(An Autonomous Institute affiliated to Savitribai Phule Pune University)

FAKE NEWS DETECTION SYSTEM USING MACHINE LEARNING

SY.B.Tech. Minor Project Report

SUBMITTED BY

SANSKAR SHARMA	**[S194059]**
PRATIKSHA SABALE	**[S194090]**
NAKUL AGGARWAL	**[S194031]**

GUIDED BY

Mrs. Kavitha

SCHOOL OF COMPUTER ENGINEERING & TECHNOLOGY

MIT ACADEMY OF ENGINEERING, ALANDI(D), PUNE-412105,

MAHARASHTRA (INDIA)

MAY, 2020

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

FAKE NEWS DETECTION SYSTEM USING MACHINE LEARNING

A Minor Project Report

Submitted in partial fulfilment of the requirement for the award of the degree

of

Bachelor of Technology

in

Computer science & engineering

By

Sanskar Sharma,
Pratiksha Sabale
&
Nakul Aggarwal

SCHOOL OF COMPUTER ENGINEERING & TECHNOLOGY

MIT ACADEMY OF ENGINEERING, ALANDI(D), PUNE-412105,

MAHARASHTRA (INDIA)

MAY, 2020

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

CERTIFICATE

It is hereby certified that the work which is being presented in the SY.B.Tech Minor Project Report entitled **"FAKE NEWS DETECTION USING MACHINE LEARNING"** **,** in partial fulfilment of the requirements for the award of the **Bachelor of Technology in Computer Engineering** and submitted to the **SCHOOL OF COMPUTER ENGINEERING & TECHNOLOGY of MIT ACADEMY OF ENGINEERING, ALANDI(D), PUNE** is an authentic record of work carried out during a period from January 2020 to July 2020 under the supervision of **Mrs. Kavitha**, **School of Computer Engineering & Technology**.

SANSKAR SHARMA	**PRN No. 0120180381**	**Exam Seat No. S194059**
PRATIKSHA SABALE	**PRN No. 0120180481**	**Exam Seat No. S194090**
NAKUL AGGARWAL	**PRN No. 0120180186**	**Exam Seat No. S194031**

Date:

Signature of Project Advisor
Project Adviser
School of Computer Engineering and Technology,
MIT Academy of Engineering, Alandi(D), Pune

Signature of Dean
Dean
School of Computer Engineering and Technology,
MIT Academy of Engineering, Alandi(D), Pune

(STAMP/SEAL)

Signature of Internal examiner/s
Name…………………………………
Affiliation…………………………

Signature of External examiner/s
Name…………………………………
Affiliation…………………………

ACKNOWLEDGEMENT

We want to express our heartfelt gratitude to **Dr. Rajeshwari Goudar ma'am** for her continuous support. From searching the domain to deciding the mathematical model for our selected problem statement she has been of great help. Our undying gratitude to **Mr. Jayvant H. Devare sir** for his continuous updates and suggestions. We would also like to express our utmost gratitude to **Mrs. Kavitha ma'am** for her relentless guidance and availability throughout the course of this project. It was their belief and mentorship that kept us as a team and maintained pace in the completion of the project.

We would also like to thank some seniors, faculty members and the lab instructors for their respective support that led to the successful accomplishment of this project.

Group members	**Signature**
Sanskar Sharma	…………...
Pratiksha Sabale	…………...
Nakul Aggarwal	…………...

ABSTRACT

In recent years, mainly with the rise of social media, fake news has become a society problem, on some occasions spreading more and faster than the true information. Recent political events have led to an increase in the popularity and spread of fake news. As demonstrated by the widespread effects of the large onset of fake news, humans are inconsistent if not outright poor detectors of fake news. With this, efforts have been made to automate the process of fake news detection. The most popular of such attempts include **''blacklists''** of sources and authors that are unreliable. While these tools are useful, in order to create a more complete end to end solution, we need to account for more difficult cases . As such, the goal of this project was to create a tool for detecting the language patterns that characterize fake and real news through the use of machine learning. The results of this project demonstrate the ability for machine learning to be useful in this task. We have built a model that catches many intuitive indications of real and fake news.

There are many SCAMMERS on the internet trying to get people's information. They do all kinds of things with it. Don't provide information via the internet until after you research the company and contact them to ensure they posted the ad and there is an actual job opening. A best practice is to mail a resume after you verify the company is real. If they use an ATS (applicant tracking system) then will call you to "apply for the job" AFTER they look at your resume. Each year, the FTC takes a hard look at the number of reports people make to the Consumer Sentinel Network. In fact, during 2019, they got more than 3.2 million reports to the FTC. This project solely concerns the news of fake job postings that have been frauding the customers globally. The digitalization has increased the online job opportunities which invites the interns and graduates who'd definitely share their personal details including banking details as in reference for the job they have come across online. How can he rely whether the job being offered or posted is valid or not? This might be a fraudulent case? Such suspicion definitely leaves a factor of doubt.

LIST OF FIGURES

Figures	Page number
Figure 1 Block diagram	16
Figure 2 Use Case diagram	17
Figure 3 Random forest classifier	18
Figure 4 Entropy	19
Figure 5 Correlation Heat Map	25
Figure 6 AUC(Area under curve) diagram	26
Figure 7 Flowchart	27
Figure 8 % Fraud job posting news grouped by Education req. in jobs	28
Figure 9 % Fraud job posting news grouped by Education experience in jobs	29
Figure 10 % Fraud job posting news grouped by type of employment.	31
Figure 11 Percentage of fraud news in top 11 locations with most data entries	33
Figure 12 using default parameters to build random forest models	34
Figure 13 using the best parameters found from RandomizedSearchCV	34
Figure 14 using the best parameters found from GridSearchCV	35
Figure 14 Feature importance diagram	35
Figure 15 GUI (for user's interaction)	36

LIST OF TABLES

Tables	Page number
Table 1 Features and their datatypes	24
Table 2 count of fraud and not fraud news with req education	28
Table 3 count of fraud and not fraud news with req experience wise	29
Table 4 count of fraud and not fraud news with employment wise	31
Table 5 Output table 1	37
Table 6 Output table 2	37

CONTENTS

Acknowledgements			4
Abstract			5
List of Figures			6
List of Tables			7
1.	Introduction		9-10
	1.1	Motivation for the project	9
	1.2	Problem Statement	10
	1.3	Objectives and Scope	10
	1.4	Organization of the report	10
2.	Literature Survey		11-15
3.	System Design		16-23
	3.1	Block diagram/ Proposed System setup	16
	3.2	Use Case Diagram (If Applicable)	17
	3.3	Related mathematical modelling	18-22
	3.4	Hardware and Software Requirements	23
4.	Implementation and Results		24-39
	4.1 Algorithm and flowcharts (If applicable)		24-27
	4.2 Results		28-37
	4.3 Discussion		38-39
5.	Conclusion and Future scope		40-41
References			42

1. INTRODUCTION

These days' fake news is creating different issues from sarcastic articles to fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is "fake news " but lately blathering social media's discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints. The term 'fake news' became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views. So we came up with an idea of making fake news detection systems to avoid such problems.

Fake news is a phenomenon which is having a significant impact on Our social life, in particular in the political world. Fake news detection is an Emerging research area which is gaining interest but involves some challenges Due to the limited amount of resources i.e., Datasets, published literature) Available. Popular methods are used to detect fake news : Naïve Bayes, Logistic regression, Decision tree, random forest classifier. Social and psychological factors play an important role in fake news gaining public trust and further facilitating the spread of fake news. .For instance ,humans have been proven to be irrational and vulnerable when differentiating between truth and falsehood while overloaded with deceptive information. Studies in social psychology and communications have demonstrated that human ability to detect deception is only slightly better than chance :typical accuracy rates are in the 55%-58%range ,with a mean accuracy of 54%over 1,000 participants in over 100 experiments.

The current project involves utilizing machine learning techniques to create a model that can expose documents that are, with high probability, fake news articles. . We also cautiously provide a clear broad and narrow definition for fake news in view of the current available resources and public concerns, respectively giving the minimum and overall requirements for some information to be fake news.

Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive readers". Social media and news outlets publish fake news to increase readership or as part of psychological warfare. In general, the goal is profiting through clickbait. Clickbait lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This exposition analyzes the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts.

1.1) Motivation for the project

Sometimes politicians and professional journalists even quote fake news stories. It is observed that most of the time the fake news gets more views than real news. Due to these fake news people engage in illegal and violent behavior as a result of believing a fake news story. Media literacy was not taught at many schools now and people are gullible and can't distinguish real news from fake as shown by the studies and they view the media as a moonlit.

1.2) Problem Statement

Based on various types of heterogeneous information sources, including both textual contents/profile/descriptions and the authorship and article subject relationships among them, we aim at identifying fake news. We formulate the fake news detection problem as a credibility inference problem, where the real ones will have a higher credibility while unauthentic ones will have a lower one instead. The problem of detecting not-genuine sources of information through content based analysis is considered solvable at least in the domain of spam detection , spam detection utilizes statistical machine learning techniques to classify text (i.e. tweets or emails) as spam or legitimate. These techniques involve pre-processing of the text, feature extraction (i.e. bag of words), and feature selection based on which features lead to the best performance on a test dataset.

1.3) Objectives and Scope

The main goal of this project was the creation of a visualization tool for classification of fake and real news. The goal of this project has been to comprehensively and extensively review, summarize, compare and evaluate the current research on fake news ,which includes the qualitative and quantitative analysis of fake news as well as detection and intervention strategies for fake news from four perspectives: the false knowledge fake news communicates, its writing style ,its propagation patterns ,and its credibility, main fake news characteristics (authenticity, intention, and being news) that allow distinguishing it from other related concepts (e.g., misinformation, disinformation, or rumors), various news-related(e.g. ,head line ,body-text, creator, and publisher)and social-related(e.g., comments, propagation paths and spreaders) information that can be exploited to study fake news across its lifespan (being created, published ,or propagated), feature-based and relation-based techniques for studying fake news and available resources ,e.g. ,fundamental theories, websites, tools, and platforms, to support fake news studies. To achieve 95-100% accuracy in data prediction as fake or real. Proper analysis of the data set and finding out the grains that would help us understand a pattern or trend of fake or real news.

Our system considers the features available or can be derived from the example data set only though further features can be added later as and when required. Not 100% accurate result. Accuracy of above 85% is achieved.

1.4) Organization of report

1. Jan 2020 : selection of problem statement , completed literature survey
2. Feb 2020 : completed system design by the help of UML DIAGRAMS , worked on methodology
3. March 2020 : wrapped up with the pre documentation work of the project (assignments)
4. April 2020 : completed with the back end implementation of the project
5. May 2020 : implementation (both front end and back end)was completed along with final project report and presentation.

2. LITERATURE SURVEY

Literature survey 1:

FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network

Published on: 10^{th} August, 2019

Authors:

- Jiawei Zhang (IFM Lab, Department of Computer Science, Florida State University, FL, USA)
- Bowen Dong, Philip S. Yu (BDSC Lab, Department of Computer Science, University of Illinois at Chicago, IL, USA)

Summary:

Fake news denotes intentionally presents misinformation or hoaxes spreading through both traditional print news media and recent online social media. Fake news has been existing for a long time, since the "**Great moon hoax**" was published in 1835.

In recent years, due to the booming developments of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. With deceptive words, online social network users can get infected by these online fake news easily, which has brought about tremendous effects on the offline society already

This paper includes the study of fake news detection (like articles, creators and subjects) problems in online social networks. Based on various types of heterogeneous information sources, including both textual contents/profile/descriptions and the authorship and article subject relationships among them, it aims at identifying fake news from the online social networks simultaneously. This paper formulates the fake news detection problem as a credibility inference problem, where the real ones will have a higher credibility whereas the unauthenticated ones will have a lower one instead. This work is also supported in part by NSF through grants IIS-1763365 and IIS-1763325.

Based on the news from heterogeneous social networking sites, a set of explicit and latent features capable enough to classify news as fake or real can be extracted from the textual information of news articles, creators and subjects respectively. Furthermore, based on the connections among news articles, creators and news subjects, a deep diffusive network model has also been proposed to incorporate the network structure information into model learning. This paper also introduces a new diffusive unit model, namely **GDU**. Model GDU accepts multiple inputs from different sources simultaneously, and can effectively fuse these input for output generation with content **"forget"** and **"adjust" gates**.

Literature survey 2:

Machine Learning for Detection of Fake News

Published on: 17^{th} May, 2018

Authors:

- Nicole O'Brien (Master of Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology)

Summary:

Recent political events has led to an increase in the popularity and spread of fake news. As demonstrated by the widespread effects of the large onset of fake news, humans are inconsistent if not outright poor detectors of fake news. With this, efforts have been made to automate the process of fake news detection. The most popular of such attempts include "**blacklists**" of sources and authors that are unreliable.

The goal of this research paper was to create a tool for detecting the language patterns that characterize fake and real news through the use of machine learning and natural language processing techniques.

The main contribution of this paper is support for the idea that **machine learning** could be useful in a novel way for the task of classifying fake news. Its findings show that after much pre-processing of a relatively small dataset, a simple **CNN** is able to pick up on a diverse set of potentially subtle language patterns that a human may (or may not) be able to detect. Many of these language patterns are intuitively useful in a human manner of classifying fake news. Some such intuitive patterns that our model has found to indicate fake news include generalizations, colloquialisms and exaggerations.

Other contributions of this paper includes the creation of a dataset for the task and the creation of an application that aids in the visualization and understanding of the neural nets classification of a given body text. It could also be useful in researchers trying to develop improved models through the use of improved and enlarged datasets, different parameters, etc.

Literature survey 3:

Fake News: A Survey of Research, Detection Methods, and Opportunities

Published on: 2nd December, 2018

Authors:

- XINYI ZHOU (Syracuse University, USA)
- REZA ZAFARANI (Syracuse University, USA)

Summary:

The rise of fake news during the 2016 U.S. Presidential Election highlighted not only the dangers of the effects of fake news but also the challenges presented when attempting to separate fake news from real news. Recent political events have led to an increase in the popularity and spread of fake news. As demonstrated by the widespread effects of the large onset of fake news, humans are inconsistent if not outright poor detectors of fake news. These days' fake news is creating different issues from sarcastic articles to fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is "fake news " but lately blathering social media's discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints. Fake news is one of the biggest scourges in our digitally connected world. That is no exaggeration. It is no longer limited to little squabbles – fake news spreads like wildfire and is impacting millions of people every day.

"Fake news (also known as junk news, pseudo-news, or hoax news) is a form of news consisting of deliberate disinformation or hoaxes spread via traditional news media (print and broadcast) or online social media."

An Overview of this Survey, This survey aims to present a comprehensive framework to study fake news. Fake news can be studied with respect to four perspectives:

1. knowledge-based
2. style-based
3. credibility-based

This survey compares several fake-news related terms and concepts. Besides this it also provides a definition for fake news. This survey provides the most comprehensive list of fundamental theories that can be utilized when studying fake news.

The goal of this survey has been to comprehensively and extensively review , summarize ,compare and evaluate the current research on fake news. The open issues and challenges are also presented in this survey with potential research tasks that can facilitate further development in fake news research.

Literature survey 4:

Fake News Detection Using Machine Learning

Published on: 05 April 19

Authors:

- Lilapati Waikhoma (Department of Computer Science & Engineering, NIT, Arunachal Pradesh, India)
- Rajat Subhra Goswami (Department of Computer Science & Engineering, NIT, Arunachal Pradesh, India)

Summary:

The Internet has become compulsory in our life. It is now very easy to access the Internet than it was before. There is no doubt that many young people prefer the internet to get their news rather than the newspaper, radio, etc. The Internet provides many opportunities for us, we can search for anything on the internet to clear our doubts and for research purposes also. Simply saying, we can't even imagine our life without the internet. In a diverse country like India where Internet access has become cheap as compared to the past decade , a lot of people have a convenient access of news through their digital devices relevant to the field of interest If it is about the news, the internet plays a very important role because through the internet, the news widespread very fast. There are so many consequences of fake news, it can cause harm to innocent people. Fake news may be made intentionally or accidentally to give harms to an individual or a group for any purposes, such as for political issues, for religious purposes and so on.

Automatic fake news detection has already been studied for some years. Rubin, et.al in his book along with N.J Conroy and Y. Chen titled "Automatic deception detection: Methods for finding fake news" gave a hybrid approach which combines the linguistic features of a language with the network analysis approach which may not be always suitable as the network information may be restricted or not available. Majority of the datasets available contain short statements as the language used for political information broadcasting on TV interviews, social media posts and tweets which are mostly short length statements, that's why the detection of fake news is more challenging. Following are the important methodologies that play a crucial role in the making of fake news detection algorithm :

- Textual features extraction(includes bag of words concept , N-grams, TFIDF[term frequency inverse document frequency])
- Categorial features (include hot encoding and label encoding)
- Numerical features (including the calling and normalization)

In this paper, we present the task of automatic detection of fake news. We have used a new publicly available fake news dataset, the LIAR-dataset. The classification of fake news from the real news is a very crucial task nowadays. Our best performing models achieved accuracies that are comparable to the human ability to spot fake content.

Literature survey 5:

Fake News Detection using Machine Learning and Natural Language Processing

Published on: March 25, 2019

Authors:

- Kushal Agarwalla (Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.)
- Shubham Nandan (Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.)
- Shubham Nandan (Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.)
- D. Deva Hema (Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.)

Summary:

Modern life has become quite suitable and the people of the world have to thank the vast contribution of the internet technology for transmission and information sharing. There is no doubt that the internet has made our lives easier and access to surplus information viable.

But at the same time it unfocussed the line between true media and maliciously forged media. Today anyone can publish content – credible or not – that can be consumed by the world wide web. Sadly, fake news accumulates a great deal of attention over the internet, especially on social media. People get deceived and don't think twice before circulating such mis-informative pieces to the far end of the world. This kind of news vanishes but not without doing the harm it intended to cause. Various models are used to provide an accuracy range of 60-75%. Which comprises Naïve Bayes classifier. Linguistic features based, Bounded decision tree model, SVM etc. The parameters that are taken in consideration do not yield high accuracy. The motive of the following paper tends to increase the accuracy of detecting fake news more than the present results that are available.

The following were the relational models that are found useful for making of the algorithm :

1. Logistic regression: The LR model uses gradient descent to converge onto the optimal set of weights (θ) for the training set.
2. Support vector machine : A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression purposes. SVMs are mostly used in classification problems.
3. Naïve Bayes Classification with Lid stone smoothing :In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with powerful (naive) independent assumptions between the features. A lot of our results circle back to the need for acquiring more accuracy. Generally speaking, simple algorithms perform better on less (less variant) data. Since we had a huge set of data, SVM, Naive Bayes and Logistic Regression underperformed.

3. SYSTEM DIAGRAMS

3.1) Block diagram/ Proposed System setup

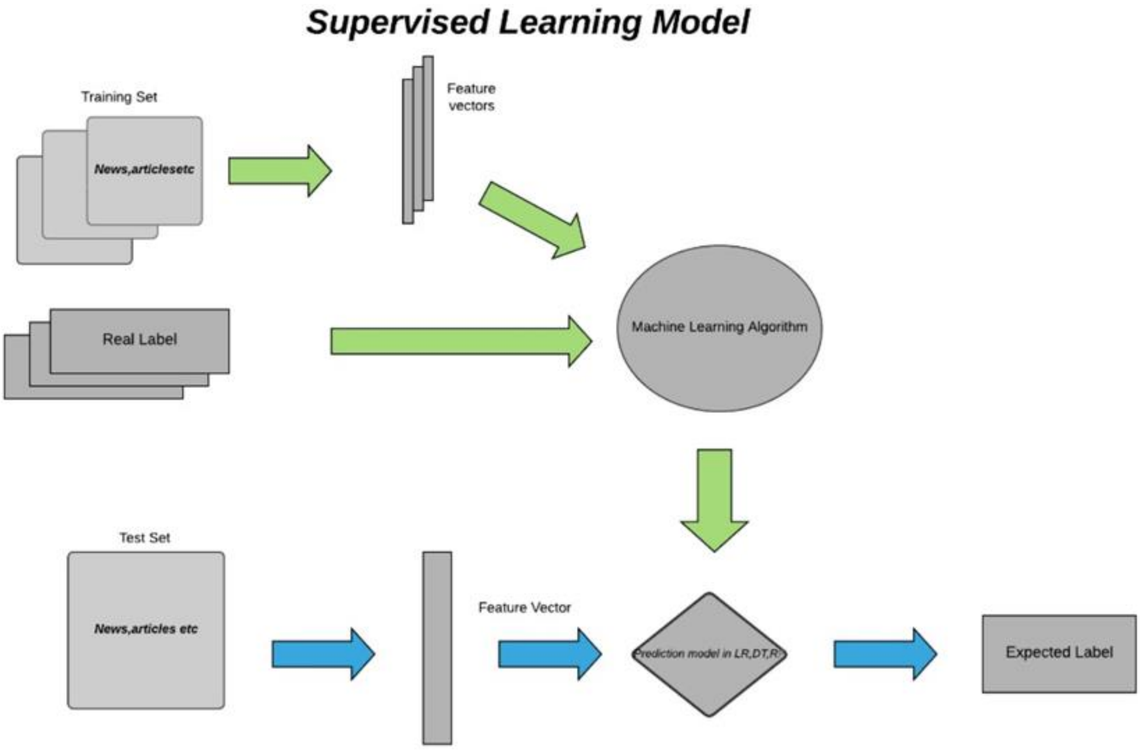


Figure 16 Block diagram

3.2) Use Case Diagram

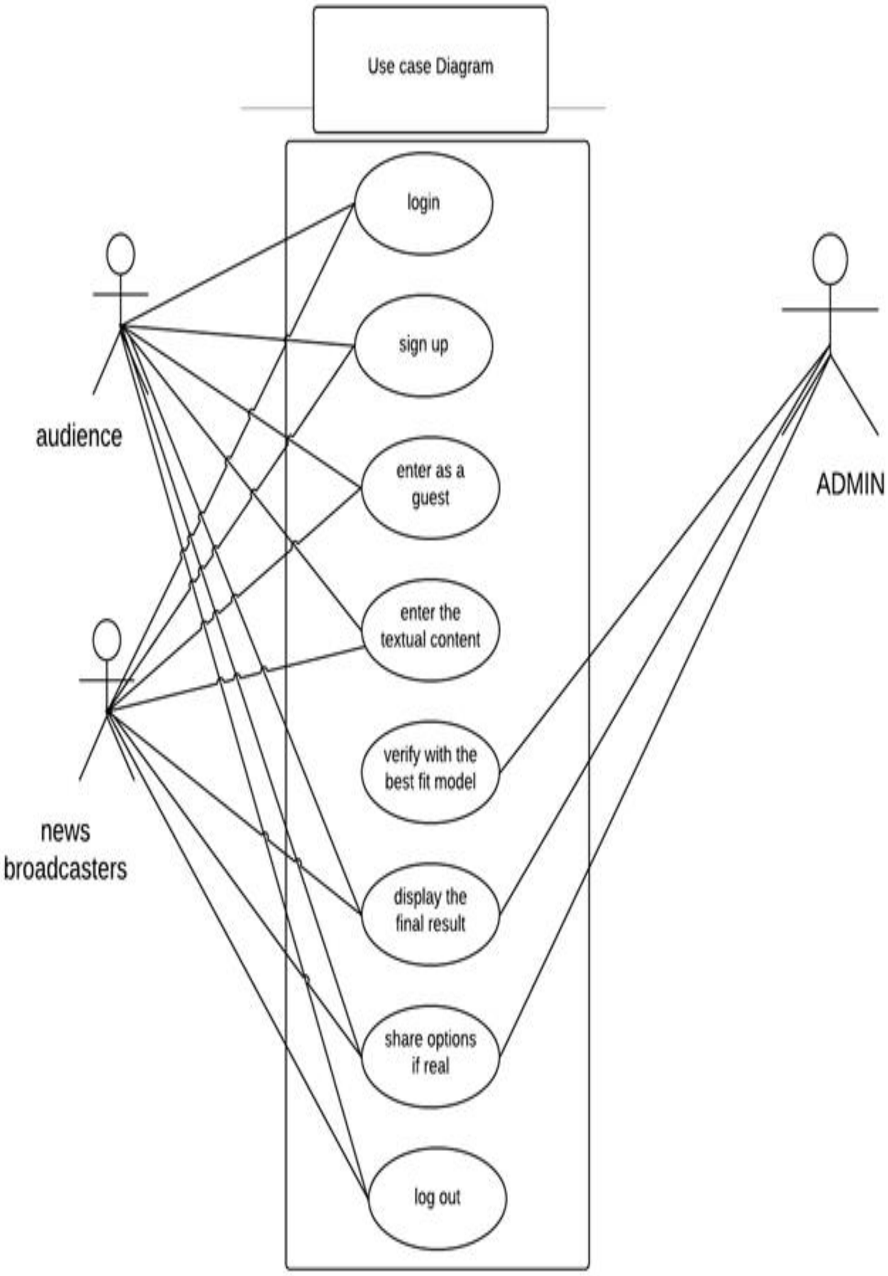


Figure 17 Use Case diagram

3.3) Related mathematical modelling

Random Forest:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction

Bagging (Bootstrap Aggregation) :

Decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures. Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees. This process is known as bagging.

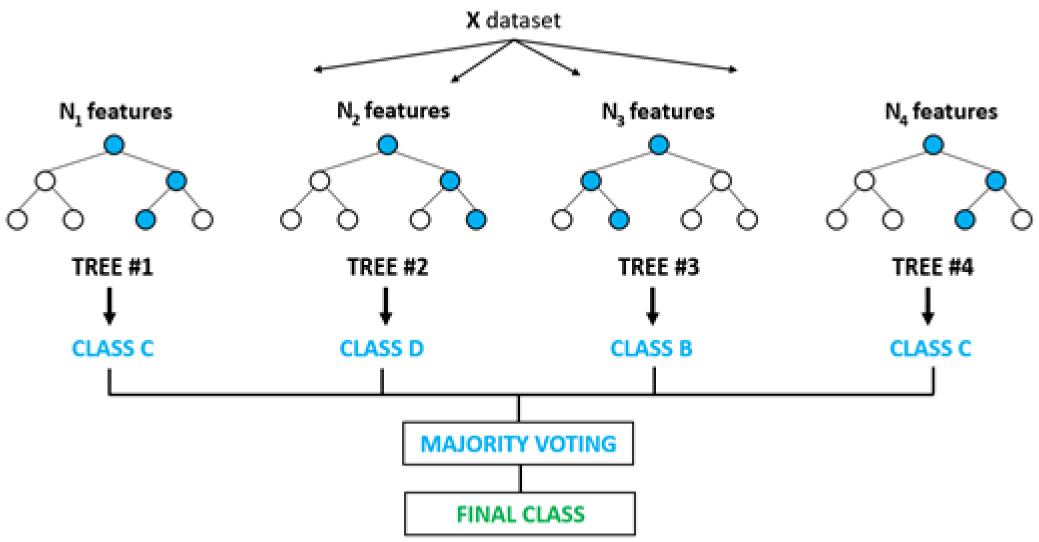


Figure 18 Random forest classifier

Decision Tree algorithm

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter. (Algorithms are ID3, gini etc)

Decision Tree consists of :

Nodes : Test for the value of a certain attribute.

Edges/ Branch : Correspond to the outcome of a test and connect to the next node or leaf.

Leaf nodes : Terminal nodes that predict the outcome (represent class labels or class distribution).

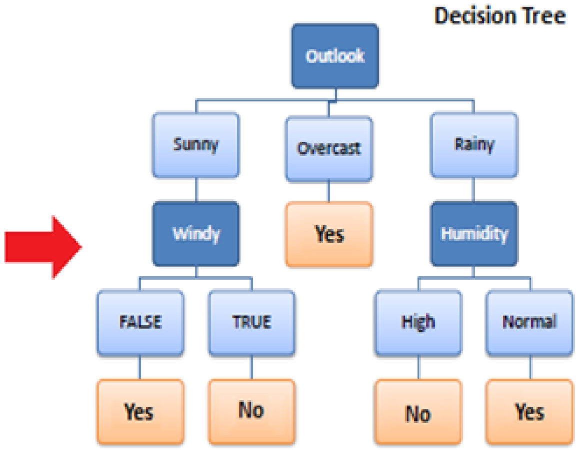
There are two main types of Decision Trees:

- Classification Trees. (Entropy and Information Gain method)
- **Regression Trees. (Standard Deviation Reduction method)**

SDR (Standard Deviation Reduction for classification)

Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

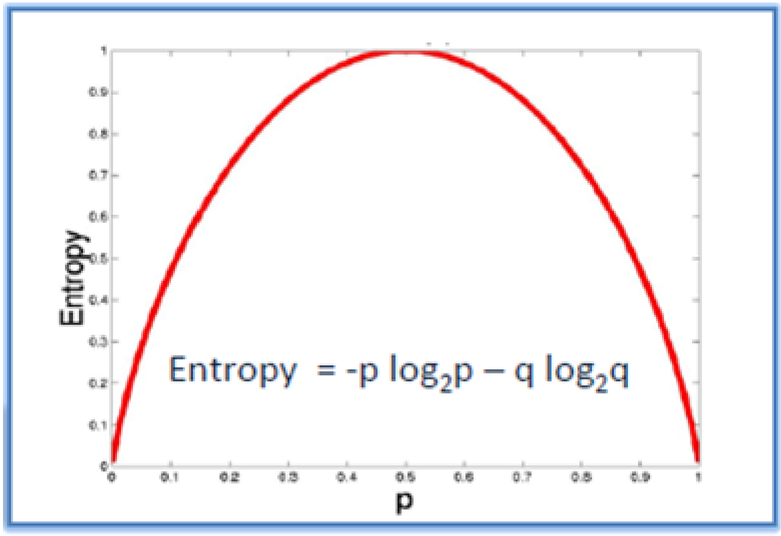


Algorithm:

The core algorithm for building decision trees called **ID3** by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses *Entropy* and *Information Gain* to construct a decision tree.

Entropy:

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is equally divided it has entropy of one.



$$\text{Entropy} = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$$

Figure 19 Entropy

To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

1. Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

2. Entropy using the frequency table of two attributes:

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = P(Sunny)*E(3,2) + P(Overcast)*E(4,0) + P(Rainy)*E(2,3)
= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971
= 0.693

Information Gain:

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding the attribute that returns the highest information gain (i.e., the most homogeneous branches).

- Calculate entropy of the target.
- The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

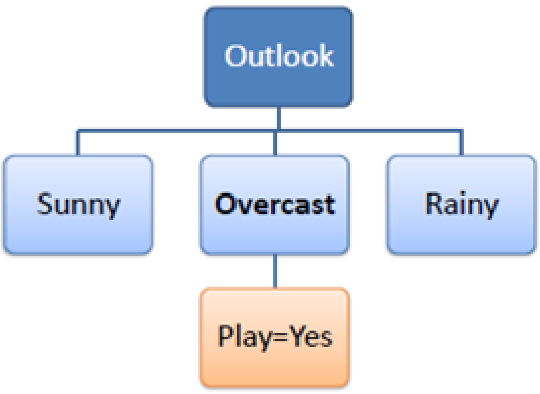
= 0.940 – 0.693 = 0.247

- Choose the attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

		Play Golf	
★		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

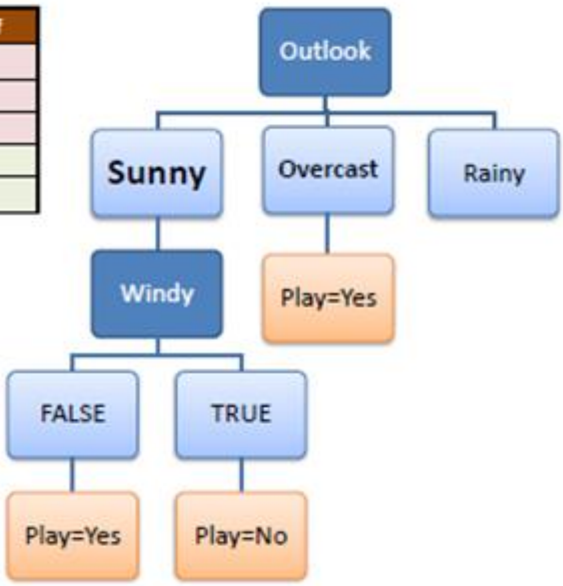
- A branch with entropy of 0 is a leaf node.

Temp	Humidity	Windy	Play Golf
Hot	High	FALSE	Yes
Cool	Normal	TRUE	Yes
Mild	High	TRUE	Yes
Hot	Normal	FALSE	Yes



- A branch with entropy more than 0 needs further splitting.

Temp	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No



- The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

Decision Tree to Decision Rules

A decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one.

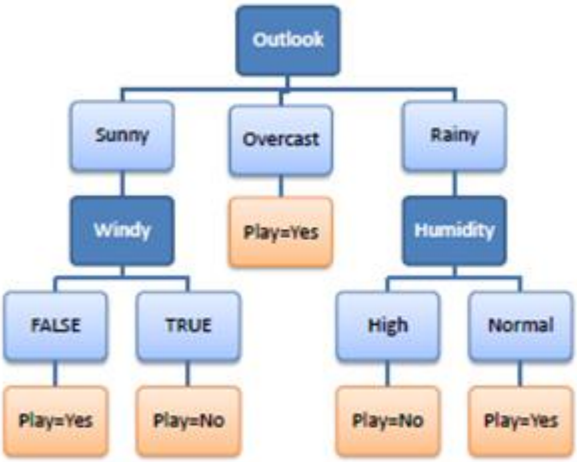
R_1: IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

R_2: IF (Outlook=Sunny) AND (Windy=TRUE) THEN Play=No

R_3: IF (Outlook=Overcast) THEN Play=Yes

R_4: IF (Outlook=Rainy) AND (Humidity=High) THEN Play=No

R_5: IF (Outlook=Rain) AND (Humidity=Normal) THEN Play=Yes



3.4) Hardware and Software Requirements

No hardware resources required.

Software resources required are:

- **Anaconda**

  Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS.

- **Jupyter notebook**

  Jupyter Notebook is a web-based interactive computational environment for creating Jupyter notebook documents. The "notebook" term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context. A Jupyter Notebook document is a JSON document, following a versioned schema, and containing an ordered list of input/output cells which can contain code, text mathematics, plots and rich media, usually ending with the ".ipynb" extension. A Jupyter Notebook can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library or "jupyter nbconvert" command line interface in a shell.

- **Python (language used)**

  Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured(particularly, procedural), object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library. Python interpreters are available for many operating systems.

Some of the main modules used are as follows:

Numpy, Matplotlib, Seaborn, Pandas, Scikit-learn, Scipy, Sklearn, openpyxl, tkinter etc.

4. IMPLEMENTATION AND RESULTS

4.1) Algorithm and flowcharts

Algorithm

1. Data collection
2. Sort the data (tabular form)
3. Input the data (.csv format)
4. Import **pandas** and **numpy**

```
import pandas as pd
import numpy as np
import csv
Jobpostings = pd.read_csv('F:/Zulu/My Btech/Semester 4/Minor/fake_job_postings.csv',encoding='utf8' )
Jobpostings.head()
```

5. Proceed with data analysis and cleaning
6. Import matplotlib and seaborn
7. Provide all necessary data visualisations.
8. Derive conclusions on labels of various job posting news.
9. Convert **object columns** to **int64** or **float**.

Table 6 Features and their datatypes

```
----------------------All the features of Merged Main Table and their datatypes----------------------
17533 x 14
job_id                 int64
title                 object
location              object
department            object
company_profile       object
description           object
requirements          object
benefits              object
has_company_logo       int64
required_experience   object
required_education    object
industry              object
function              object
fraudulent             int64
dtype: object
job_id                int64
title                 int64
location              int64
department            int64
company_profile       int64
description           int64
requirements          int64
benefits              int64
has_company_logo      int64
required_experience   int64
required_education    int64
industry              int64
function              int64
fraudulent            int64
dtype: object
* features:
['job_id', 'title', 'location', 'department', 'company_profile', 'description', 'requirements', 'benefits', 'has_comp
any_logo', 'required_experience', 'required_education', 'industry', 'function']
```

10. Split the attributes of the data set as **features** & **targets.**

```
#first n-1 col as features, and the last one as target
df1=df.iloc[:,0:n]
features = list(df1.columns[:(n-1)])
print("* features:", features, sep="\n")
df1.rename(columns={'fraudulent':'Target'}, inplace=True)
list(df1)
df1

y = df1["Target"]
X = df1[features]
```

11. Create the heatmap using seaborn library to better visualise correlation between various features.

```
------------------------------------------ Correlation Matrix ------------------------------------------
Co-relation between features
(greenish implies low significance while the reddish implies high significance)
SEE Between target values and others
```

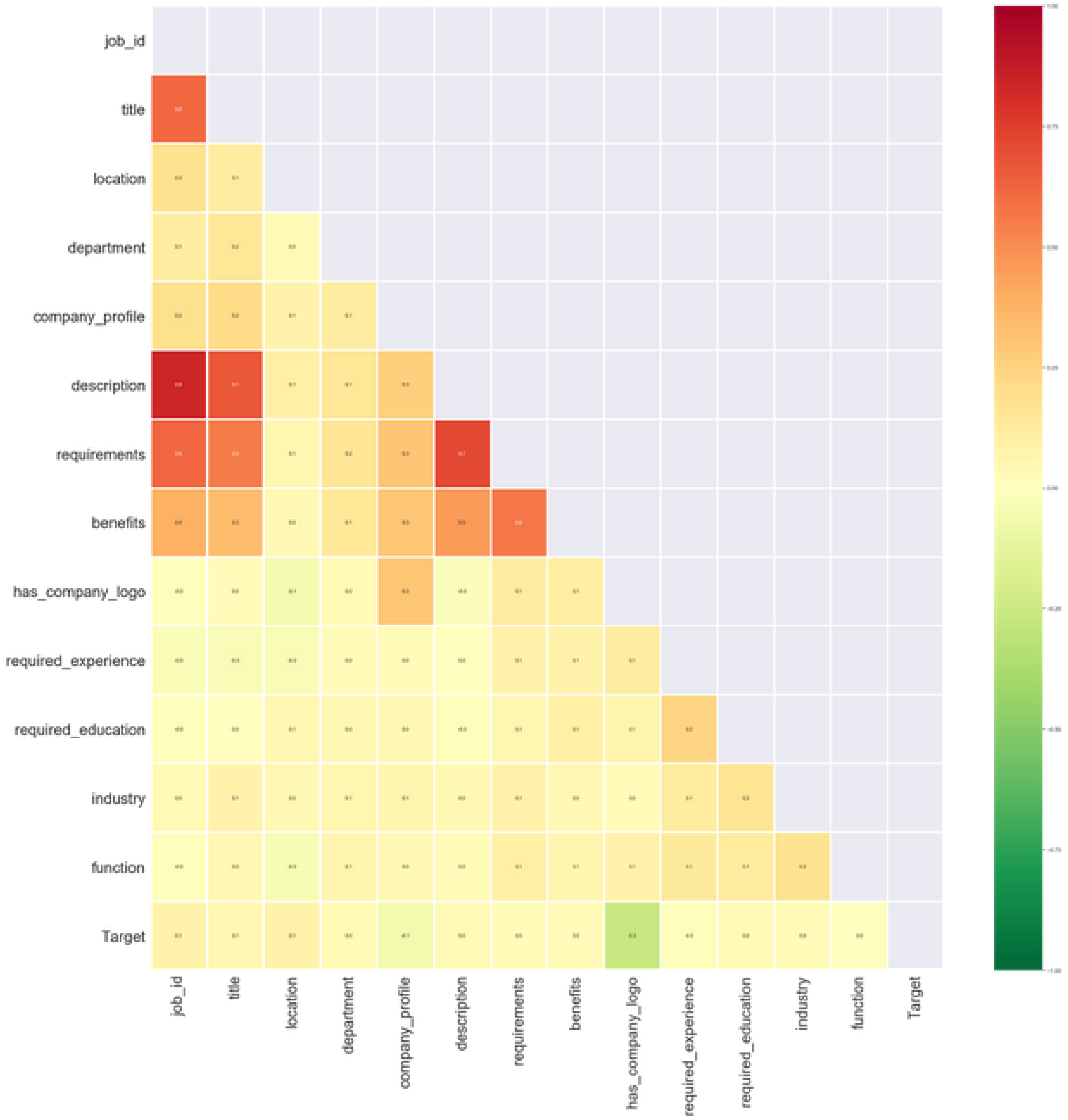


Figure 20 Correlation Heat Map

12. Construct a model to predict credibility of news.
13. Import **scipy** and **sklearn.**
14. Import **RandomizedSearchCV** and **GridSearchCV** from sklearn
15. **Random forest algorithm** with hyperparameter tuning.
16. Compare model’s accuracy of three models

    1. using **default** parameters to build random forest models.

    2. use the **best parameters** found from **RandomizedSearchCV.**

    3. use the **best parameters** from **GridSearchCV.**

17. Plot **AUC(Area under curve)** for all the three models to better visualise.

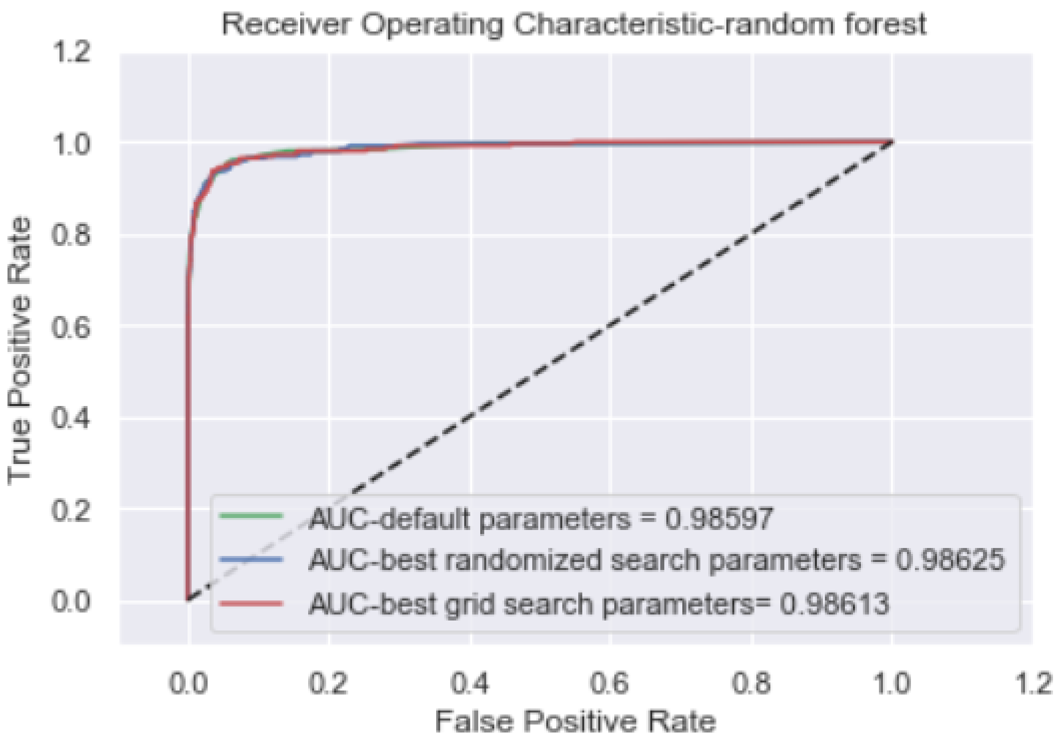


Figure 21 AUC(Area under curve) diagram

18. Print the feature ranking graph to obtain significant and insignificant features.
19. Import **openpyxl** and **tkinter.**
20. Develop **GUI** application for user input.
21. Print the output of user inputs as fraudulent or not fraudulent.

Flowchart

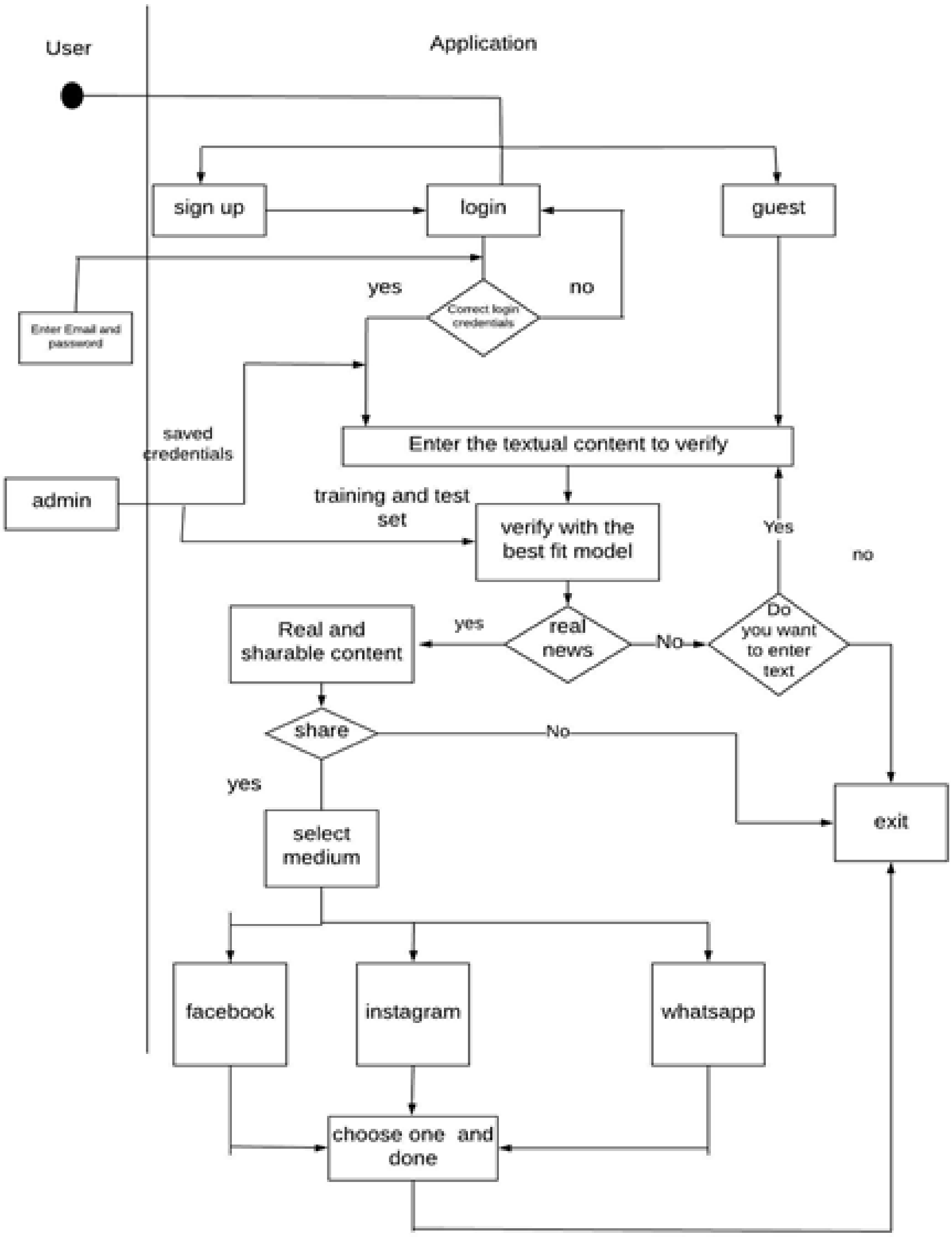


Figure 22 Flowchart

4.2) Results

Some analysis and visualization results

Required Education asked in News

Table 7 count of fraud and not fraud news with req education

```
Number of not fraud and fraud customers with required_education wise:
fraudulent  required_education
0           0                                    7385
            Bachelor's Degree                    5014
            High School or equivalent            1898
            Unspecified                          1332
            Master's Degree                       379
            Associate Degree                      266
            Certification                         150
            Some College Coursework Completed      98
            Professional                           70
            Vocational                             48
            Doctorate                              25
            Vocational - HS Diploma                 9
            Some High School Coursework             7
            Vocational - Degree                     6
1           0                                     435
            High School or equivalent             169
            Bachelor's Degree                      98
            Unspecified                            60
            Master's Degree                        31
            Some High School Coursework            20
            Certification                          19
            Associate Degree                        6
            Professional                            4
            Some College Coursework Completed       3
            Doctorate                               1
Name: required_education, dtype: int64
```

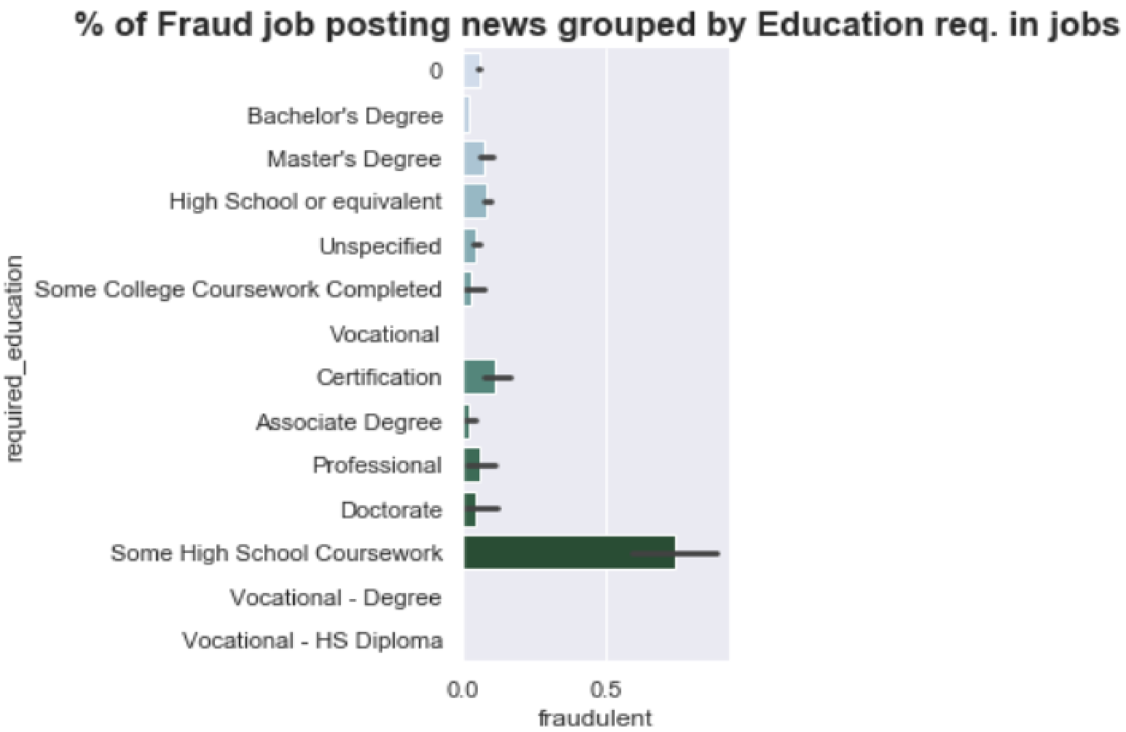


Figure 23 % Fraud job posting news grouped by Education req. in jobs

Required experience asked in News

Table 8 count of fraud and not fraud news with req experience wise

```
Number of not fraud and fraud customers with required_experience wise:
fraudulent  required_experience
0           0                      6377
            Mid-Senior level       3663
            Entry level            2508
            Associate              2241
            Not Applicable         1040
            Director                369
            Internship              358
            Executive               131
1           0                       419
            Entry level             177
            Mid-Senior level        113
            Not Applicable           60
            Associate                41
            Director                 17
            Internship               10
            Executive                 9
Name: required_experience, dtype: int64
```

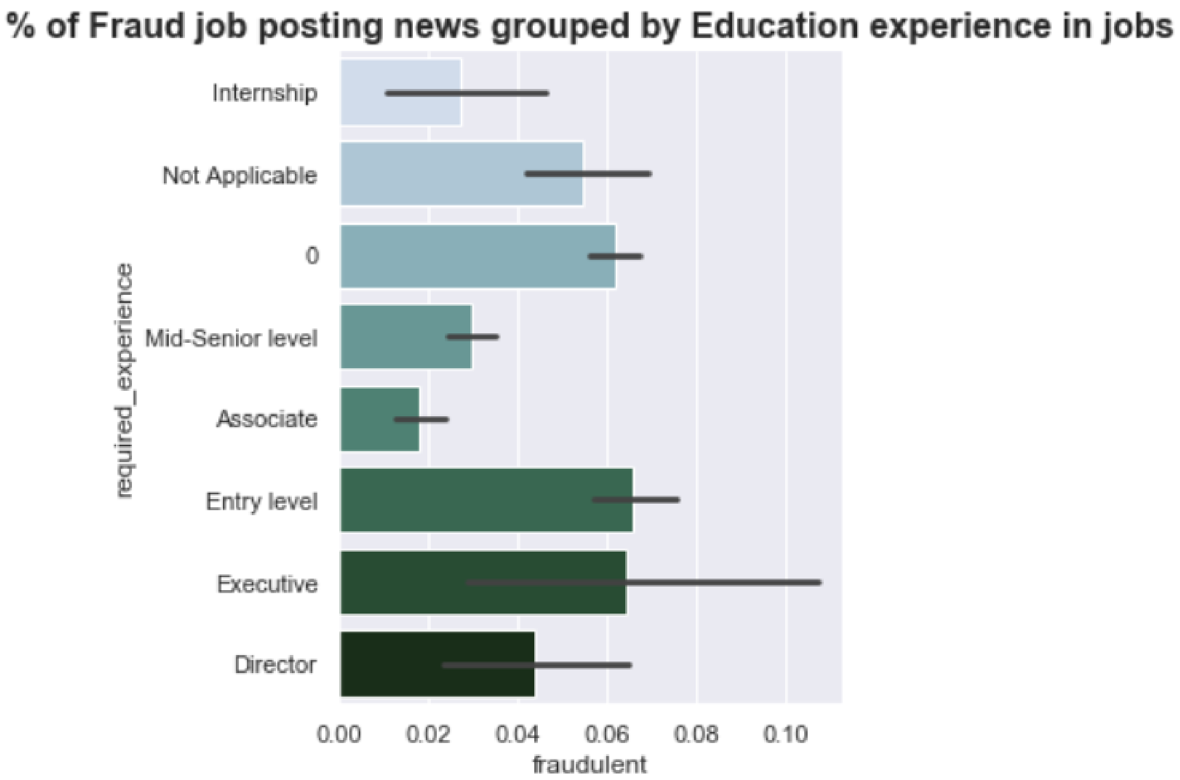
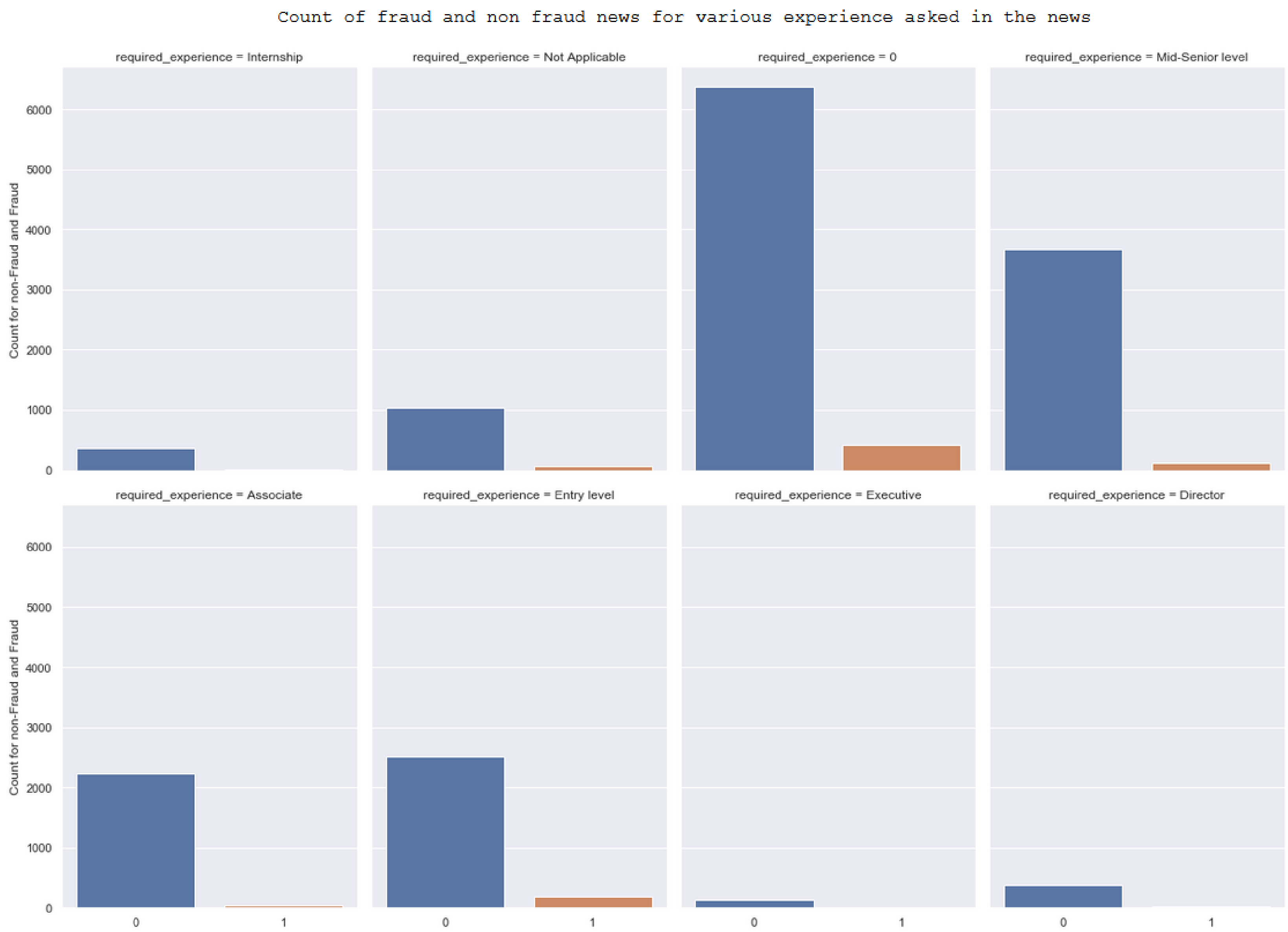


Figure 24 % Fraud job posting news grouped by Education experience in jobs


Count of fraud and non fraud news for various experience asked in the news
required_experience = Internship
required_experience = Not Applicable
required_experience = 0
required_experience = Mid-Senior level
required_experience = Associate
required_experience = Entry level
required_experience = Executive
required_experience = Director
Count for non-Fraud and Fraud
6000
5000
4000
3000
2000
1000
0
0
1

Employment type asked in News

Table 9 count of fraud and not fraud news with employment wise

```
Number of not fraud and fraud customers with employment_type wise:
fraudulent  employment_type
0           Full-time          11039
            0                   3028
            Contract            1470
            Part-time            709
            Temporary            237
            Other                204
1           Full-time            485
            0                    228
            Part-time             74
            Contract              42
            Other                 15
            Temporary              2
Name: employment_type, dtype: int64
```

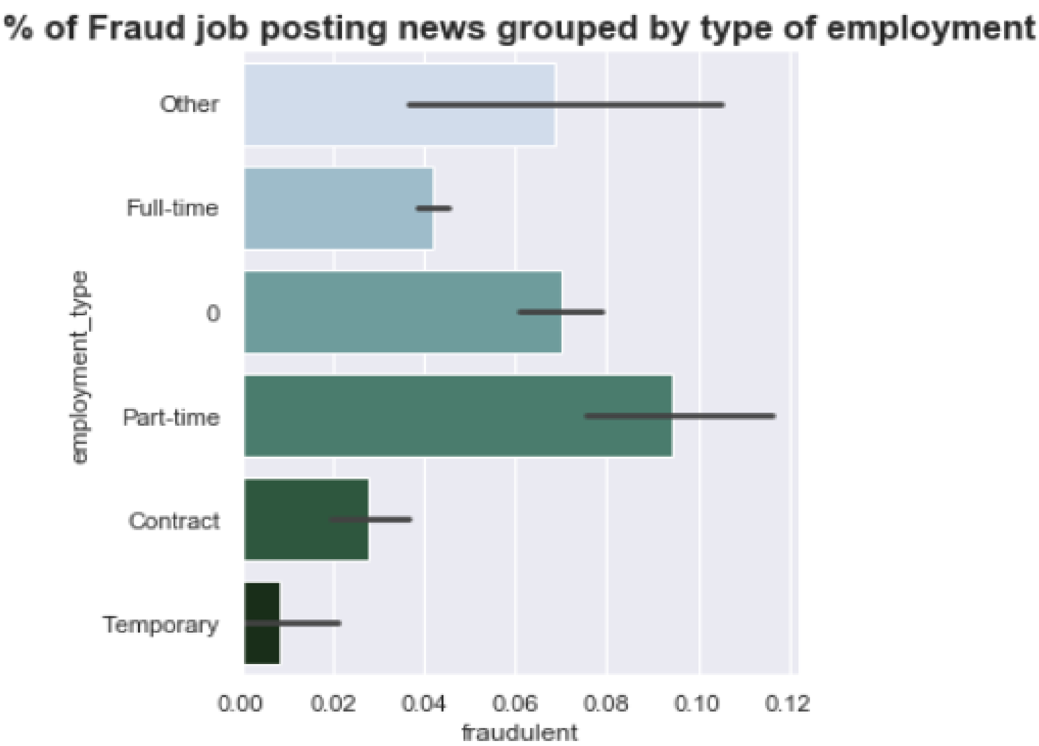
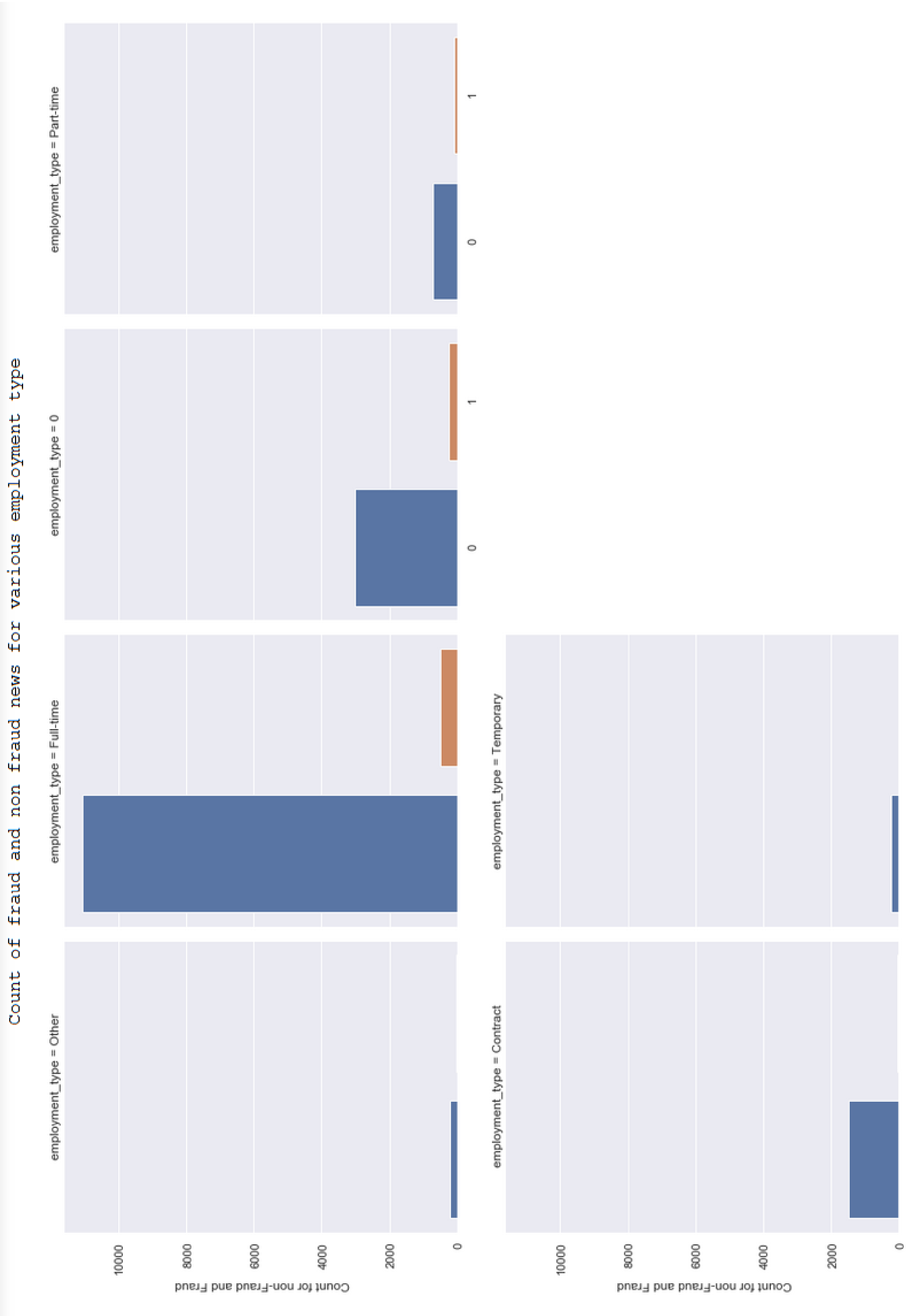


Figure 25 % Fraud job posting news grouped by type of employment.


Count of fraud and non fraud news for various employment type
employment_type = Other
employment_type = Full-time
employment_type = 0
employment_type = Part-time
employment_type = Contract
employment_type = Temporary
Count for non-Fraud and Fraud
10000
8000
6000
4000
2000
0
1

Percentage of fraud news in top 11 locations with most data entries

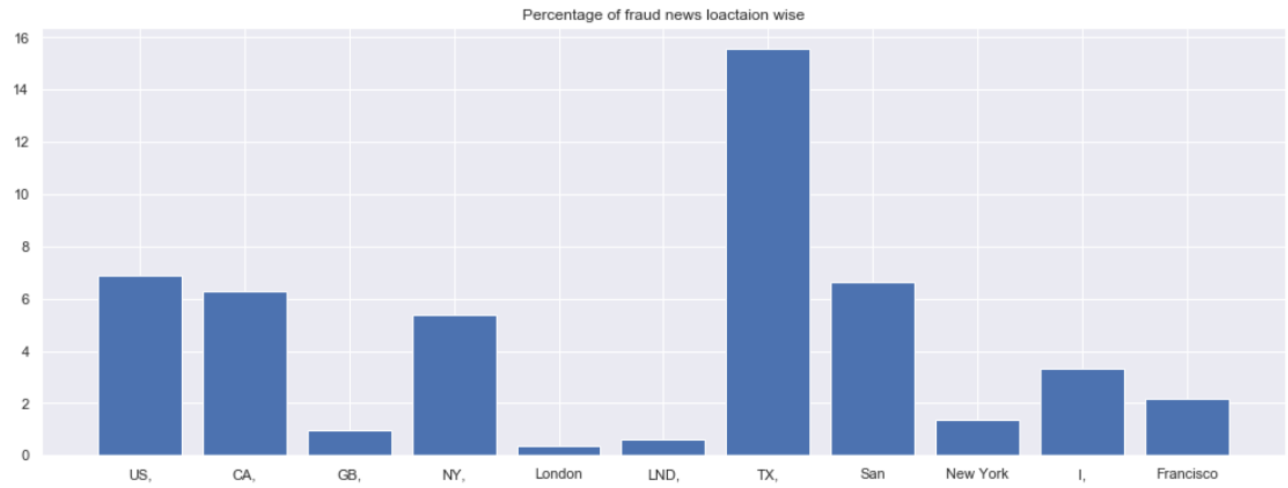


Figure 26 Percentage of fraud news in top 11 locations with most data entries

Random forest models

```
Automatically created module for IPython interactive environment
----------------------Best Parameter search using Random and Grid Search----------------------

C:\Users\hp\anaconda3\lib\site-packages\sklearn\model_selection\_search.py:823: FutureWarning: The parameter 'iid' is
deprecated in 0.22 and will be removed in 0.24.
  "removed in 0.24.", FutureWarning

RandomizedSearchCV took 5.82 seconds for 20 candidates parameter settings.
Model with rank: 1
Mean validation score: 0.935 (std: 0.030)
Parameters: {'bootstrap': False, 'criterion': 'entropy', 'max_depth': None, 'max_features': 9, 'min_samples_split':
4}

Model with rank: 2
Mean validation score: 0.925 (std: 0.028)
Parameters: {'bootstrap': False, 'criterion': 'entropy', 'max_depth': None, 'max_features': 7, 'min_samples_split':
2}

Model with rank: 3
Mean validation score: 0.925 (std: 0.021)
Parameters: {'bootstrap': True, 'criterion': 'gini', 'max_depth': None, 'max_features': 7, 'min_samples_split': 5}

GridSearchCV took 17.28 seconds for 72 candidate parameter settings.
Model with rank: 1
Mean validation score: 0.933 (std: 0.021)
Parameters: {'bootstrap': False, 'criterion': 'entropy', 'max_depth': None, 'max_features': 10, 'min_samples_split':
3}

Model with rank: 2
Mean validation score: 0.931 (std: 0.021)
Parameters: {'bootstrap': False, 'criterion': 'gini', 'max_depth': None, 'max_features': 10, 'min_samples_split': 2}

Model with rank: 3
Mean validation score: 0.931 (std: 0.019)
Parameters: {'bootstrap': False, 'criterion': 'gini', 'max_depth': None, 'max_features': 10, 'min_samples_split': 3}

C:\Users\hp\anaconda3\lib\site-packages\sklearn\model_selection\_search.py:823: FutureWarning: The parameter 'iid' is
deprecated in 0.22 and will be removed in 0.24.
  "removed in 0.24.", FutureWarning
```

using default parameters to build random forest models

```
In [193]: #using default parameters to build random forest model

from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier()
clf.fit(X, y)

from sklearn import datasets
from sklearn import metrics
expected = test["Target"]
X1 = test[features]
predicted1 = clf.predict(X1)

print(metrics.classification_report(expected, predicted1))
print(metrics.confusion_matrix(expected, predicted1))
```

```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5007
           1       0.99      0.62      0.77       253

    accuracy                           0.98      5260
   macro avg       0.98      0.81      0.88      5260
weighted avg       0.98      0.98      0.98      5260

[[5005    2]
 [  95  158]]
```

```
In [194]: # roc1 for default parameters
probas1_ = clf.fit(X, y).predict_proba(X1)

from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt

false_positive_rate1, true_positive_rate1, thresholds = roc_curve(expected, probas1_[:, 1])
roc_auc1 = auc(false_positive_rate1, true_positive_rate1)
roc_auc1

Out[194]: 0.9859698398526647
```

Figure 27using default parameters to build random forest models

using the best parameters found from RandomizedSearchCV

```
#use the best parameters found from RandomizedSearchCV
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(criterion='entropy', max_features=4, bootstrap= False, min_samples_split=3,
     max_depth=20, min_samples_leaf=3)
clf.fit(X, y)
from sklearn import datasets
from sklearn import metrics
expected = test["Target"]
X1 = test[features]
predicted2 = clf.predict(X1)

print(metrics.classification_report(expected, predicted2))
print(metrics.confusion_matrix(expected, predicted2))
```

```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5007
           1       0.98      0.66      0.79       253

    accuracy                           0.98      5260
   macro avg       0.98      0.83      0.89      5260
weighted avg       0.98      0.98      0.98      5260

[[5003    4]
 [  85  168]]
```

```
# roc2 for RandomizedSearchCV
probas2_ = clf.fit(X, y).predict_proba(X1)
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
false_positive_rate2, true_positive_rate2, thresholds = roc_curve(expected,probas2_[:, 1])
roc_auc2 = auc(false_positive_rate2, true_positive_rate2)
roc_auc2

0.9862508693362889
```

Figure 28using the best parameters found from RandomizedSearchCV

use the best parameters from GridSearchCV

```
#use the best parameters from GridSearchCV
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(criterion='gini', max_features=3, bootstrap= False, min_samples_split=4,
                             max_depth=20, min_samples_leaf=3)
clf.fit(X, y)

from sklearn import datasets
from sklearn import metrics
expected = test["Target"]
print(features)
X1 = test[features]
predicted3 = clf.predict(X1)

print(metrics.classification_report(expected, predicted3))
print(metrics.confusion_matrix(expected, predicted3))
#0 income<=50k
#1 income > 50k
```

```
['job_id', 'title', 'location', 'department', 'company_profile', 'description', 'requirements', 'benefits', 'has_comp
any_logo', 'required_experience', 'required_education', 'industry', 'function']
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5007
           1       0.99      0.62      0.76       253

    accuracy                           0.98      5260
   macro avg       0.98      0.81      0.87      5260
weighted avg       0.98      0.98      0.98      5260

[[5005    2]
 [  97  156]]
```

```
# roc GridSearchCV
probas3_ = clf.fit(X, y).predict_proba(X1)

from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
false_positive_rate3, true_positive_rate3, thresholds = roc_curve(expected,probas3_[:, 1])
roc_auc3 = auc(false_positive_rate3, true_positive_rate3)
roc_auc3
```

```
0.9861312739240162
```

Figure 29use the best parameters from GridSearchCV

Feature importance diagram

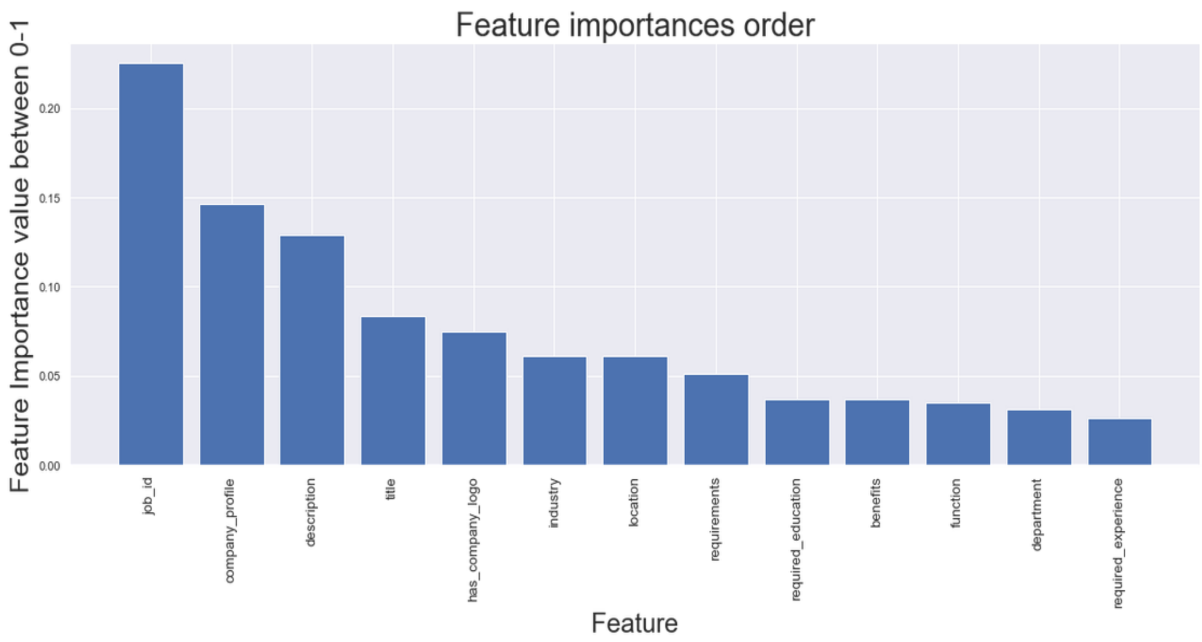


Figure 30Feature importance diagram

Output:

GUI (for user's interaction)

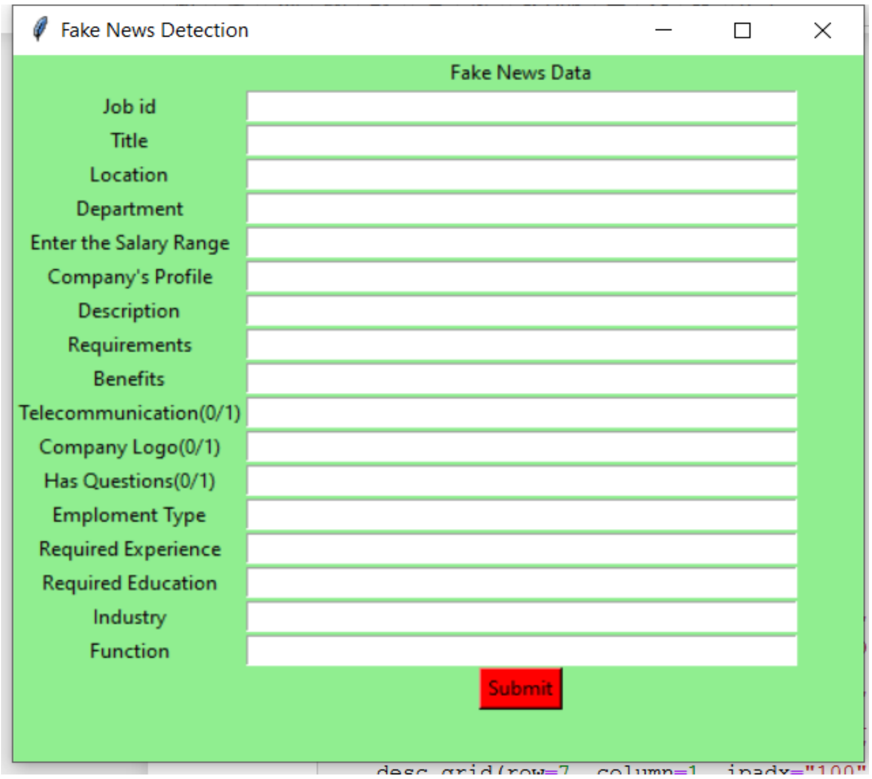
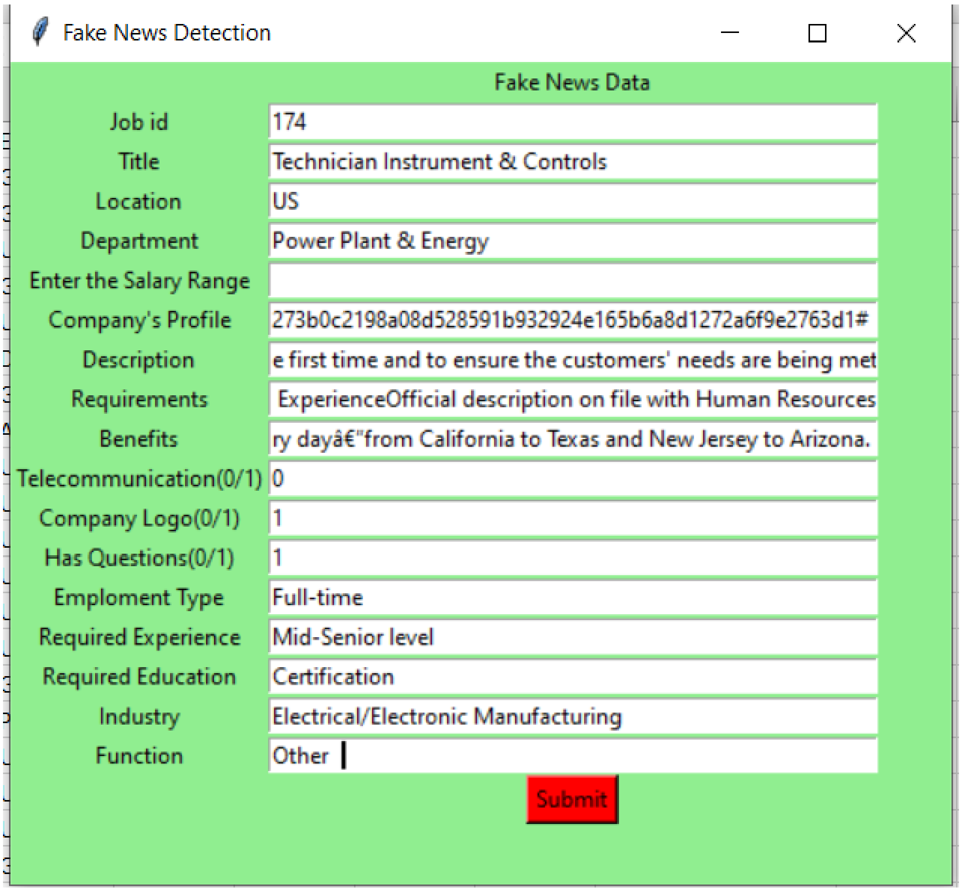


Figure 31GUI (for user's interaction)

User input:



Predicted label:

Table 10 Output table 1

	job_id	title	location	department	company_profile	description	requirements	benefits	has_company_logo	required_experience	re
0	1	sdg	df	gdf	df	h	NaN	hgfh	fgh	gfh	
1	2	dfh	gfh	gfj	j	gh	k	ghj	jh	j	
2	2	dfgdf	g	NaN	gfh	gf	hgfh	gfh	fgh	gfh	
3	4	dfg	df	h	gfh	fg	h	gfh	h	gf	
4	5	dgdfggds	g	df	dfh	gdf	h	gfh	gf	df	
5	6	khjk	hjkgghj	hg	jk	hj	l	kjl	m	hjl	
6	7	gj	ghk	NaN	hjkhjk	hjk	hjk	hjk	k	khjk	
7	8	ytujgyjyu	utyuytu	tutyuy	gyugiku	fjgykghk	ftjuugy	gjghikhj	fjgyuh	gkjghkhj	
8	9	Payroll Data Coordinator Positions - Earn $100...	US, KS, Abbyville	NaN	NaN	We are a full-service marketing and staffing f...	RequirementsAll you need is access to the Inte...	This is an entry level position and we offer f...	0	NaN	
9	10	Technician Instrument & Controls \n	US\n	Power Plant & Energy\n	Edison International and Refined Resources hav...	Technician Instrument & ControlsLocation D...	JOB QUALIFICATIONS-Ability to understand proce...	we are a team of almost 8,000 employees who he...	1	Mid-Senior level\n	

to be continued…

Table 11 Output table 2

file	description	requirements	benefits	has_company_logo	required_experience	required_education	industry	function	FraudNews
df	h	NaN	hgfh	fgh	gfh	gf	hf	hf	False
j	gh	k	ghj	jh	j	ghj	ghgh	j	False
gfh	gf	hgfh	gfh	fgh	gfh	gfh	NaN	gfh	False
gfh	fg	h	gfh	h	gf	hf	h	fghf	False
dfh	gdf	h	gfh	gf	df	h	gfh	gf	False
jk	hj	l	kjl	m	hjl	j	lhjlh	lhjjhl	False
khjk	hjk	hjk	hjk	k	khjk	hjk	hjk	hjkklhj	False
jiku	fjgykghk	ftjuugy	gjghikhj	fjgyuh	gkjghkhj	uhkhul	fjghkghk	gfjghkghk	False
laN	We are a full-service marketing and staffing f...	RequirementsAll you need is access to the Inte...	This is an entry level position and we offer f...	0	NaN	NaN	NaN	NaN	False
son and ned av...	Technician Instrument & ControlsLocation D...	JOB QUALIFICATIONS-Ability to understand proce...	we are a team of almost 8,000 employees who he...	1	Mid-Senior level\n	Certification\n	Electrical/Electronic Manufacturing\n	Other\n	True

Therefore, all the user entered news postings is **"False"** indicating **NOT FRAUDULENT NEWS** whereas the last entered news is **"True"** indicating it as **FRAUDULENT NEWS.**

4.2) Discussion

Exploratory analysis of the data:

The given dataset is a record of job news postings uniquely identified by their job ids. It consists of news from about 2800+ locations of 1300+ departments with their descriptions, requirements and benefits they're offering. This data also checks whether the news has a company logo, reviewed questions etc. This data set contains extracted headlines of news such as employment type offered, required ed and experience, job function, the industry/company offering the job and the label whether that news is fraudulent or not.

On studying the data thoroughly and applying necessary operations certain observations were made such as,

1. Most null valued columns and it's percentage

	Number of NA	Percent NA
salary_range	15012	83.96
department	11547	64.58
required_education	8105	45.33

2. Dataset has **16687** not fraud news & **846** fraud news.
3. Low significant features were removed to increase accuracy of models such as **telecommuting**, **salary_range**, **has_questions** and **employment_type**.

Summary of trends of data /visualisation and significant and insignificant factors of a fraud(**Red** color indicates **fraud** while **Green** **not fraud**):

1. The job postings which ask for **Some High School Coursework** as **required_education** is observed showing most **fraudulent** nature with **above 75% fraudit**y followed by **Certification** with **nearly 10% fraudity**.
2. Whereas, news with **required_education** as **Vocational**, **Vocational - Degree** and **Vocational - HS Diploma** have shown **NO case of fraudity**. **(NOTE: Though, all of them have less than 10 job posting news)**
3. The job postings which ask for **Entry level, Executive** and "**NULL**" as **required_experience** is observed to show maximum **fraudulent** nature than any other with **above 60% fraudity** with Entry level being the most out of three**.**
4. Whereas, news with **required_experience as Internship and Associate** have shown **least case of fraudity.**
5. The job postings which ask for **Part-time** as **employment_type** is observed to show most **fraudulent** nature with **above 80% fraudity** followed by **Other** and "**NULL**" with **above 60% fraudity.**
6. Whereas, news with **employment_type** as **Temporary** has shown **least case of fraudity (less than 10%).**
7. The job postings news from **TX** as **one of the location** is observed to show most **fraudulent** nature with **above 15% fraudity** followed by **US** and **CA(Canada)** with **above 6% fraudity.**
8. Whereas, news with **one of the locations** as **London** and **LND** have shown **least case of fraudity. (less than 0.5 %)**
9. In general, job postings news with **required_education** as **Vocational** or **Vocational - Degree** or **Vocational - HS Diploma ; required_experience as Internship** or

Associate ; employment_type as **Temporary ; one of the locations** as **London** or **LND** are the **most authentic news** and they're **LEAST PROBABLE FRAUD NEWS** or **NO FRAUD NEWS.**

10. **Whereas,** job postings news with **Some High School Coursework** as **required_education ; Entry level** or **Executive** or "**NULL**" as **required_experience ; Part-time** as **employment_type ; TX** or **US** or **CA(Canada)** as **one of the location** are the **least authentic news** and they're **MOST PROBABLE FRAUD NEWS** or **DIRECTLY FRAUD NEWS.**

An overview of model's implementation and success percentages.

The problem statement demanded a predictive model on Job Posting news details predicting them being fraudulent or not fraudulent.

Since, we've been given the labelled data so we proceeded with Supervised Learning. Problem statement demanded classification so I chose to use the most accurate and effective algorithm, Random Forest. Python provides a very comfortable and efficient platform for implementation of such glorious algorithms and also visualizations for better understanding.

Correlation Matrix was used to identify the relations between pairs of two attributes.

It started with importing some python defined algorithms from sklearn and under various modules like model_selection, ensemble, datasets.

- from sklearn.model_selection import GridSearchCV
- from sklearn.model_selection import RandomizedSearchCV
- from sklearn.datasets import load_digits
- from sklearn.ensemble import RandomForestClassifier

5. CONCLUSION AND FUTURE SCOPE

Conclusion

The main contribution of this project is support for the idea that machine learning could be useful in a novel way for the task of classifying fake news. Our findings show that after much pre-processing of a relatively small dataset.
It started with importing some python defined algorithms from sklearn and under various modules like model_selection, ensemble, datasets.

- from **sklearn.model_selection** import **GridSearchCV**
- from **sklearn.model_selection** import **RandomizedSearchCV**
- from **sklearn.datasets** import **load_digits**
- from **sklearn.ensemble** import **RandomForestClassifier**

It started with splitting the dataset as **test(0.2) and train(0.8)**. Then, using 20 Decision trees under random forest to predict for the test data using the train data and showed the accuracy result of top 3 decision tree models out of 20. It was the first raw approach towards the prediction model which resulted in an accuracy of **98.59%**

Now, to get the best accurate results and reduce the overfitting of data “Hyper Parameter Tuning” was a must. So, by importing the **GridSearchCV** and **RandomizedSearchCV** algorithms from sklearn.model_selection two other models were prepared with enhanced accuracy rates as **99.61%** and **98.63%** respectively. Receiver Operating characteristic curves were plotted for all 3 Random forest models of True and False positive rates. Lastly, by using GridSearchCV the best features or important features were identified.

As such, this seems to be a really good start on a tool that would be useful to augment humans ability to detect Fake News. Other contributions of this project include the creation of a dataset for the task and the creation of an application using **tkinter** module that aids in the visualization and understanding of the Random forest classification of a given data set. This application could be a tool for humans trying to classify fake news. It could also be useful in researchers trying to develop improved models through the use of improved and enlarged datasets, different parameters, etc. The application also provides a way to see manually how changes in the body text affect the classification. The classification of fake news from the real news is a very crucial task nowadays. It is becoming an imminent threat in some situations to be unable to discern real and fake news. Our best performing models achieved accuracies that are comparable to the human ability to spot fake content.

Future scope

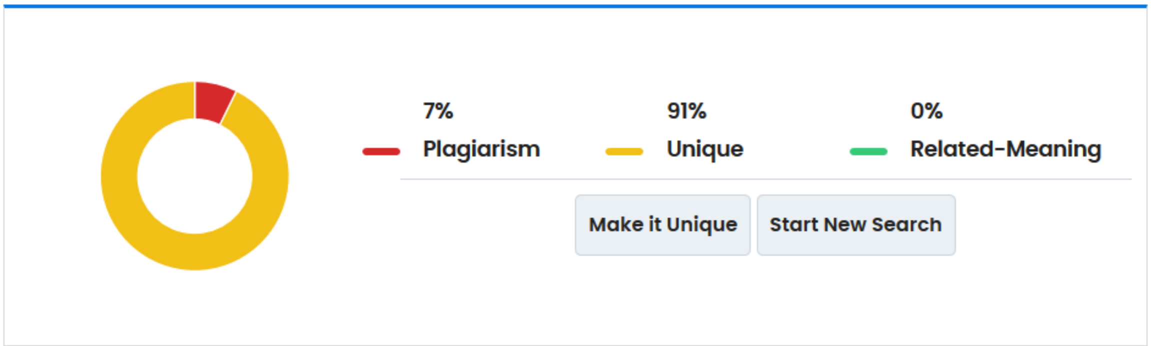
Through the work done in this project, we have shown that machine learning certainly does have the capacity to pick up on sometimes subtle language patterns that may be difficult for humans to pick up on. The next steps involved in this project come in three different aspects. The first aspect that could be improved in this project is augmenting and increasing the size of the dataset. We feel that more data would be beneficial in ridding the model of any bias based on specific patterns in the source. There is also question as to weather or not the size of our dataset is sufficient.
The second aspect in which this project could be expanded is by comparing it to humans performing the same task. Comparing the accuracies would be beneficial in deciding whether or not the dataset is representative of how difficult the task of separating fake from real news is. If humans are more accurate than the model, it may mean that we need to choose more deceptive fake news examples. Because we acknowledge that this is only one tool in a toolbox that would really be required for an end-to-end system for classifying fake news, we expect that its accuracy will never reach perfect. However, it may be beneficial as a stand-alone application if its accuracy is already higher than human accuracy at the same task. In addition to comparing the accuracy to human accuracy, it would also be interesting to compare the phrases/trigrams that a human would point out if asked what they based their classification decision on. Then, we could quantify how similar these patterns are to those that humans find indicative of fake and real news.

REFERENCES

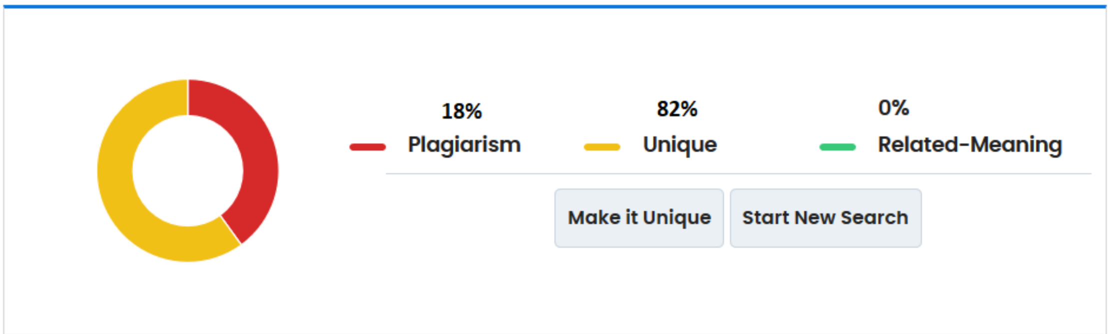
- Lilapati Waikhom , Rajat Subhra Goswami, “Fake News etection system using machine learning”, International Conference on Advancements in Computing & Management (ICACM-2019),page : 1-4 ,2019
- Xinyi Zhou, REZA Zafarani, “Fake news : A survey of research , detection and opportunities”, Association for Computing Machinery,Page: 3-7 , 2019
- Jiawei Zhang , Bowen Dong, Philip S. Yu, “FakeDetector : Effective fake news Detection with deep diffusive Neural network” , BDSC lab , Department of computer science , university of Chicago ,IL,USA, Page :3-7 , 2019

- https://www.kaggle.com/mrisdal/fake-news
- https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535
- https://becominghuman.ai/image-data-pre-processing-for-neural-networks-498289068258
- https://machinelearningmastery.com/logistic-regression-for-machine-learning/
- https://www.geeksforgeeks.org/decision-tree/

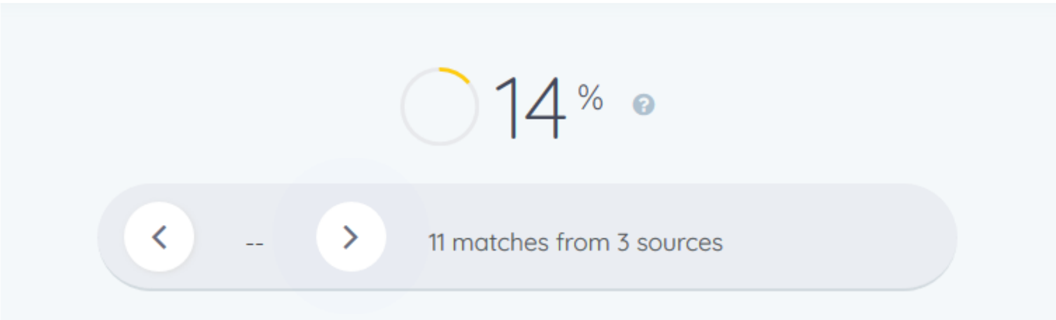
Plagiarism Checker



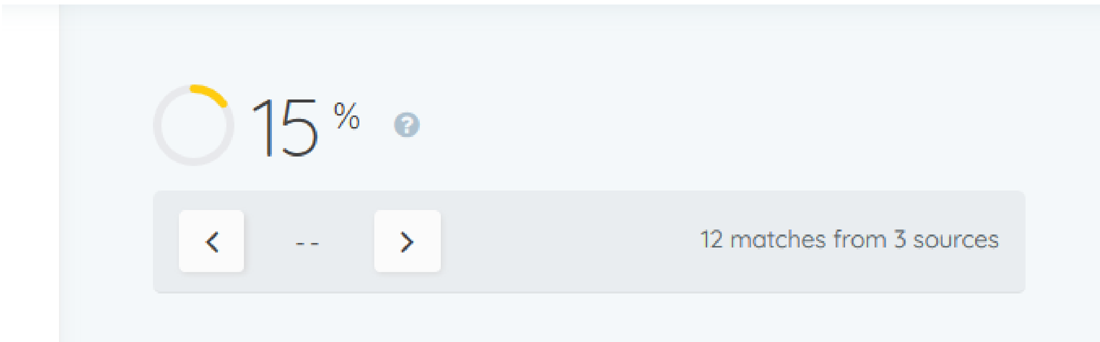
https://www.duplichecker.com/



https://www.duplichecker.com/



https://www.quetext.com/results/2825904e9e4d6d9433ec



https://www.quetext.com/results/2825904e9e4d6d9433ec