (or)
* Google Collab

**8.2 Hardware Requirements**

Minimum hardware requirements are very dependent on the particular software being developed by a given Enthought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

Operating system : Windows, Linux

Processor : minimum intel i3

Ram : minimum 4 GB

Hard disk : minimum 250GB

**CHAPTER 9**

**FUNCTIONAL REQUIREMENTS**

**9.1 Output Design**

Outputs from computer systems are required primarily to communicate the results of processing to users. They are also used to provides a permanent copy of the results for later consultation. The various types of outputs in general are:

* External Outputs, whose destination is outside the organization
* Internal Outputs whose destination is within organization and they are the
* User’s main interface with the computer.
* Operational outputs whose use is purely within the computer department.
* Interface outputs, which involve the user in communicating directly.

**9.2 Output Definition**

The outputs should be defined in terms of the following points:

* Type of the output
* Content of the output
* Format of the output
* Location of the output
* Frequency of the output
* Volume of the output
* Sequence of the output

It is not always desirable to print or display data as it is held on a computer. It should be decided as which form of the output is the most suitable.

**9.3 Input Design**

Input design is a part of overall system design. The main objective during the input design is as given below:

* To produce a cost-effective method of input.
* To achieve the highest possible level of accuracy.
* To ensure that the input is acceptable and understood by the user.

**9.4 Input Stages**

The main input stages can be listed as below:

* Data recording
* Data transcription
* Data conversion
* Data verification
* Data control
* Data transmission
* Data validation
* Data correction

Keeping in view the above description of the input types and input media, it can be said that most of the inputs are of the form of internal and interactive. As

Input data is to be the directly keyed in by the user, the keyboard can be considered to be the most suitable input device.

**Error Avoidance**

At this stage care is to be taken to ensure that input data remains accurate form the stage at which it is recorded up to the stage in which the data is accepted by the system. This can be achieved only by means of careful control each time the data is handled.

The system immediately prompts the user and the user has to again key in the data and the system will accept the data only if the data is correct. Validations have been included where necessary.

The system is designed to be a user friendly one. In other words the system has been designed to communicate effectively with the user. The system has been designed with popup menus.

**User Interface Systems Can Be Broadly Clasified As:**

* User initiated interface the user is in charge, controlling the progress of the user/computer dialogue. In the computer-initiated interface, the computer selects the next stage in the interaction..

**User Initiated Intergfaces**

User initiated interfaces fall into two approximate classes:

* Command driven interfaces: In this type of interface the user inputs commands or queries which are interpreted by the computer.
* Forms oriented interface: The user calls up an image of the form to his/her screen and fills in the form. The forms-oriented interface is chosen because it is the best choice.

**Computer-Initiated Interfaces**

The following computer – initiated interfaces were used:

* The menu system for the user is presented with a list of alternatives and the user chooses one; of alternatives.
* Questions – answer type dialog system where the computer asks question and takes action based on the basis of the users reply.

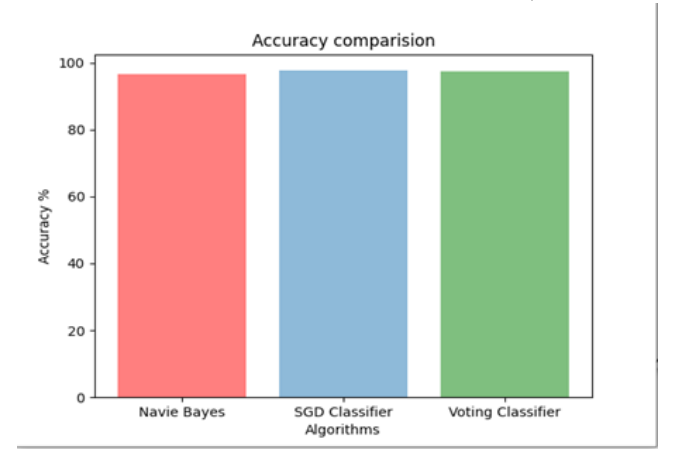
Right from the start the system is going to be menu driven, the opening menu displays the available options. Choosing one option gives another popup menu with more options.

**CHAPTER 10**

**RESULTS AND DISCUSSION**

**10.1 RESULTS AND DISCUSSION**

The system was created using Windows 11 as well as a 64-bit processor with 8 GB of RAM. The model implemented with the help of Python v3.7.8

****

14

Accuracy comparison between algorithms from the values calculated in the confusion matrix, an accuracy graph is generated for each algorithm for comparison for best algorithm with highest accuracy.

**CHAPTER 11**

**CONCLUSION**

**CONCLUSION**

Job scam detection has become a great concern all over the world at present. In this project, we have analyzed the impacts of job scam which can be a very prosperous area in research filed creating a lot of challenges to detect fraudulent job posts. We have experimented with EMSCAD dataset which contains real life fake job posts. In this paper, we have experimented both machine learning algorithms SVM, KNN, Naive Bayes, Random Forest and a neural network concept called MLP. This work shown the evaluation of machine learning and MLP-based classifiers.

**REFERENCES**

1. B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection,” J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009
2. S. U. Habiba, M. K. Islam and F. Tasnim, "A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 543-546, doi: 10.1109/ICREST51555.2021.9331230.
3. Amaar, A., Aljedaani, W., Rustam, F. et al. Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches. Neural Process Lett 54, 2219–2247 (2022). <https://doi.org/10.1007/s11063-021-10727-z>
4. S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, “Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset”, Future Internet 2017, 9, 6; doi:10.3390/fi9010006.
5. Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, IEEE 36th International Conference on Data Engineering (ICDE), 2020.
6. Mehboob, A., Malik, M.S.I. Smart Fraud Detection Framework for Job Recruitments. Arab J Sci Eng 46, 3067–3078 (2021). <https://doi.org/10.1007/s13369-020-04998-2>