Detecting Fraud Calls *vis-à-vis* Natural Language Processing

Anurag Dutta[1*]

[1*]Undergraduate, Computer Science and Engineering, Government College of Engineering and Textile Technology, 12, William Carey Road, Serampore, 712201, Calcutta, India.

Corresponding author(s). E-mail(s): anuragdutta.research@gmail.com;

Abstract

Purpose: Fraud is defined in law as the willful use of deception to obtain unfair or illegal gain or to deny a victim of a legitimate right. Fraud can be illegal under criminal law or civil law. Scam may not always result in a loss of money, property, or legal rights but still be a component of another civil or criminal wrong. A victim of fraud may sue the offender to stop the fraud or receive monetary compensation. The goal of fraud may be financial gain or other benefits, such as getting a passport, travel document, or driver's licence, or it may be mortgage fraud, when the offender makes false claims in an effort to qualify for a mortgage. The collection of enormous volumes of data combined with predictive analytics or forensic analytics, the use of electronic data to reconstruct or detect fraud, makes the detection of fraudulent acts on a wide scale conceivable. Particularly when using computer-based analytical techniques, errors, inconsistencies, inefficiencies, irregularities, and biases can be exposed, which frequently pertain to fraudsters favouring particular currency amounts in order to bypass internal control thresholds. Scam calls are false calls that persons or businesses make in an effort to deceive recipients into parting with their money or private information. Scammers frequently play down the situation by calling it a normal call or even by lying to the person they were trying to con. In this work, we will try to detect fraud calls, by making use of Supervised Machine Learning Algorithms. For the same, we have collected a total of 5925 Data points. We have used the data to train the model. For obvious reasons, we have kept the source of the data as anonymous. This work would help in early detection of scam calls and would help a

lot of innocent lives from getting tied by the knot of fraudulences.

Methods: Numerous Machine Learning Algorithms have been used for the work, namely Support Vector Machine, ***k*** - Nearest Neighbours, Gaussian Naive Bayes, etc. Further, these Algorithms have been compared on the basis of their performance of prediction. We have also used the corpus, from Natural Language Toolkit Python Library, for preprocessing of data.

Results: According to the work proposed in the article, k Nearest Neighbour and the Support Vector Machine are best for detection of Fraud Calls, as the model modelled using k Nearest Neighbour Classifier and the Support Vector Classifier gives an accuracy of 99%.

Conclusion: Fraudulences have caught great height in current work of advancing informatics. The betterments in the Information Technology can be taken into account to avoid innocent people from getting trapped. Government can impose laws and regulations advising the Telecom Exchange companies to introduce the Scam Call Classifier Technique as their key element. This would adversely impact the surging rate of Fraudulence.

Keywords: Supervised Learning, Support Vector Machine, Gaussian Naive Bayes, *k* - Nearest Neighbours, Fraud Detection, Criminology, Natural Language Processing

JEL Classification: C810 , C820 , O390

1 Introduction

Fraud is a purposeful act of deception that leads to unethical or unfair behaviour, according to the law. Fraud is frequently categorised as a criminal offence. Fraudulent behaviour Kaltenbrunner et al (2018) is motivated for a variety of causes. Others fabricate the assertions for their own personal gain, which cannot be done without committing a crime Rotaru et al (2022), while some people want to benefit from value. According to the most popular definition, fraud is "a deception with the goal to benefit financially or personally." Different legal systems have different definitions of what constitutes a valid explanation for fraud, which may be a civil wrong requiring proof and a rationale, or a criminal infraction if the intentional deception has been justified. Financial scams are common and cost the people who commit them a lot of money. For instance, a business or bank may trick a customer into paying for extra services. The Wells Fargo bank in the United States experienced one of the same situations. A bank is the target of a bank fraud, which is typically carried out by making false claims and utilising fictitious documentation. Everyone is concerned about bank fraud. It is a highly delicate subject since it

impacts public trust, which is the foundation of the entire banking system Zhu et al (2022a). The term "bank fraud" refers to a variety of thefts, embezzlements, and falsifications of negotiable documents such checks, bank draughts, bills of exchange, statements of accounts, stocks, etc. The booming banking industry has led to an increase in bank fraud. Indian Penal Code Paramasivan et al (2022) addresses bank fraud. By using a fake signature on the check, frauds are frequently conducted in the domain of checks. The majority of bank frauds presumably occur in this manner. Hypothecation fraud is another type of fraud in which money is fraudulently obtained in exchange for a security of some sort. A fair financial system that enjoys the public's confidence is a prerequisite for an equitable economic system. To prevent it, a number of preventive steps can be done, such as proper recruiting, constant attention, adherence to regulations, training programmes, etc. Telephonic Frauds are an another strata of Fraudulences that affect many people round the world. People are constantly communicating. Sadly, they don't always communicate in the most polite manner. While some attempts at communication result in violent or unpleasant interactions, other attempts at communication result in frauds. Scams occur on a variety of platforms, including the internet and the phone. An individual is frequently deceived into divulging personal information unknowingly when a fraud occurs. Although there are many other ways available, intimidation tactics are frequently used. In particular, the general public is compelled to accept criminals when they employ tech assistance as a front, lest they leave their computers vulnerable to malware attacks. In essence, there are many other types of scams, and this is just one of the most recent ones that customers should be aware of. Scams, whether they are conducted over the phone or in another way, need a vulnerability to operate. There are many different kinds of breaches, ranging from data to financial. The 2014 public awareness campaign on the Telephone Technical Support Scam involved an unintentional disclosure. Technical support frightened people into providing personal information when they contacted, leading them to believe they were chatting with experts who would clean their machines of malware Rafiq et al (2022) or viruses Lwoff (1967). They also kept sending money to a person they thought was a government official after learning that the technical support was a hoax. Despite the fact that they shared information, they had no intention of sharing it with a fraudster. Studies make sure that the Telephonic Phone is compromising the security and privacy of many individuals. In this work, we would build a model that can predict a call as being Fraud or Normal. From the problem statement, it's clear that it's a classification problem. So, for modelling the same, we would make use of Supervised Learning Algorithms, like Gaussian Naive Bayes Pan et al (2022), Support Vector Machine He et al (2022), k - Nearest Neighbours Uddin et al (2022), etc. For solving any Machine Learning Problem, before we design the Model, we will have to pre-process the data. Since, the data may have a lot of noise, and nuisances, Natural Language Processing Russo et al (2022) Techniques have been employed in our work to prepare the model that could lead to higher accuracies.

2 Dataset

The Dataset, that we have used for the work, is a collection of 05925 Data points. To avoid compensation of private sources, the source of Dataset is kept private. Given are, a few instances of the dataset. Figure 1 explains the prediction mechanism of data points by the Machine Learning Algorithms.

INDEX	LABEL	FEATURE
0	fraud	Hello, i m bank manager of SBI, ur debit card ...
1	fraud	Todays Vodafone numbers ending with 4882 are s...
2	normal	Please don't say like that. Hi hi hi
3	normal	Thank you!
4	normal	Oh that was a forwarded message. I thought you...
..	..	..
..	..	..
..	..	..
5920	fraud	to get 1000 INR voucher please call on 8898655...
5921	fraud	to get free access of google cloud account hit...
5922	fraud	to get free AWS cloud account hit on given mes...
5923	fraud	to get free access of Microsoft Azure hit on g...
5924	fraud	hello sir, we are from your bank have you fill...

Machine Learning Algorithms takes feature as input, and returns back the Label as an output.

3 Natural Language Processing

The study of how computers interact with human language, particularly how to design computers to process and analyse huge amounts of natural language data, is known as natural language processing (NLP), an interdisciplinary subject of linguistics, computer science, and artificial intelligence Olsson et al (2022). The ultimate goal is to create a machine that is able to "understand" the contents of documents, including the subtle subtleties of language used in different contexts. Once the information and insights are accurately extracted from the documents, the technology can classify and arrange the documents themselves. In many different commercial disciplines and places, personal assistants employ this technology. The system analyses the verbal input from the user, deconstructs it for accurate comprehension, and then processes it as necessary. Due to the fact that it is a very current and successful strategy, there is a huge demand for it right now. It has already been possible to have interactive conversations with a human being and work with smart devices thanks to advancements in the field of natural language processing Wahab et al (2021).

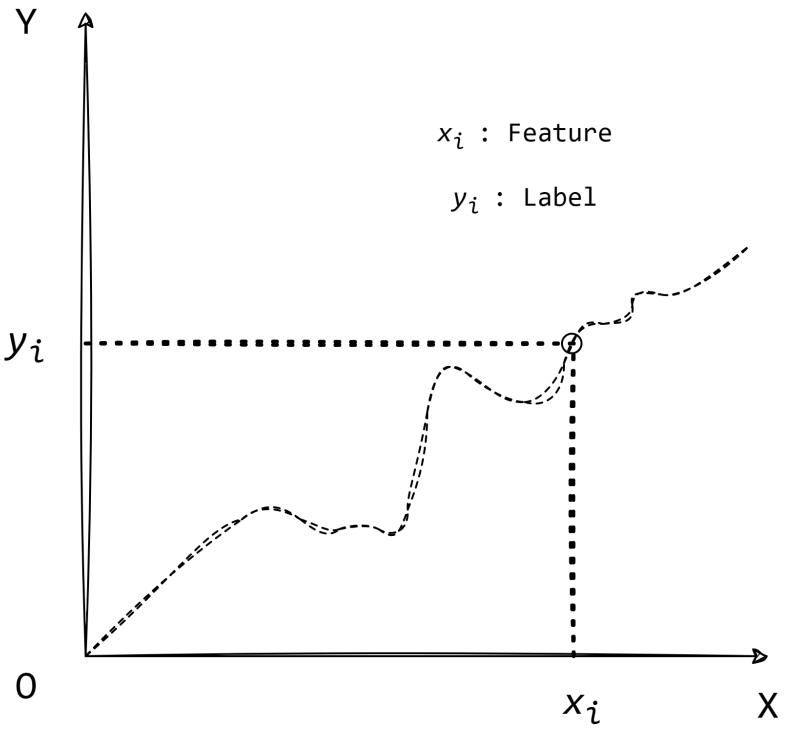

Fig. 1 Machine Learning Algorithms to Predict Label from a Feature. Algorithms, take the feature, x_i as input, and predict the label, y_i as output. The performance of the Machine Learning Algorithms is a measure of how well does the curve mimic the real time scenario. The equation representing the curve is $y = \phi(x) + \epsilon$. $\phi(x)$ is also known as Target Function, and ϵ is the Error Term.

The focus of AI applications in NLP was on knowledge representation, logical reasoning Houdé and Tzourio-Mazoyer (2003), and constraint fulfilment. Here, it was used first for semantics and then for grammar. A dramatic shift in NLP research over the past ten years has led to the extensive use of statistical techniques like machine learning Ward et al (2022) and data mining Martinez-Morata et al (2022). Due to the amount of work that needs to be done these days, automation is constantly needed. When it comes to automated applications, NLP is a really positive component. NLP is one of the most popular ways to deploy machine learning due of its applications. In order to better understand how computers and people communicate using natural language, the area of "natural language processing" (NLP) combines computer science, linguistics, and machine learning. The objective of NLP is to enable computers to comprehend and produce human language. This not only increases the effectiveness of human work but also facilitates communication with machines. NLP

fills the communication gap between people and machines. The Natural Language Toolkit, or more simply NLTK, is a collection of Python-coded tools and applications for symbolic and statistical natural language processing (NLP) of English. It was created by Steven Bird and Edward Loper at the University of Pennsylvania's Department of Computer and Information Roth et al (2022) Science. NLTK contains sample data and graphical demos. In addition to a cookbook, it comes with a book that describes the fundamental ideas behind the language processing jobs Zhu et al (2022b) that the toolkit supports. The goal of NLTK is to facilitate research and instruction in NLP or closely related fields including information retrieval, cognitive science Lordén et al (2022), artificial intelligence, and empirical linguistics Maiorani et al (2022). NLTK has been effectively used as a teaching tool, a tool for solitary study, and as a foundation for developing research systems. 25 nations, including 32 colleges in the US, use NLTK in their classes. Functionalities for categorising, tokenizing, stemming, tagging, parsing, and semantic reasoning are supported by NLTK. According to it's Stable Release available at GitHub, NLTK source code is distributed under the `Apache 2.0 License`.and the documentation is distributed under the `Creative Commons 3.0` license. Natural Language Procesing. Anyways, NLTK has a lot of useful functionalities. In this work, we will make use of a few, namely, `WordNetLemmatizer`(), and `TfidfVectorizer`().

3.1 `WordNetLemmatizer` Functionality

The process of stemming entails creating morphological variations of a root or base word. Stemming methodologies or stemmers are other names for stemming programmes. Stemming is a method used in linguistic Bromham et al (2022) morphogenesis and information retrieval that reduces inflected or occasionally derived words to their word stem, base, or root form typically a written word form. The stem need not be the same as the word's morphological root; it is typically enough that related words map to the same stem, even if that stem is not a legitimate root in and of itself. Since the 1960s, computer science has investigated stemming algorithms. Many search engines use conflation, or query expansion, to treat terms with the same stem as synonyms. *Julie Beth Lovins* penned the first stemmer that was ever published in 1968. For its early publication date, this report was noteworthy, and it greatly influenced subsequent research in this field. Her paper makes reference to three earlier significant attempts at stemming algorithms, including those made by Professor John W. Tukey of Princeton University, Michael Lesk at Harvard University working under Professor Gerard Salton, and James L. Dolby of R and D Consultants in Los Altos, California. The terms "chocolates," "chocolatey," and "choco" are condensed by a stemming algorithm to the base word, "chocolate".

Lemmatization Ramezani et al (2022) is far more effective than stemming. When applying morphological analysis to words, which aims to eliminate just inflectional endings and return the base or dictionary Rajkomar et al (2022) form of a word, known as the lemma, it looks beyond word reduction and takes

into account a language's entire lexicon. Some examples of Lemmatization would be

ORIGINAL WORD	LEMMA
meeting	meet
was	be
mice	mouse
cats	cat
tried	tried
..	..
..	..
..	..
etc.	etc.

The dataset that we are using in the work, may contain several stopwords. Stop words are terms in a halt list or negative lexicon that are removed from natural textual information or text prior to or after processing because they are unimportant. It is true that no natural language processing technologies use a truly universal list of stop words, nor are there any established guidelines for detecting them. NLTK by default uses a set of around 40 stopwords. The underwritten code fragment would print all the stopwords in the English Corpus of NLTK.

```
"""
Created on Mon Dec 26, 2022

Author: Anurag Dutta (anuragdutta.research@gmail.com)

"""

import nltk
from nltk.corpus import stopwords

stop_words = set(stopwords.words('english'))
print(stop_words)
```

Some are, 'yourselves', 'if', 'ain', 'its', etc. Note, NLTK can also filter out stopwords from languages other than English. We will just have to instantiate the stopwords.words('language'), parameter with different languages.

3.2 `TfidfVectorizer` Functionality

Term frequency–inverse document frequency, or $\texttt{TF} - \texttt{IDF}$, is a numerical statistic used in information retrieval to indicate a word's importance to a document. In information retrieval, text mining, and user modelling searches, it is frequently employed as a weighting factor. To account for the fact that some words are used more frequently than others overall, the $\texttt{TF} - \texttt{IDF}$ value rises according to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the term. One of the most common term-weighting techniques used nowadays is $\texttt{TF} - \texttt{IDF}$. According to a 2015 survey, $\texttt{TF} - \texttt{IDF}$ is used by 83% of text-based recommender systems in digital libraries. Mathematically, $\texttt{TF} - \texttt{IDF}$ is a product of two statistics, term frequency and inverse document frequency. Term Frequency, $\mathfrak{F}(\sqcup,\ \lceil)$ is the proportion of times term $\sqcup$ appears in document $\lceil$. Mathematically,

$$\mathfrak{F}(\sqcup,\ \lceil) = \frac{\Phi_{\sqcup,\ \lceil}}{\sum_{\sqcup' \in\ \lceil} \left(\Phi_{\sqcup',\ \lceil}\right)} \tag{1}$$

where,
$\Phi_{\sqcup,\ \lceil}$ is the the frequency with which phrase $\sqcup$ appears in document $\lceil$.
$\sum_{\sqcup' \in\ \lceil} \left(\Phi_{\sqcup',\ \lceil}\right)$ is the whole term count in document $\lceil$, tallying every instance of a given term separately.

The amount of information a word conveys, or whether it is common or uncommon across all texts, is indicated by the inverse document frequency. It is the inverse logarithmically scaled percentage of documents that contain the word, calculated by dividing the total number of documents by the number of documents that contain the phrase, and then calculating the logarithm of that quotient. Mathematically,

$$\mathfrak{I}(\sqcup,\ \mathcal{D}) = log\left(\frac{|\mathcal{D}|}{1 + |\lceil \in \mathcal{D},\ \sqcup \in \lceil|}\right) \tag{2}$$

where,
$|\mathcal{D}|$ is the count of documents in the corpus as a whole.
$|\lceil \in \mathcal{D},\ \sqcup \in \lceil|$ is number of texts that use the word $\sqcup$. Now, if $|\lceil \in \mathcal{D},\ \sqcup \in \lceil| = 0$, it would lead to Division by Zero error. To normalize that, 1 is added to the term.

Now, $\texttt{TF} - \texttt{IDF}$ mathematically, would be equal to,

$$\mathfrak{F}\mathfrak{I}(\sqcup,\ \lceil, \mathcal{D}) = \left(\frac{\Phi_{\sqcup,\ \lceil}}{\sum_{\sqcup' \in\ \lceil} \left(\Phi_{\sqcup',\ \lceil}\right)} \times log\left(\frac{|\mathcal{D}|}{1 + |\lceil \in \mathcal{D},\ \sqcup \in \lceil|}\right)\right) \tag{3}$$

Here, we use the $\texttt{TF} - \texttt{IDF}$ functionality of the NLTK Package to make a list of all the terms in the dataset as a vector. We take into account a word's

entire document weightage in $\texttt{TF} - \texttt{IDF}$. It aids us in coping with the most common words. We can punish them with it. The word counts are weighted by a measure of how frequently they appear in the documents by $\texttt{TF} - \texttt{IDF}$. The underwritten code fragment shows the implementation of $\texttt{TF} - \texttt{IDF}$ functionality of NLTK.

```
"""
Created on Mon Dec 26, 2022

Author: Anurag Dutta (anuragdutta.research@gmail.com)

"""

import nltk
from sklearn.feature_extraction.text import TfidfVectorizer

TF_IDF = TfidfVectorizer()
datum = TF_IDF.fit_transform(corpos).toarray()
print(datum)
```

4 Results

A machine learning paradigm known as supervised learning (SL) is used to solve issues when the data at hand consists of labelled instances, which means that each data point has variables, features, and a label. The aim of supervised learning algorithms is to learn a function from example input-output pairs that maps feature vectors, inputs, to labels, output. From labelled training data made up of a collection of training instances, it infers a function. Each example in supervised learning is a pair made up of an input object and the supervisory signal, which is the desired output value. An inferred function is generated by a supervised learning algorithm from the training data, which may then be used to map fresh samples. The algorithm will be able to accurately determine the class labels for instances that are not yet visible in an ideal environment. This calls for the learning algorithm to "reasonably" generalise from the training data to hypothetical situations. The so-called generalisation error is used to gauge an algorithm's statistical performance. The Fraud Call detection is a Classification based Supervised Learning Problem. The Algorithms used hereby for the classification are,

1. *Support Vector Machines*: Support Vector Machines are supervised learning models in machine learning that analyse data for regression and classification using a learning algorithm. Developed by AT&T Bell Laboratories

employees, notably Vladimir Vapnik. SVM is one of the most reliable statistical learning - based prediction systems. Given a sequence of training examples, each of which corresponds to a single binary category, the SVM training procedure generates a model that assigns subsequent instances to either group. Non - probabilistic bipolar linear classifier describes this model. SVM converts training data into space-based points in order to increase the separation between two categories. Additional samples are projected to the same space and are anticipated to fit into a particular group based on which side of the gap they are on. In addition to linear classification, SVMs may do non - linear classification by implicitly translating the input to a high-dimensional feature space using so called kernel techniques. For Support Vector Machine, the results are as

```
==========================
[ Support Vector Machine ]
==========================

Accuracy:  0.979746835443038
Recall:  0.9981006647673314
```

Figure 2 represents the Precision Recall Curve, while Figure 3 shows the Confusion Matrix.

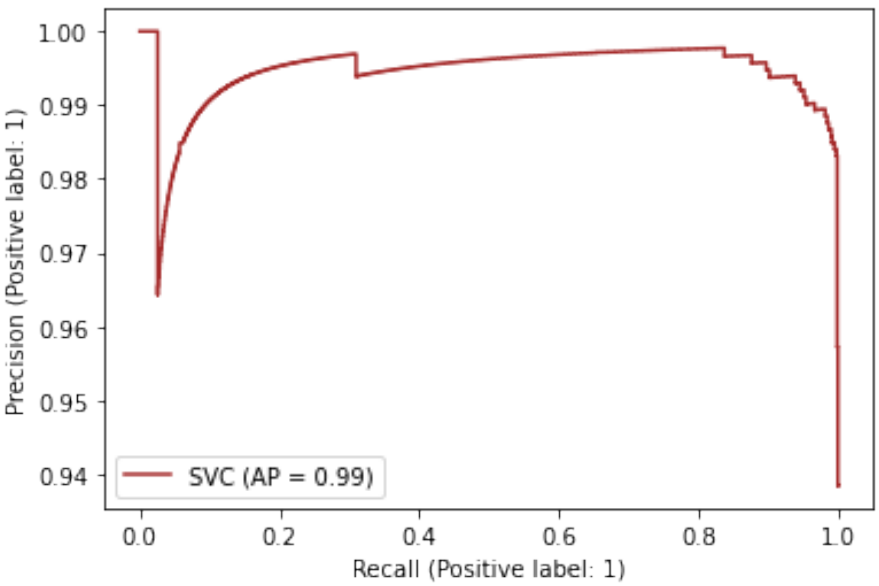

Fig. 2 Precision Recall Curve for the Support Vector Machine.

2. *Random Forest*: During training, several decision trees are built using the Random Forest (Random Decision Forest) ensemble learning approach,

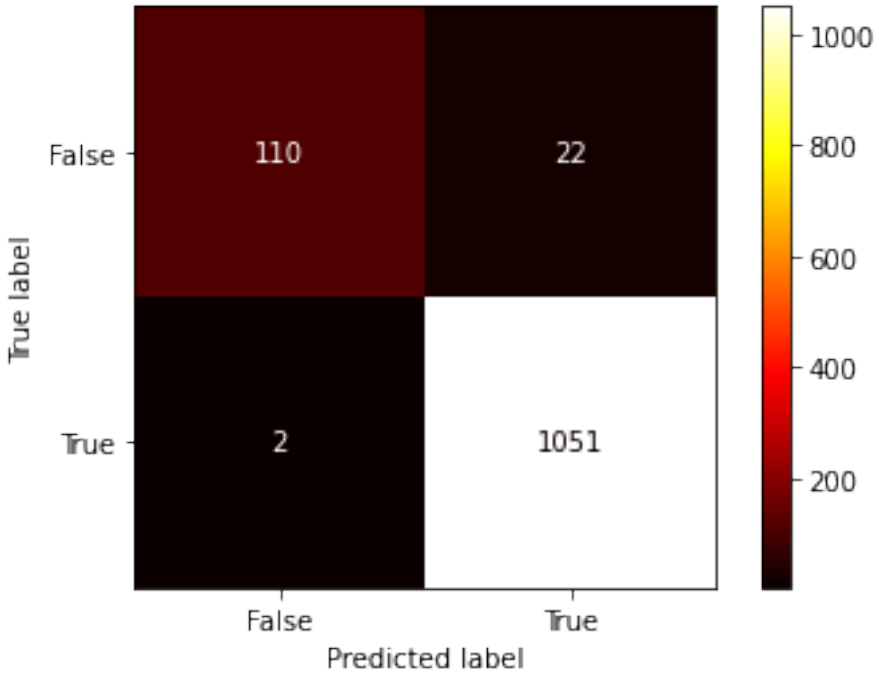

Fig. 3 Confusion Matrix for the Support Vector Machine Classification. Here, False Positive (Fake test results that claim to have found particular diseases or attributes is 2, True Positive (An examination outcome that accurately demonstrates the existence of a state or property) is 1051, False Negative (Misleading test results that claim specific conditions or attributes are absent) is 110, and True Negative (An accurate test result that shows the absence of a condition or trait) is 22.

which can be used for classification, regression, and other tasks. A random forest's categorization output is the majority of trees' classes. Regression jobs yield tree means or forecasts. Random Decision Forest corrects decision tree overfitting. Gradient-boosted trees outperform random forests. Data properties may affect performance. For Random Forest, the results are as

```
==========================
[ Random Decision Forest ]
==========================

Accuracy:  0.9789029535864979
Recall:  0.9971509971509972
```

Figure 4 represents the Precision Recall Curve, while Figure 5 shows the Confusion Matrix.

3. *k-Nearest Neighbours*: K-Nearest Neighbor is one of the most basic supervised learning-based non-parametric machine learning algorithms. the

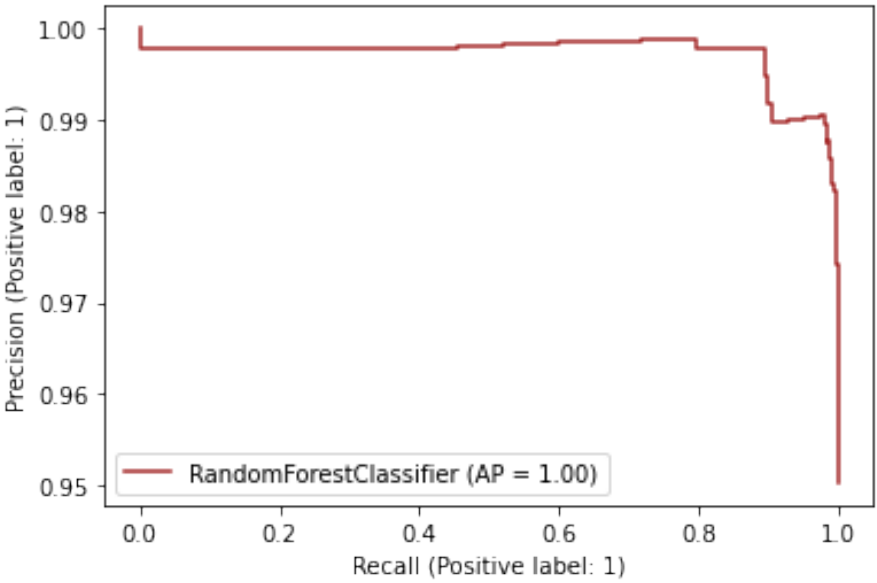

Fig. 4 Precision Recall Curve for the Random Decision Forest.

classification of new instances into categories that are most similar to existing categories, the storage of all pertinent data, and the assumption that new cases and data are comparable to old cases. As a result, using the K-NN approach, it is simple to classify new data into the appropriate categories. For k-Nearest Neighbours, the results are as

```
==========================
[ k - Nearest Neighbours ]
==========================

Accuracy:  0.9324894514767933
Recall:  0.9990503323836657
```

Figure 6 represents the Precision Recall Curve, while Figure 7 shows the Confusion Matrix.

4. *Gaussian Naive Bayes*: It's easy to create classifiers using the Naive Bayes method. A model that selects class labels from a finite set and represents class labels as a vector of feature values gives class labels to problem occurrences. Although there isn't one specific strategy for training these classifiers, there is a series of algorithms built on similar ideas. All naive Bayesian classifiers use the assumption that, given the class variables, the value of a single feature is independent of the value of another feature. For instance, an apple is a round, red fruit with a 10 centimetre diameter. Any potential correlations between the colour, roundness, and diameter variables are also evaluated, as well as the likelihood that the fruit is an

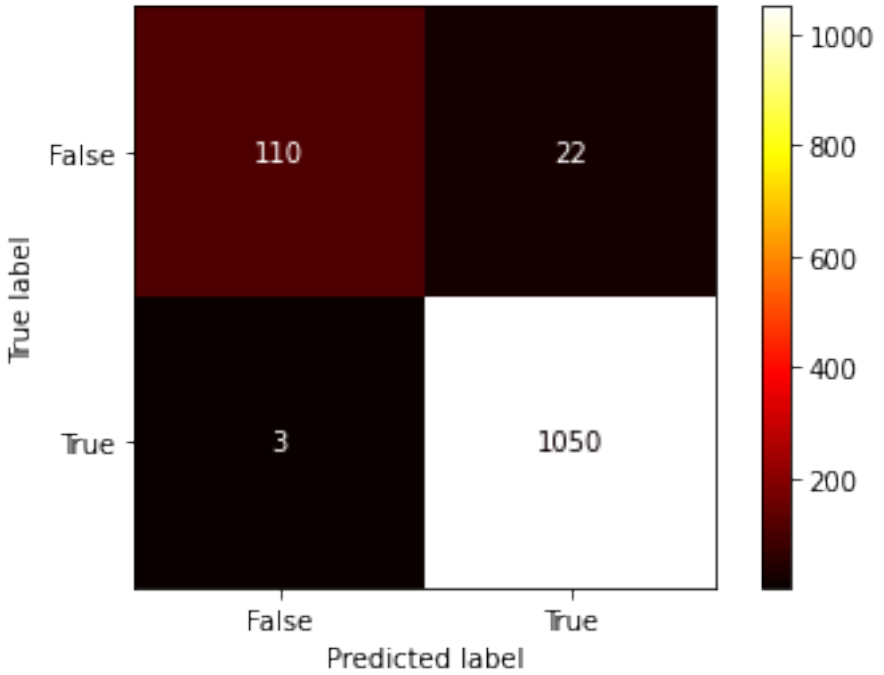

Fig. 5 Confusion Matrix for the Random Decision Forest Classification. Here, False Positive (Fake test results that claim to have found particular diseases or attributes is 3, True Positive (An examination outcome that accurately demonstrates the existence of a state or property) is 1050, False Negative (Misleading test results that claim specific conditions or attributes are absent) is 110, and True Negative (An accurate test result that shows the absence of a condition or trait) is 22.

apple, by a straightforward Bayesian classifier. Many practical applications employ maximum likelihood techniques for parameter estimation of simple Bayesian models. In other words, simple Bayesian models can be used without adopting or employing Bayesian probability. Despite their basic design and extremely simplified assumptions, naive Bayesian classifiers perform well in a range of difficult real-world circumstances. There are solid theoretical foundations for the ostensibly exceptional performance of straightforward Bayesian classifiers, according to a study of Bayesian classification challenges published in 2004. Bayesian classification, however, triumphs over other strategies like boosted trees and random forests in a comprehensive comparison with other classification methods in 2006. Naive Bayes' benefit is that it just needs a little amount of training data to estimate the classification parameters. For Gaussian Naive Bayes, the results are as

```
========================
[ Gaussian Naive Bayes ]
========================
```

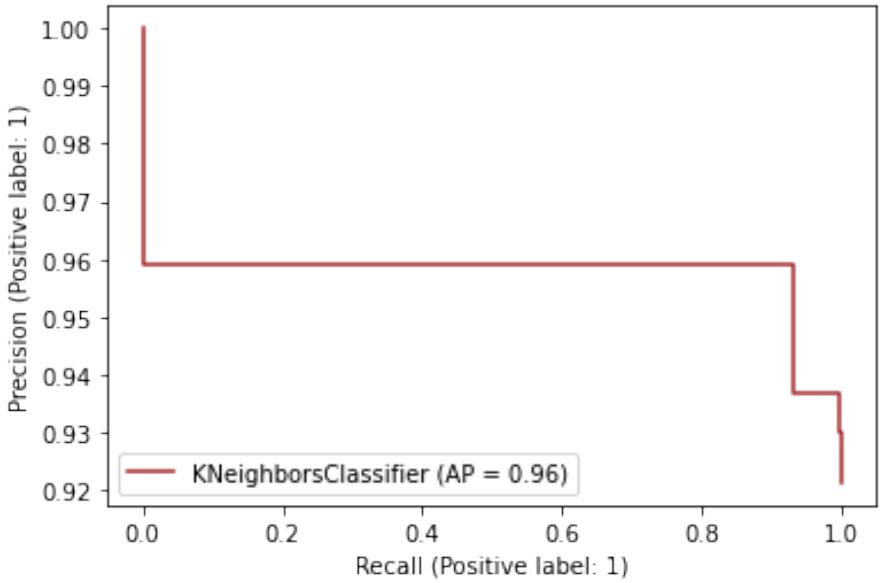

Fig. 6 Precision Recall Curve for the k-Nearest Neighbours.

```
Accuracy:  0.8514767932489451
Recall:  0.8499525166191833
```

Figure 8 represents the Precision Recall Curve, while Figure 9 shows the Confusion Matrix.

References

Bromham L, Dinnage R, Skirgård H, et al (2022) Global predictors of language endangerment and the future of linguistic diversity. Nature ecology & evolution 6(2):163–173

He H, Quan G, Zhu H, et al (2022) Evaluation modeling of highway collapse hazard based on rough set and support vector machine. Scientific reports 12(1):1–10

Houdé O, Tzourio-Mazoyer N (2003) Neural foundations of logical and mathematical cognition. Nature Reviews Neuroscience 4(6):507–514

Kaltenbrunner M, Hochegger R, Cichna-Markl M (2018) Sika deer (cervus nippon)-specific real-time pcr method to detect fraudulent labelling of meat and meat products. Scientific reports 8(1):1–11

Lordén G, Wozniak JM, Doré K, et al (2022) Enhanced activity of alzheimer disease-associated variant of protein kinase cα drives cognitive decline in a

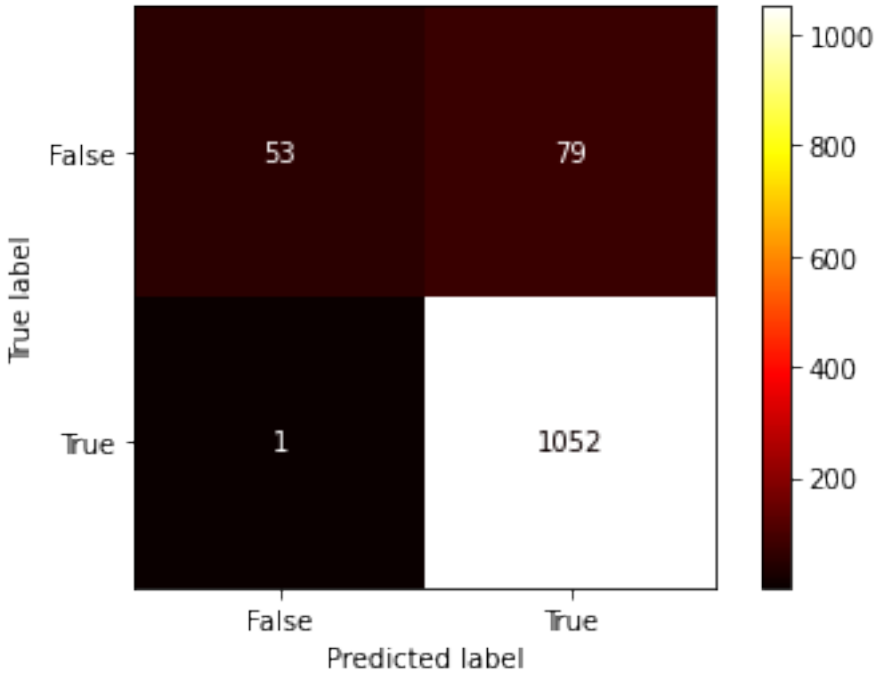

Fig. 7 Confusion Matrix for the k-Nearest Neighbours Classification. Here, False Positive (Fake test results that claim to have found particular diseases or attributes is 1, True Positive (An examination outcome that accurately demonstrates the existence of a state or property) is 1052, False Negative (Misleading test results that claim specific conditions or attributes are absent) is 53, and True Negative (An accurate test result that shows the absence of a condition or trait) is 79.

mouse model. Nature Communications 13(1):1–16

Lwoff A (1967) Principles of classification and nomenclature of viruses. Nature 215(5096):13–14

Maiorani A, Bateman JA, Liu C, et al (2022) Towards semiotically driven empirical studies of ballet as a communicative form. Humanities and Social Sciences Communications 9(1):1–18

Martinez-Morata I, Bostick BC, Conroy-Ben O, et al (2022) Nationwide geospatial analysis of county racial and ethnic composition and public drinking water arsenic and uranium. Nature Communications 13(1):1–12

Olsson H, Kartasalo K, Mulliqi N, et al (2022) Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nature Communications 13(1):1–10

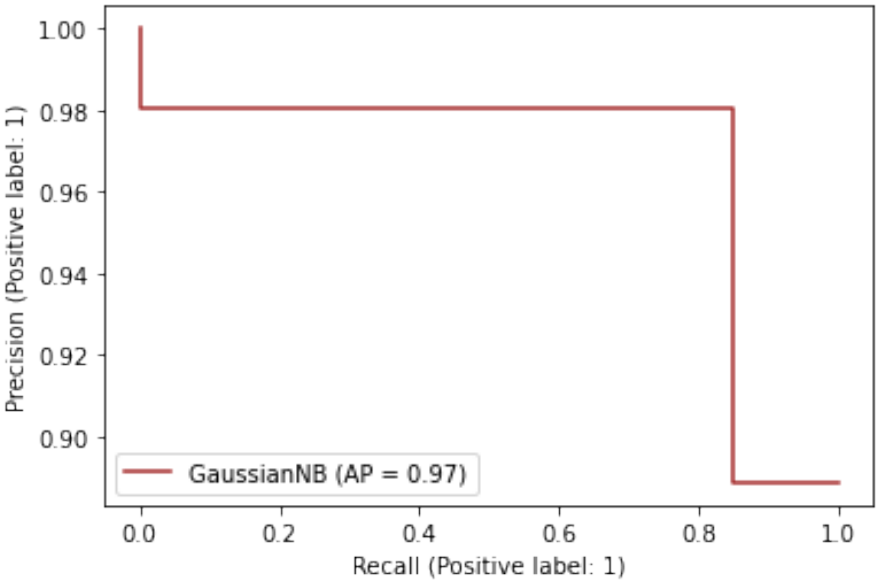

Fig. 8 Precision Recall Curve for the Gaussian Naive Bayes.

Pan Hy, He Sn, Zhang Tj, et al (2022) Application of an improved naive bayesian analysis for the identification of air leaks in boreholes in coal mines. Scientific Reports 12(1):1–15

Paramasivan K, Subburaj R, Jaiswal S, et al (2022) Empirical evidence of the impact of mobility on property crimes during the first two waves of the covid-19 pandemic. Humanities and Social Sciences Communications 9(1):1–14

Rafiq H, Aslam N, Aleem M, et al (2022) Andromalpack: enhancing the ml-based malware classification by detection and removal of repacked apps for android systems. Scientific Reports 12(1):1–18

Rajkomar A, Loreaux E, Liu Y, et al (2022) Deciphering clinical abbreviations with a privacy protecting machine learning system. Nature Communications 13(1):1–14

Ramezani M, Feizi-Derakhshi MR, Balafar MA (2022) Text-based automatic personality prediction using kgrat-net; a knowledge graph attention network classifier. arXiv preprint arXiv:220513780

Rotaru V, Huang Y, Li T, et al (2022) Event-level prediction of urban crime reveals a signature of enforcement bias in us cities. Nature human behaviour 6(8):1056–1068

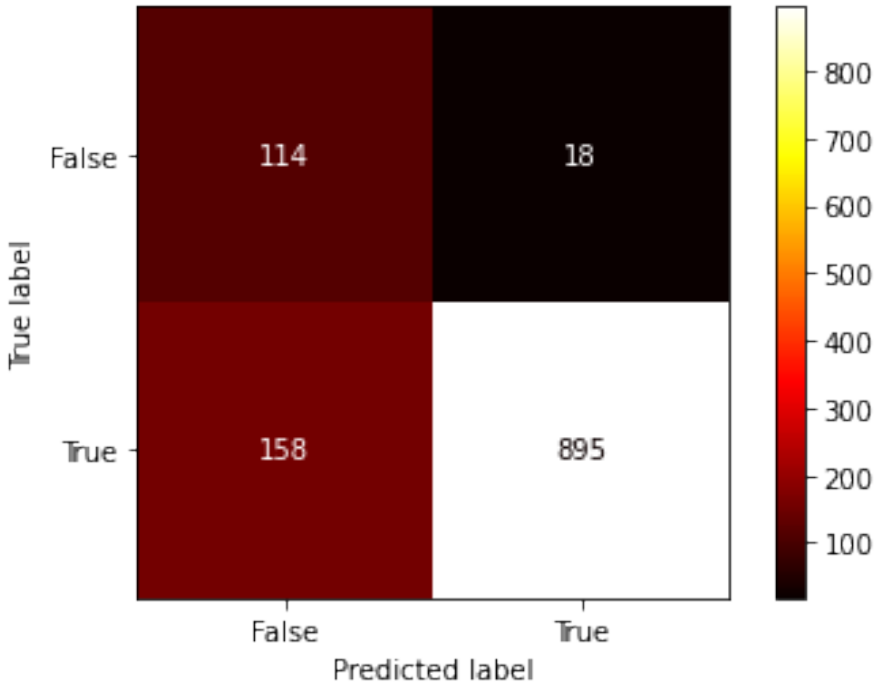

Fig. 9 Confusion Matrix for the Gaussian Naive Bayes Classification. Here, False Positive (Fake test results that claim to have found particular diseases or attributes is 158, True Positive (An examination outcome that accurately demonstrates the existence of a state or property) is 895, False Negative (Misleading test results that claim specific conditions or attributes are absent) is 114, and True Negative (An accurate test result that shows the absence of a condition or trait) is 18.

Roth ZN, Kay K, Merriam EP (2022) Natural scene sampling reveals reliable coarse-scale orientation tuning in human v1. Nature communications 13(1):1–13

Russo AG, Ciarlo A, Ponticorvo S, et al (2022) Explaining neural activity in human listeners with deep learning via natural language processing of narrative text. Scientific Reports 12(1):1–9

Uddin S, Haque I, Lu H, et al (2022) Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction. Scientific Reports 12(1):1–11

Wahab A, Tayara H, Xuan Z, et al (2021) Dna sequences performs as natural language processing by exploiting deep learning algorithm for the identification of n4-methylcytosine. Scientific Reports 11(1):1–9

Ward EN, Hecker L, Christensen CN, et al (2022) Machine learning assisted interferometric structured illumination microscopy for dynamic biological imaging. Nature Communications 13(1):1–10

Zhu X, Wang Y, Li J (2022a) What drives reputational risk? evidence from textual risk disclosures in financial statements. Humanities and Social Sciences Communications 9(1):1–15

Zhu Y, Xu M, Lu J, et al (2022b) Distinct spatiotemporal patterns of syntactic and semantic processing in human inferior frontal gyrus. Nature Human Behaviour pp 1–8