# CHAPTER 1 INTRODUCTION

## MACHINE LEARNING AND NLP

* + 1. **MACHINE LEARNING**

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory, and application domains to the field of machine learning. "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.” This is Alan Turing’s definition of machine learning.

## NATURAL LANGUAGE PROCESSING

NLP is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, how to program computers to fruitfully process large amounts of natural language data. Natural language processing (NLP) is a sub field of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural)

languages, how to program computers to process and analyse large amounts of natural language data.

## OBJECTIVE

There are numerous factors that contribute to the dissemination of fake news. The first is due to a lack of data among the public. The readers are uninformed of the sources' legitimacy, and hence the news' veracity. Being listened to This will have a huge detrimental influence on the public . The lack of automatic fact-checking procedures is the second reason. Fake news detection is attempted by websites such as Politifact1, Full Fact2, and AltNews3, but the time-consuming human process is just too slow to prevent the initial dissemination of false information. Detecting false news automatically is a difficult task that defines present contentbased analysis method.

## MOTIVATION OF WORK

Social media facilitates the creation and sharing of information that uses computer-mediated technologies. This media changed the way groups of people interact and communicate. It allows low cost, simple access and fast dissemination of information to them. The majority of people search and consume news from social media rather than traditional news organizations these days. On one side, where social media have become a powerful source of information and bringing people together, on the other side it also 1 put a negative impact on society. Look at some examples herewith; Facebook Inc’s popular messaging service, WhatsApp became a political battleplatform in Brazil’s election. False rumours, manipulated photos, decontextual videos, and audio jokes were used for campaigning. These kinds of stuff went viral on the digital platform without monitoring their origin or

reach. A nationwide block on major social media and messaging sites including Facebook and Instagram was done in Sri Lanka after multiple terrorist attacks in the year 2019. The government claimed that “false news reports” were circulating online. This is evident in the challenges the world's most powerful tech companies face in reducing the spread of misinformation. Such examples show that social media enables the widespread use of “fake news” as well. The news assimilated on social media platforms may be of low quality carrying misleading information intentionally. This sacrifices the credibility of the information. Millions of news articles are being circulated every day on the Internet – how one can trust which is real and which is fake? Thus, incredible or fake news is one of the biggest challenges in our digitally connected world. Fake news detection on social media has recently become an emerging research domain. The domain focuses on dealing with the sensitive issue of preventing the spread of fake news on social media. Fake news identification on social media faces several challenges. Firstly, it is difficult to collect fake news data. Furthermore, it is difficult to label fake news manually. Since they are intentionally written to mislead readers, it is difficult to detect them simply based on news content. Furthermore, Facebook, Whats App, and Twitter are closed messaging apps. The misinformation disseminated by trusted news outlets or their friends and family is therefore difficult to be considered as fake. It is not easy to verify the credibility of newly emerging and time-bound news as they are not sufficient to train the application data set. Significant approaches to differentiate credible users, extract useful news features and develop authentic information dissemination systems are some useful domains of research and need further investigations. If we cannot control the spread of fake news, the trust in the system will collapse. There will be widespread distrust among people. There will be nothing left that can be objectively used.

It means the destruction of political and social coherence. We wanted to build some sort of web-based system that can fight this nightmare scenario. And we made some significant progress towards that goal.

## PROBLEM STATEMENT

News consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news. It enables the wide spread of “fake news,” i.e., low quality news with intentionally false information. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Therefore, fake news detection has recently become emerging research that is attracting tremendous attention. First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content. To develop a fake news detection system using natural language processing and its accuracy will be tested using machine learning algorithms. The algorithm must be able to detect fake news in a given scenario. Now, it is extremely difficult to decide whether the news we come across is real or not. There are very few options to check the authenticity and all of them are sophisticated and not accessible to the average person. There is an acute need for a web-based factchecking platform that harnesses the power of Machine Learning to provide us with that opportunity.

# CHAPTER 2

**LITERATURE SURVEY**

1. **Title:** Automatic Deception Detection: Methods for Finding Fake News.

**Author:** Niall J, Conroy, Victoria L. Rubin, Yilin Chen.

**Year :** 2015

“Fake news detection” is defined as the task of categorizing news along a continuum of veracity, with an associated measure of certainty. Veracity is compromised by the occurrence of intentional deceptions. The nature of online news publication has changed, such that traditional fact checking and vetting from potential deception is impossible against the flood arising from content generators, as well as various formats and genres. These methods have emerged from separate development streams, utilizing disparate techniques. In this survey, two major categories of methods emerge: 1. Linguistic Approaches in which the content of deceptive messages is extracted and analyzed to associate language patterns with deception; and 2. Network Approaches in which network information, such as message meta data or structured knowledge network queries can be harnessed to provide aggregate deception measures. Both forms typically incorporate machine learning techniques for training classifiers to suit the analysis

1. **Title:** Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques.

**Author:** Hadera Ahmed, Issa Traore, Sheriff Saad. **Year :** 2015 fake news is because identifying these entities require measuring the news propagation, which has shown to be complex and resource intensive [3]. Trend Micro, a cyber security company, analyzed hundreds of fake news services provider around the globe.

They reported that it is effortless to purchase one of those services. In fact, according to the report, it is much cheaper for politicians and political parties to use those services to manipulate election outcomes and people opinions about certain topics [4, 5]. Detecting fake news is believed to be a complex task and much harder than detecting fake product reviews given that they spread easily using social media and word of mouth. We present in this paper an n-gram features based approach to detect fake news, which consists of using text analysis based on n-gram features and machine learning classification techniques.

1. **Title:** Development of Fake News Model using Machine Learning through Natural Language Processing.

**Author:** Sajjad Ahmed, Knut Finkelman, Flavio Corradino.

**Year :** 2020

Facebook post prediction through real or fake labelling can be done through naïve Bayes and it performs well [6]. A proposed method can separate fake contents in three categories: serious fabrication, large scale hoaxes and humorous fake [11]. It can also provide a way to filter, vet and verify the news. PHEME was a threeyear research project funded by the European Commission from 2014-2017, studying NLP techniques for dealing rumour detection, stance detection [8] and [9], contradiction detection and analysis of social media rumors. Fake news stories can be easily shared on social media platforms but it is difficult to identify fake content automatically.

1. **Title:** Fake News Detection on Social Media: A Data Mining Perspective.

**Author:** Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu.

**Year:** 2021

Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news from social media. On the other hand, it enables the wide spread of "fake news", i.e., low quality news with intentionally false information.

The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Therefore, fake news detection on social media has recently become an emerging research that is attracting tremendous attention. Fake news detection on social media presents unique characteristics and challenges that make existing detection algorithms from traditional news media ineffective or not applicable. First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content; therefore, we need to include auxiliary information, such as user social engagements on social media, to help make a determination. Second, exploiting this auxiliary information is challenging in and of itself as users' social engagements with fake news produce data that is big, incomplete, unstructured, and noisy. Because the issue of fake news detection on social media is both challenging and relevant, we conducted this survey to further facilitate research on the problem. In this survey, we present a comprehensive review of detecting fake news on social media, including fake news characterizations on psychology and social theories, existing algorithms from a data mining perspective, evaluation metrics and representative datasets. We also discuss related research areas, open problems, and future research directions for fake news detection on social media.

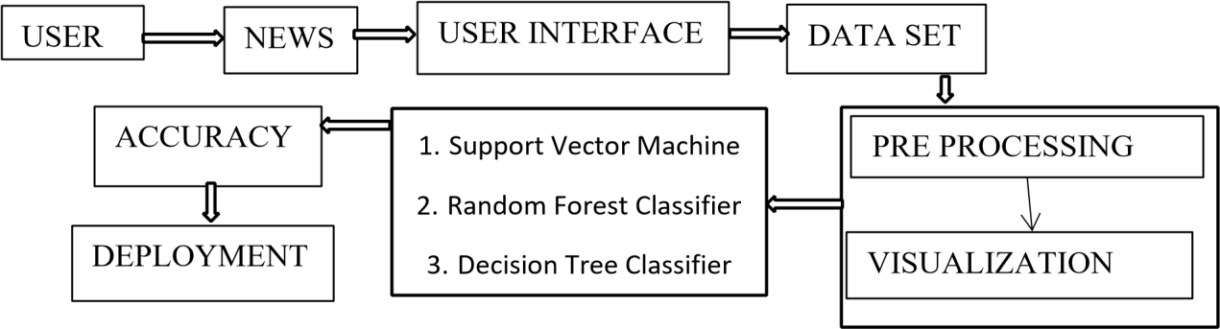
# CHAPTER 3

**SYSTEM DESIGN**

System design is the process architecture, components, modules, interfaces and data for a system to satisfy specified requirements. System design could be seen as the application of systems theory to product development. System design is the process of defining the elements of a system such as architecture, modules, and components, the different interfaces of those components and the data go through that system. It is meant to satisfy specific needs and requirements of a business or organization through the engineering of a coherent and well-running system.

## ARCHITECTURE DIAGRAM

An architectural diagram is a diagram of a system that is used to abstract the overall outline of the software system and the relationships, constraints, and boundaries between components. It provides an overall view of the physical deployment of the software system and its evolution road map.



## FIG 3.1 ARCHITECTURE DIAGRAM

* 1. **UML DIAGRAM**

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

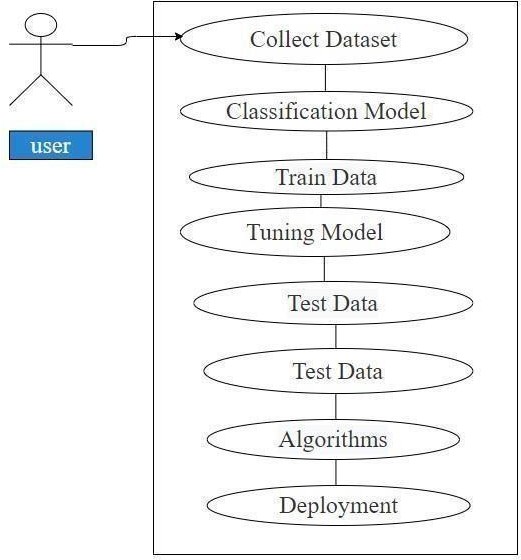
* + - * business processes
      * (logical) components
      * activities
      * programming language statements
      * database schema
      * Reusable software components.

UML combines best techniques from data modeling (entity relationship diagrams), business modeling (work flows), object modeling, and component modeling. It can be used with all processes, throughout the software development life cycle, and across different implementation technologies. UML has synthesized the notations of the Booch method, the Object-modeling technique (OMT) and

Object-oriented software engineering (OOSE) by fusing them into a single, common and widely usable modeling language. UML aims to be a standard modeling language which can model concurrent and distributed systems.

## USE CASE DIAGRAM

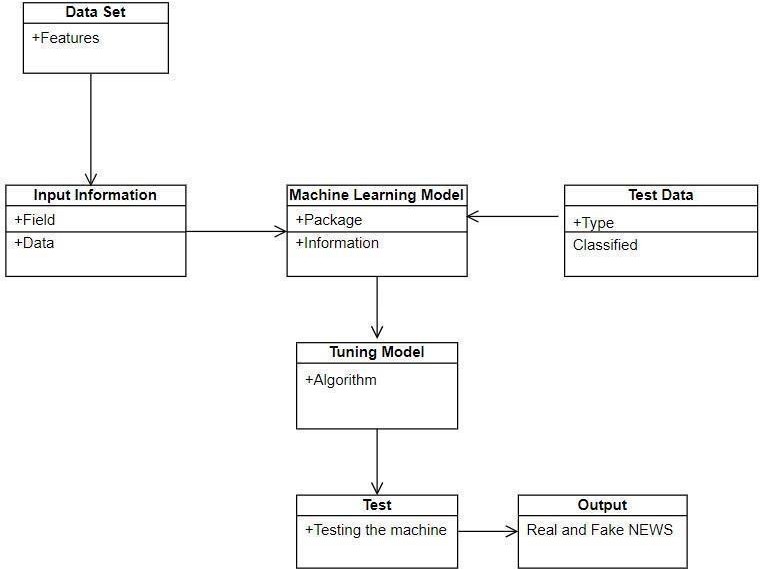
Use case diagrams are considered for high level requirement analysis of a system. So when the requirements of a system are analyzed the functionality are captured in use cases. So, it can say that uses cases are nothing but the system functionality written in an organized manner.



## FIG 3.2.1 USE CASE DIAGRAM

## CLASS DIAGRAM

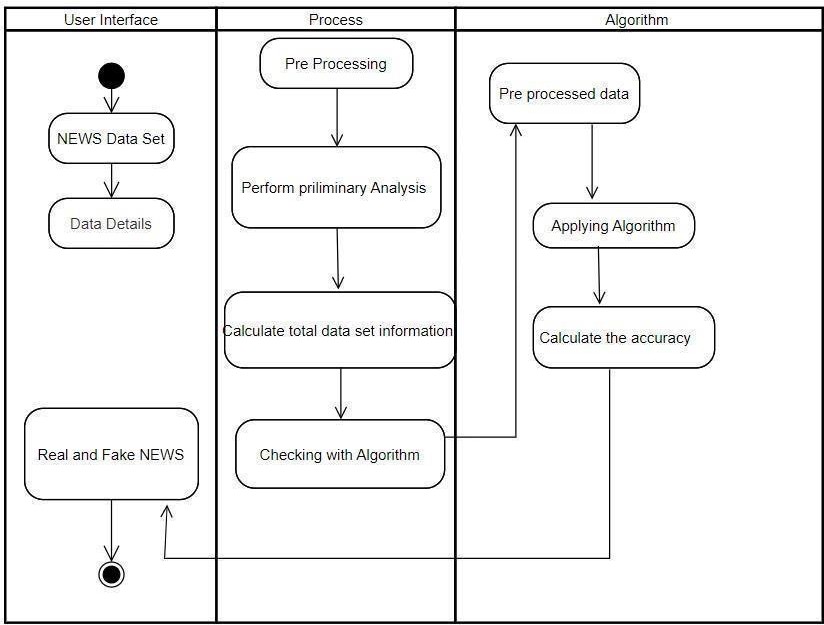
Class diagram is basically a graphical representation of the static view of the system and represents different aspects of the application. So, a collection of class diagrams represent the whole system. The name of the class diagram should be meaningful to describe the aspect of the system. Each element and their relationships should be identified in advance Responsibility (attributes and methods) of each class should be clearly identified for each class minimum number of properties should be specified and because, unnecessary properties will make the diagram complicated. Use notes whenever required to describe some aspect of the diagram and at the end of the drawing it should be understandable to the developer/coder. Finally, before making the final version, the diagram should be drawn on plain paper and rework as many times as possible to make it correct.



## FIG 3.2.2 CLASS DIAGRAM

## ACTIVITY DIAGRAM

Activity is a particular operation of the system. Activity diagrams are not only used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is some time considered as the flow chart. Although the diagrams looks like a flow chart but it is not. It shows different flow like parallel, branched, concurrent and single.

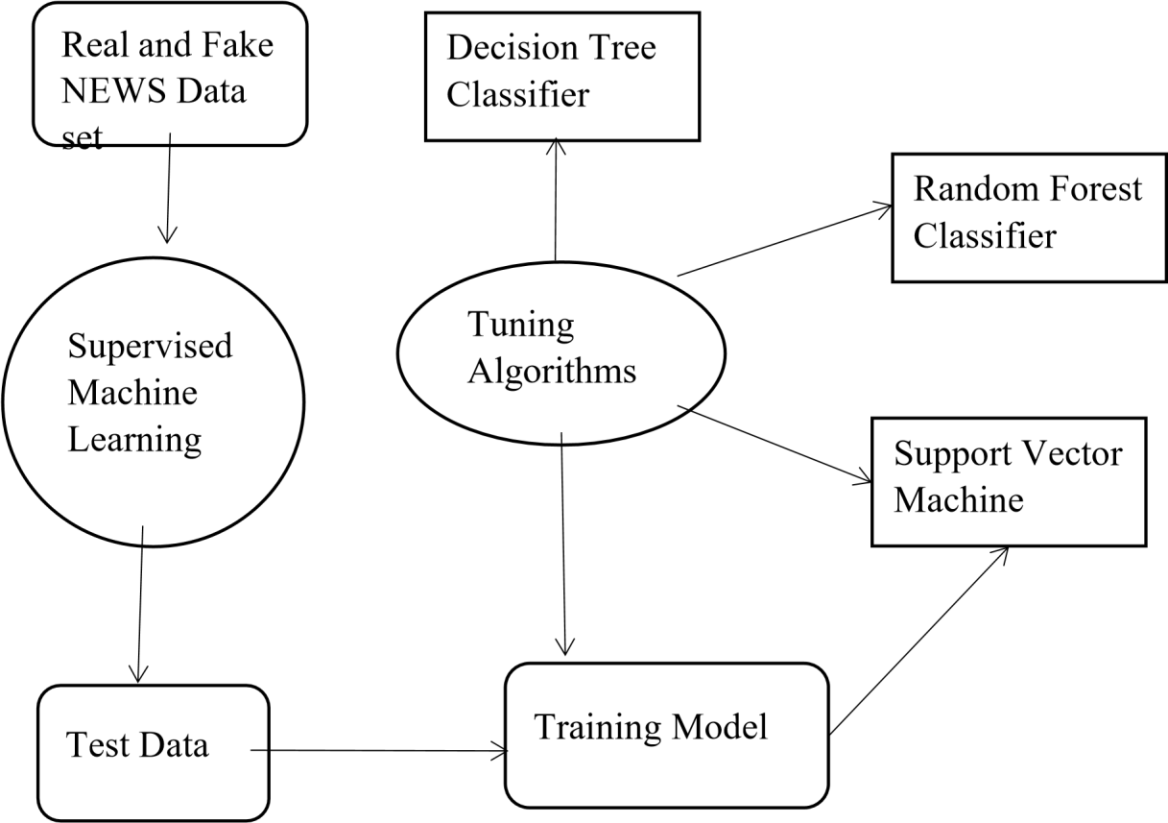


## FIG 3.2.3 ACTIVITY DIAGRAM

* 1. **ENTITY RELATIONSHIP DIAGRAM**

An entity relationship diagram (ERD), also known as an entity relationship model, is a graphical representation of an information system that depicts the relationships among people, objects, places, concepts or events within that system. An ERD is a data model technique that can help define business processes and be

used as the foundation of a relational database. Entity relationship diagrams provide a visual starting point for database design that can also be used to help determine information system requirements throughout an organization. After a relational database is rolled out, an ERD can still serve as a referral point, should any debugging or business process re-engineering be needed later.



**FIG 3.3 ENTITY RELATIONSHIP DIAGRAM**

# CHAPTER 4 MODULE DESCRIPTION

## DATA PREPROCESSING

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the data set. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given data set. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training data set while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation data set is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

## PANDAS

A number of different data cleaning tasks using Python’s pandas library and specifically, it focus on probably the biggest data cleaning task, missing values and it able to more [quickly clean data. It](https://www.dataoptimal.com/data-cleaning-with-python-2018/) wants to spend less time cleaning data, and more time exploring and modeling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It’s important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data Before, joint into code, it’s important to understand the sources of missing data. Here are some typical reasons why data is missing:

Variable identification with Uni-variate, Bi-variate and Multivariate analysis:

* import libraries for access and functional purpose and read the given data set
* General Properties of Analyzing the given data set
* Display the given data set in the form of data frame
* show columns
* shape of the data frame
* To describe the data frame
* Checking data type and information User forgot to fill in a field.
* Data was lost while transferring manually from a legacy database.
* There was a programming error.
* Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.
* on about data set
* Checking for duplicate data
* Checking Missing values of data frame
* Checking unique values of data frame
* Checking count values of data frame
* Rename and drop the given data frame
* To specify the type of values
* To create extra columns

GIVEN INPUT EXPECTED OUTPUT input

: data output : removing noisy data

## DATA VISUALIZATION

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a data set and can help with identifying patterns, corrupt data, out liners, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

* How to chart time series data with line plots and categorical quantities with bar charts.
* How to summarize data distributions with histograms and box plots. GIVEN INPUT EXPECTED OUTPUT

input : data output :

visualized data

## TREE BASEC CLASSIFICATION

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 on the basis of features of the given data set. 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, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

GIVEN INPUT EXPECTED OUTPUT

input : data output : getting accuracy

## RANDOM CLASSIFICATION

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 data set and takes the average to improve the predictive accuracy of that data set." 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 over fitting.

GIVEN INPUT EXPECTED OUTPUT

input : data output : getting accuracy

## VECTOR CLASSIFICATION:

The objective of the vector classification is to find a hyper plane in an N dimensional space that distinctly classifies the data pointed. To separate the two classes of data points, there are many possible hyper planes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

GIVEN INPUT EXPECTED OUTPUT

input : data output : getting accuracy

## DEPLOYMENT

In this module the trained machine learning model is converted into pickle data format file (.pkl file) which is then deployed for providing better user interface and predicting the Real or fake news.

GIVEN INPUT EXPECTED OUTPUT

input : data values output

: predicting output

## Flask (Web Frame Work)

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require tools or libraries. It has no database abstraction layer, form validation, or any other components where pre - existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

## FEATURES

Flask was designed to be easy to use and extend. The idea behind Flask is to build a solid foundation for web applications of different complexity. From then on you are free to plug in any extensions you think you need. Also, you are free to build your own modules. Flask is great for all kinds of projects. It's especially good for prototyping. Flask depends on two external libraries: the Jinja2 template engine and the WSGI toolkit. Still the question remains why use Flask as your web application framework if we have immensely powerful [Dja](https://quintagroup.com/services/python/django)[ngo, Pyramid, and](https://quintagroup.com/cms/python/pyramid) don’t forget web mega-framework [Turbo-gears? Tho](https://quintagroup.com/cms/python/turbogears)se are supreme [Python web](https://quintagroup.com/services/python/python-web-development.png) [frameworks BUT](https://quintagroup.com/services/python/python-web-development.png) out-of-the-box [Flask is p](http://quintagroup.com/cms/python/flask)retty impressive too with its.

* Built-In Development server and Fast debugger
* integrated support for unit testing
* REST full request dispatching
* Uses [Jinja2 Temp](https://quintagroup.com/cms/python/jinja2)lating
* support for secure cookies
* Unicode based
* Extensive Documentation
* Google App Engine Compatibility
* Extensions available to enhance features desired

Plus, Flask gives you so much more control on the development stage of your project.

It follows the principles of minimalism and let you decide how you will build your application.

* Flask has a lightweight and modular design, so it easy to transform it to the web framework you need with a few extensions without weighing it down
* ORM-agnostic: you can plug in your favourite ORM
* Basic foundation API is nicely shaped and coherent
* Flask documentation is comprehensive, full of examples and well structured. You can even try out some sample application to really get a feel of Flask. *•* It is super easy to deploy Flask in production (Flask is 100% [WSGI](http://www.python.org/dev/peps/pep-0333/) 1.0 compliant”)
* HTTP request handling functionality
* High Flexibility

# CHAPTER 5 ALGORITHM IMPLEMENTATION

## IMPLEMENTATION

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with sci kit - learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using re - sampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new data set, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracy. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is

evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

## Performance metrics to calculate:

**False Positives (FP):** A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

**False Negatives (FN):** A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

**True Positives (TP):** A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

**True Negatives (TN):** A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

True Positive Rate(TPR) = TP / (TP + FN) False Positive rate(FPR) = FP / (FP + TN)

**Accuracy:** The Proportion of the total number of predictions that is correct

otherwise overall how often the model predicts correctly defaulters and nondefaulters.

## Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric data sets where values of false positive and false negatives are almost same.

**Precision:** The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

**Recall:** The proportion of positive observed values correctly predicted.

Recall = TP / (TP + FN)

Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

**F1 Score** is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

## General Formula:

F- Measure = 2TP / (2TP + FP + FN)

## F1-Score Formula:

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

The three different algorithms are compared:

* + Decision tree classifier
  + Random forest Classifier
  + Support Vector Machine

# CHAPTER 6 SYSTEM ANALYSIS

## GENERAL

In system analysis, we analyse the hardware and software environment required for the implementation of the proposed system.

## HARDWARE ENVIRONMENT

Processor : Intel i3

Hard disk : minimum 80 GB

RAM : minimum 2 GB

## SOFTWARE REQUIREMENTS

Operating System : Windows 10 or later

Tool : Anaconda with Jupyter Notebook

## JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations and other multimedia resources, along with explanatory text in a single document.

## ANACONDA

Anaconda is an open-source Python distribution / data discovery & analytics platform that allows you to launch applications and manage conda packages, environments, and

channels without using command line interface (CLI) commands.

## Feasibility study:

**Data Wrangling**

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

## Data collection

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The

Data Model which was created using Random Forest, logistic, Decision tree algorithms, Support vector classifier (SVC), Multilayer Perceptron are applied on the Training set and based on the test result accuracy, Test set prediction is done.

## Preprocessing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done.

## Building the classification model

The prediction of Real and fake job classification, a high accuracy prediction model is effective because of the following reasons: It provides better results in classification problem.

* It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables.
* It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

## Construction of a Predictive Model

Machine learning needs data gathering have lot of past data’s. Data gathering have sufficient historical data and raw data. Before data pre-processing, raw data can’t be used directly. It’s used to pre-process then, what kind of algorithm with model. Training and testing this model working and predicting correctly with minimum errors. Tuned model involved by tuned time to time with improving the accuracy.

## CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set of higher accuracy score algorithm will be find out. The founded one is used in the application which can help to find the Real and fake news. With the development in digital world, the online fake content is increasing drastically. Being in common people’s access, such fake widespread content can cause potential set-back to journalism and democracy by misleading people. The decisions and opinions of the public are greatly influenced by false content as they spread much quicker and leave a great impact. Such contents gain popularity because most of the users on social media are unaware about certain topics and are easily deluded by fake content. Other reasons may include people’s reliance on online media platforms and the catchy draft of fake news. With ever-increasing fake content the research on the fake news detection is also making progress and various approaches from vast domains have been implemented ranging from AI to linguistic and knowledge engineering, but no ideal methodology has been devised yet that can accurately classify real news and fake news. The major challenge in this task is the growing social media content, which is increasing exponentially daily. The social sites and apps, such as Facebook, Twitter, Instagram, and other blogs have enabled people to put any random unchecked content including their personal opinions and thoughts that has become quite challenging in devising a fake news detection algorithm. In this article, a thorough study on existing fake news detection techniques along with a new taxonomy is included and major challenges of fake news detection have been discussed with future recommendations to further improve this area.

## 

## REFERENCE

1. J. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proenca, and J. Fierrez, “Ganprintr: Improved fakes and evaluation of the state of the art in

face manipulation detection,” IEEE J. Sel. Topics Signal Process.,vol. 14, no. 5, pp. 1038–1047, Aug. 2020.

1. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: three types of fakes. In: Proceedings of the 78th ASIS&T Annual Meeting:

Information Science with Impact: Research in and for the Community (ASIST 2015). Article 83, p. 4, American Society for Information

Science, Silver Springs (2015)

1. P. H. A. Faustini and T. F. Covões, “Fake news detection in multiple platforms and languages,” Expert Syst. Appl., vol. 158, Nov. 2020, Art. no.

113503.

1. Y. Liu and Y. Wu, “FNED: A deep network for fake news early detection on social media,” ACM Trans. Inf. Syst., vol. 38, no. 3, pp. 1–33, 2020.
2. Schulten, K.: Skills and Strategies—Fake News vs. Real News: Determining the Reliability of Sources. The Learning Network (2017). https://learning.blogs.nytimes.com/2015/10/02/ skillsandstrategiesfakenewsvs-real-news-determining-the-reliability-of- sources/. Accessed 16 Aug 2017.
3. N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news,” Proceedings of the Association for

Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

1. Schow, A.: The 4 Types of ‘Fake News’. Observer (2017). <http://observer.com/2017/01/>fake-news-russia-hacking-clinton-loss