Bangalore Institute of Technology

Department of Computer Science and Engineering

K R Road, V V Puram, Bengaluru-560004

Mini Project Synopsis on

REAL OR FAKE JOB POSTING

Submitted as the mini project for the subject
Data Mining and Data Warehousing (18CS641)

Submitted by
ANUSHA.M. R 1BI18CS024
ASHIKA.K 1BI18CS029
CHANDANA.C 1BI18CS039

For academic year 2020-21

Under the guidance of
Dr. M.S. BHARGAVI
ASSISTANT PROFESSOR

INDEX

INTRODUCTION

PROBLEM STATEMENT:

To create classification models which can learn about job description which are fraudulent.

INTRODUCTION:

A job posting is the official advertisement of an open position for which the company is actively seeking a new-hire.

Job postings can be made internally by utilizing company bulletin boards or announcement forums. This allows current employees to move around within the company and bid for the position prior to it being opened to the general public.

Internal job postings are what is traditionally indicated by the term. Publically-advertised job postings can be found on websites, online job boards or staffing agency availability lists. Postings are traditionally written by the human resources department and include details regarding qualifications, hours and pay.

They may also describe the work environment or where the new hire would fit into the current company structure.

There are a lot of job advertisements on the internet, even on the reputed job advertising sites, which never seem fake. But after the selection, the so-called recruiters start asking for the money and the bank details. Many of the candidates fall in their trap and lose a lot of money and the current job sometimes.

So, it is better to identify whether a job advertisement posted on the site is real or fake. Identifying it manually is very difficult and almost impossible. We can apply machine learning to train a model for fake job classification. It can be trained on the previous real and fake job advertisements and it can identify a fake job accurately.

So we will train the machine learning classifier on dataset to identify the fake job advertisements. First, we will visualize the insights from the fake and real job advertisement and then we will use the Support Vector Classifier in this task which will predict the real and fraudulent class labels for the job advertisements after successful training. Finally, we will evaluate the performance of our classifier using several evaluation metrics.

DATASET DESCRIPTION:

This dataset contains 18K job descriptions out of which about 800 are fake. The data consists of both textual information and meta-information about the jobs. The dataset can be used to create classification models which can learn the job descriptions which are fraudulent. The different columns are:

Job_id: a unique id for each job

title: description about a job/position/designation held and gives a brief idea on what the job is about.

Location: job location.

Department: part of the organization that deals with a particular area of work.

Salary_range: the payment amount between a set of low to high numbers that an employee wants to receive once they're hired by a company.

Company_profile: a statement that describes a business essential element.

Description: the main purpose of job description is to collect job-related data in order to advertise for a particular job.

Requirements: job requirements are qualifications and skills necessary for a certain position. job requirements are usually written in form of a list that contains the most important qualifications that a candidate must possess in order to be able to perform certain job duties.

Benefits: employee benefits are non-salary compensation that can vary from company to company. benefits are indirect and non-cash payments within a compensation package. they are provided by organizations in addition to salary to create a competitive package for the potential employee.

Telecommuting: 0 if work from home is not allowed, and 1 if work from home is allowed

Has_company_logo: 1 if the job posting has company logo and 0 otherwise

Has_questions: 1 if there are questions in the posting and 0 otherwise

Employment_type: full_time or other(part-time)

Required_experience: experience requirements is the level experience on the job.

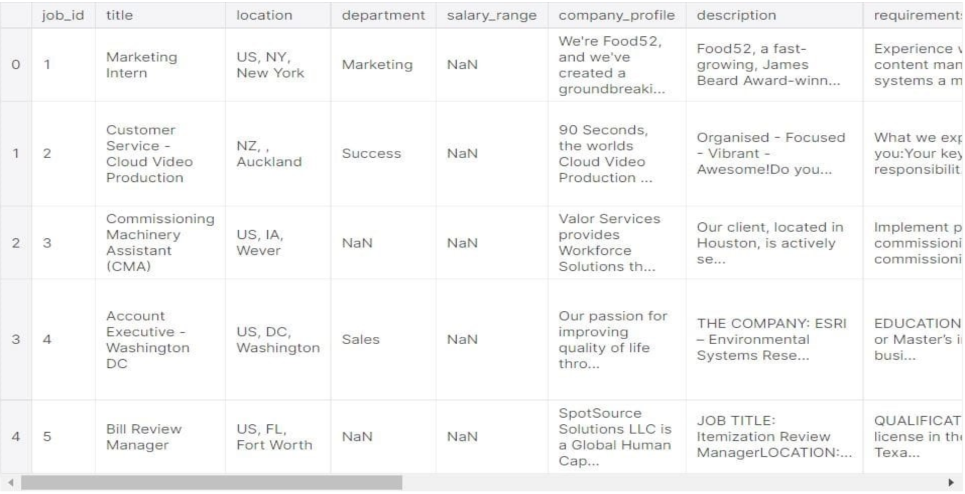
Required_education: The prerequisite education for the job

Industry: the industry under which the job falls

Function: Job function is the list of competencies expected for the job

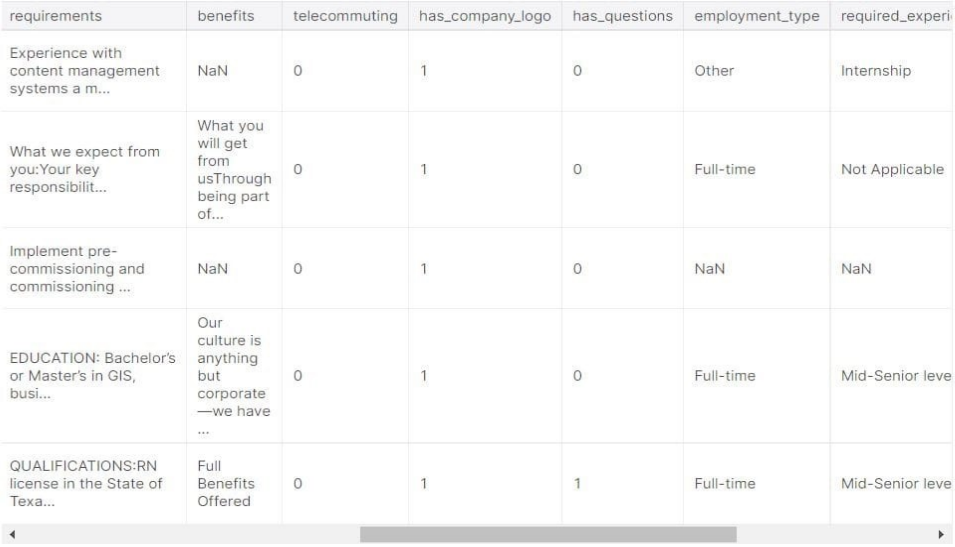
Fraudulent: 1 if the job posting is fake and 0 otherwise

DATASET SNAPSHOT:



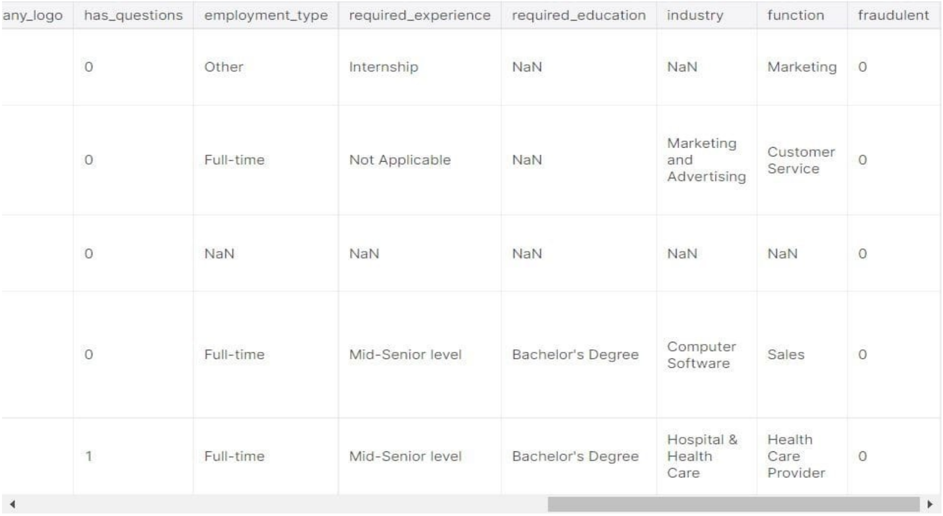
	job_id	title	location	department	salary_range	company_profile	description	requirement:
0	1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience v content man systems a m
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we exp you:Your key responsibilit
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement p commissioni commissioni
3	4	Account Executive - Washington DC	US, DC, Washington	Sales	NaN	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION or Master's i busi...
4	5	Bill Review Manager	US, FL, Fort Worth	NaN	NaN	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICAT license in th Texa...

Figure 1: Snapshot of Job_id



requirements	benefits	telecommuting	has_company_logo	has_questions	employment_type	required_experi
Experience with content management systems a m...	NaN	0	1	0	Other	Internship
What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...	0	1	0	Full-time	Not Applicable
Implement pre-commissioning and commissioning ...	NaN	0	1	0	NaN	NaN
EDUCATION: Bachelor's or Master's in GIS, busi...	Our culture is anything but corporate—we have ...	0	1	0	Full-time	Mid-Senior leve
QUALIFICATIONS:RN license in the State of Texa...	Full Benefits Offered	0	1	1	Full-time	Mid-Senior leve

Figure 2: Snapshot of Requirements



any_logo	has_questions	employment_type	required_experience	required_education	industry	function	fraudulent
	0	Other	Internship	NaN	NaN	Marketing	0
	0	Full-time	Not Applicable	NaN	Marketing and Advertising	Customer Service	0
	0	NaN	NaN	NaN	NaN	NaN	0
	0	Full-time	Mid-Senior level	Bachelor's Degree	Computer Software	Sales	0
	1	Full-time	Mid-Senior level	Bachelor's Degree	Hospital & Health Care	Health Care Provider	0

Figure 3: Snapshot of Types

TOOLS AND TECHNOLOGIES USED:

- Python –

Reasons for using Python:

1. Python is easy to learn and understand and has a simple syntax.
2. The programming language is scalable and flexible.
3. It has a vast collection of libraries for numerical computation and data manipulation.
4. Python provides libraries for graphics and data visualization to build plots.
5. It has broad community support to help solve many kinds of queries.

- Natural language processing (NLP).
- Machine Learning (scikit-learn library).

  Scikit-learn is an open source Python library that has powerful tools for data analysis and data mining

  SciPy, an ecosystem consisting of various libraries for completing technical computing tasks.

- Numpy and Pandas.

  Numpy stands for "numerical python". It offers pre-compiled functions for numerical routines. Pandas is perfect for data analysis, manipulation and visualisation. It allows high-level data structures and some tools to manipulate them.

  NumPy, a library for manipulating multi-dimensional arrays and matrices. It also has an extensive compilation of mathematical functions for performing various calculations.

- Jupyter Notebook

  Jupyter notebook is a flexible tool that helps us create readable analysis, as we can keep code, images, comments, formulae and plots together.

PREPROCESSING TECHNIQUE:

A new position in an organization requires a detailed and complete job description. A job description is an internal document that details the responsibilities, authority, nuances, decision authority and working conditions associated with the job at hand.

A job posting, on the other hand, is an advertisement meant to attract applicants. The posting, therefore, should be a slimmed down, concise, yet hyped-up version of the description designed to attract applicants.

Put another way, a job description is akin to a legal document outlining the responsibilities and duties of a position. A job posting is the advertisement “selling” the position to potential applicants.The most effective job ads successfully brand the company and sell the position.

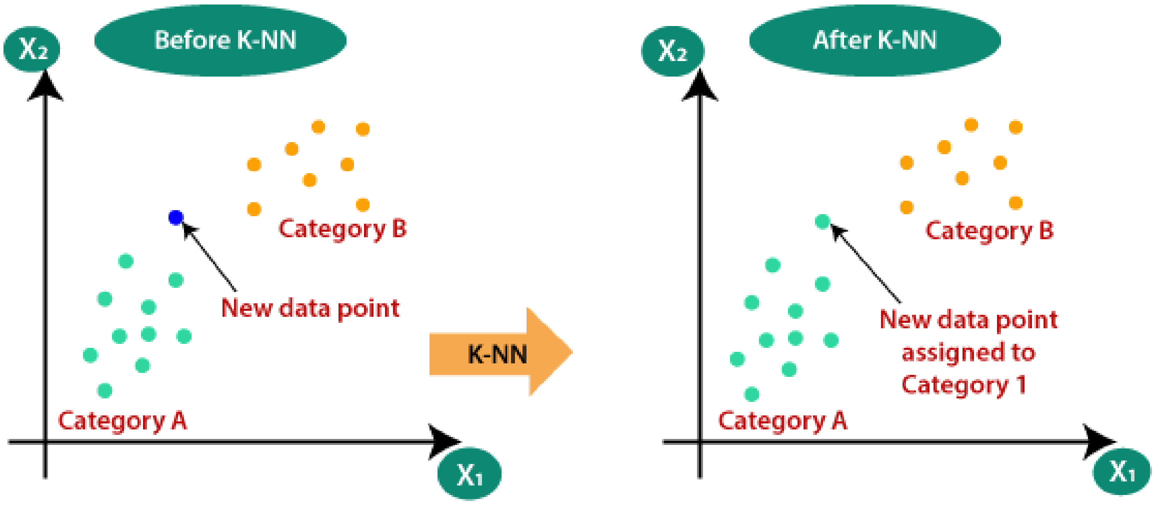
A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different language.

CLASSIFICATION/CLUSTERING TEHCNIQUE:

1.K-NEAREST NEIGHBOUR

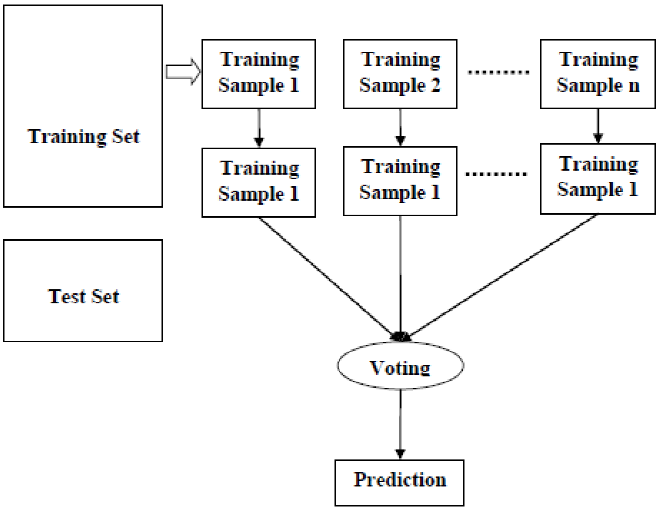
K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories’-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.



2.RANDOM FOREST

Random forest is a supervised learning algorithm which is used for both classification as well as regression.

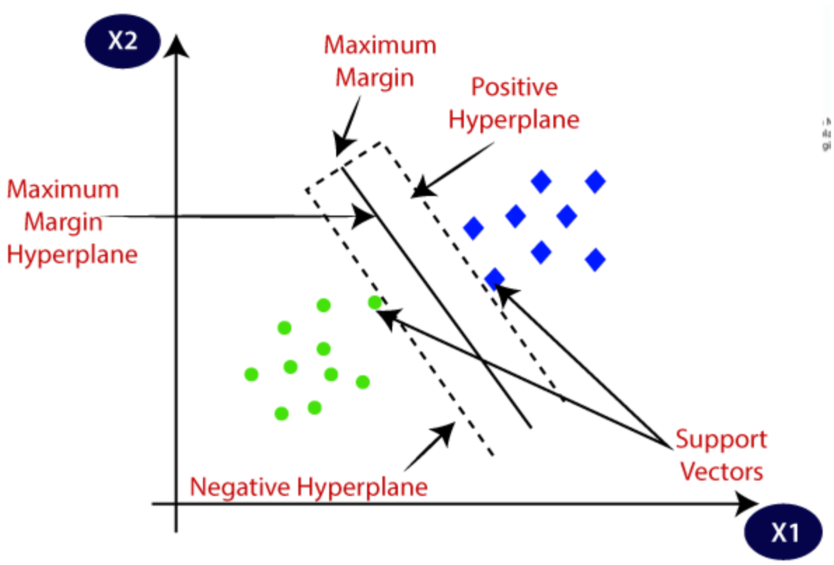
Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.



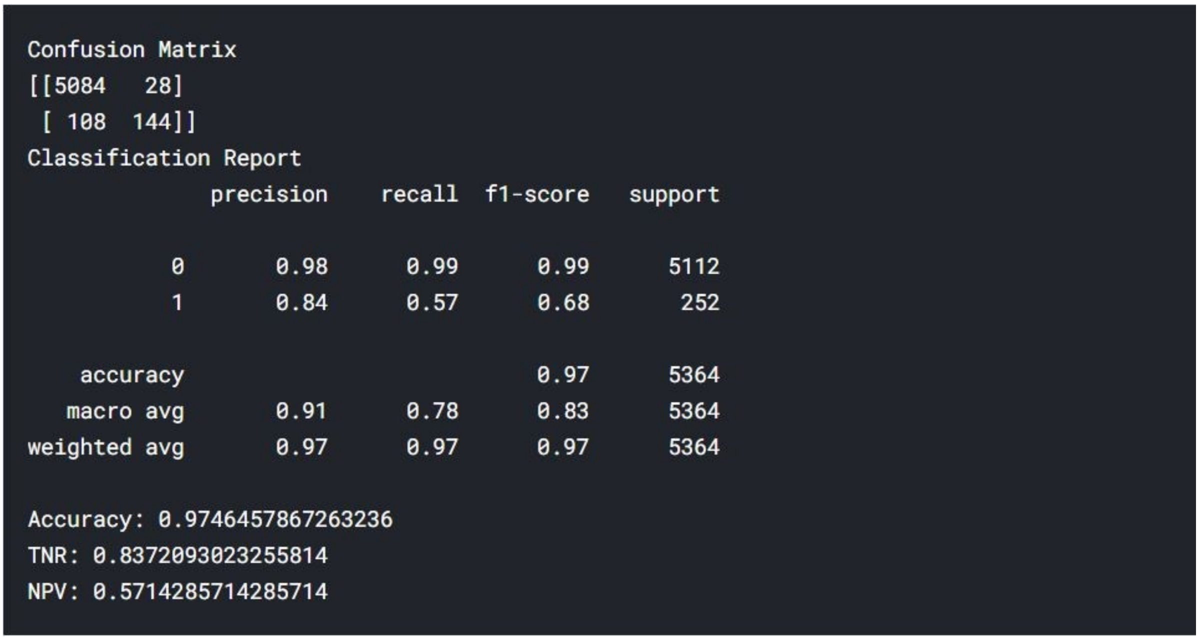
3.SUPPORT VECTOR MACHINE

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

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane



Results:



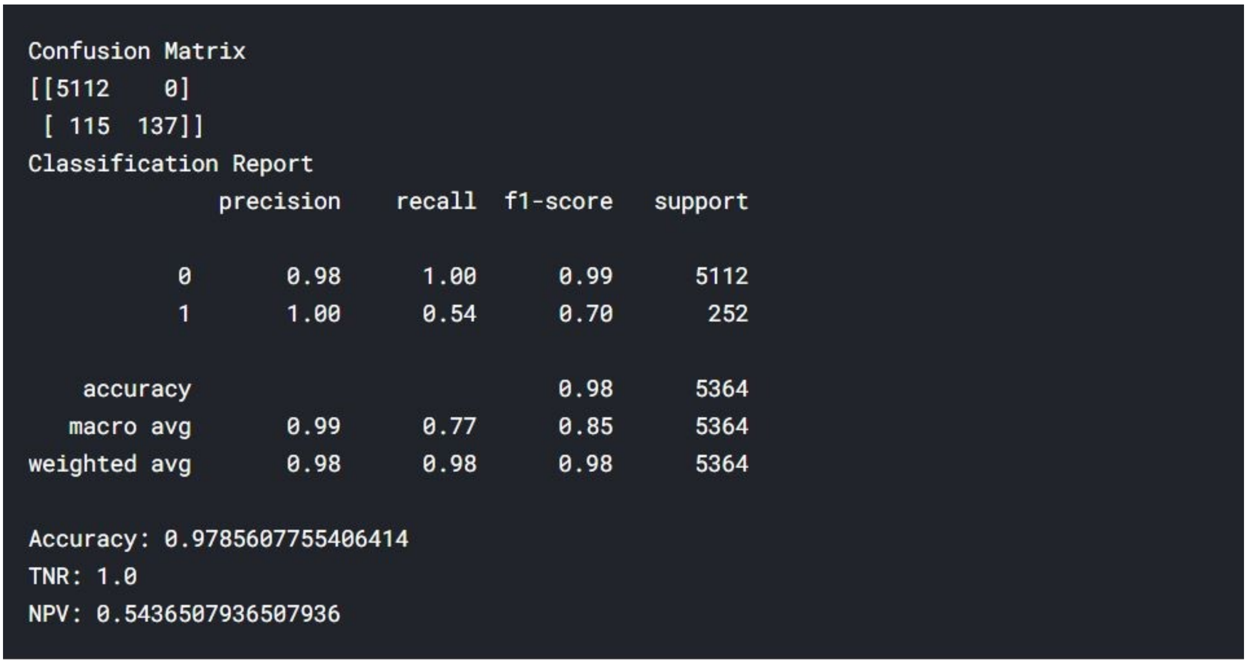
```
Confusion Matrix
[[5084   28]
 [ 108  144]]
Classification Report
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      5112
           1       0.84      0.57      0.68       252

    accuracy                           0.97      5364
   macro avg       0.91      0.78      0.83      5364
weighted avg       0.97      0.97      0.97      5364

Accuracy: 0.9746457867263236
TNR: 0.8372093023255814
NPV: 0.5714285714285714
```

Figure 1: Result of KNN algorithm



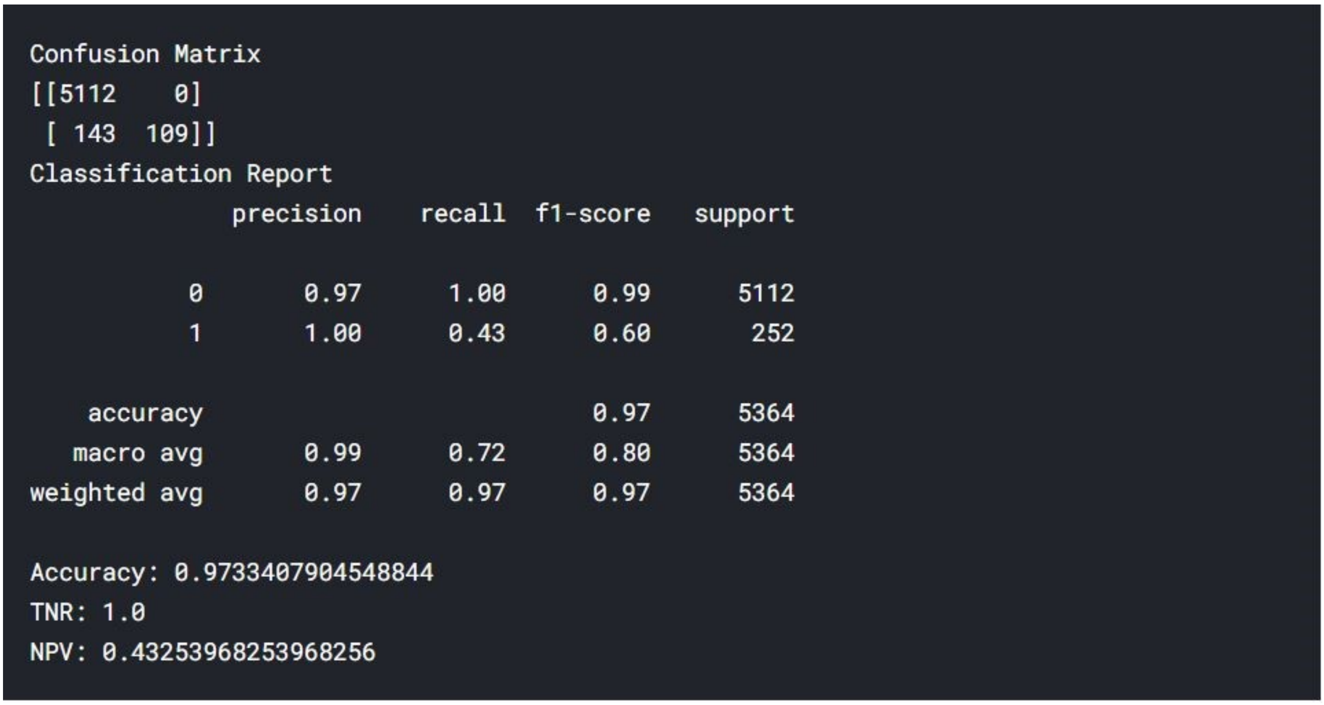
```
Confusion Matrix
[[5112    0]
 [ 115  137]]
Classification Report
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5112
           1       1.00      0.54      0.70       252

    accuracy                           0.98      5364
   macro avg       0.99      0.77      0.85      5364
weighted avg       0.98      0.98      0.98      5364

Accuracy: 0.9785607755406414
TNR: 1.0
NPV: 0.5436507936507936
```

Figure 2: Result of Random Forest Algorithm



```
Confusion Matrix
[[5112    0]
 [ 143  109]]
Classification Report
              precision    recall  f1-score   support

           0       0.97      1.00      0.99      5112
           1       1.00      0.43      0.60       252

    accuracy                           0.97      5364
   macro avg       0.99      0.72      0.80      5364
weighted avg       0.97      0.97      0.97      5364

Accuracy: 0.9733407904548844
TNR: 1.0
NPV: 0.43253968253968256
```

Figure 3: Result of SVM algorithm