International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 7, Issue 11, pp 42-49, November-2023
https://doi.org/10.47001/IRJIET/2023.711007

"DOCU SAFE" Secure Data Management System Using Machine Learning

[1]R.M.P.S.Rajapaksha, [2]S.A.A.T.S.Sooriyamali, [3]L.D.C.Jayarathna, [4]H.S.D De Silva, [5]Ms.Chethana Liyanapathirana [6]Dr. Lakmal Rupasinghe

[1,2,3,4,5,6]Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Abstract **- A data management system is crucial to businesses as it efficiently organizes, secures, and leverages data, enabling informed decision-making, streamlined operations, and improved competitiveness. Machine learning enhances data management by automating insights extraction, predictive analytics, and pattern recognition, optimizing data utilization and driving informed business strategies. DOCU SAFE is a web base secure data management system that uses machine learning and NLP to enable organizations to store, manage, and analyze their data securely. The system addresses the challenges associated with data uploading, moving, and processing, including size/storage/volume of data, inconsistency and variety of data, formatting of data, security of data, analyzing and mapping of data, and cost management issues. DOCU SAFE provides a scalable solution for data management, enabling organizations to handle large volumes of data efficiently. The system uses machine learning algorithms to ensure that data is consistent, accurate, and secure, providing organizations with insights that can be used to make informed decisions. The data we enter here can be directed to four categories of classification, hidden and highlight, encryption, and hygiene solutions according to our needs. They currently have many tools in this regard, but there are problems with them. In addition, there are separate tools for data classification, data hidden and highlight, data encryption, data hygiene solutions, and there is no possibility to do everything with one tool/system. Here, all the above issues are the provided by same tool. Overall, DOCU SAFE is an effective solution for organizations that handle sensitive data and need a secure and efficient data management system.**

Keywords: machine learning algorithms, data hygiene, data classification, data highlight and hidden, data encryption.

I. INTRODUCTION

Data management has become an essential part of any organization. It involves the processes of collecting, storing, processing, and retrieving data to ensure the smooth running of an organization's operations. However, data management is a challenging task due to the volume, variety, and complexity of data. The need for secure data management systems has become increasingly important due to the increase in cyber threats and data breaches. DOCU SAFE is a secure data management system that uses machine learning and NLP to address the challenges faced in data management. This paper review aims to provide an overview of DOCU SAFE, its features, and how it addresses the problems associated with data uploading, moving, and processing in the Data classification, Data hygiene , Data Highlight and hidden and data encryption methods. Specially this system can be used for both image and text input and You can get voice activated data cleaning and real time data backup facilities which are not available in other tools.

II. LITERATURE REVIEW

They currently have many tools in this regard, but there are problems with data uploading, data moving and data processing. In addition, there are separate tools for data classification, data hidden and highlight, data encryption, data hygiene solutions, and there is no possibility to do everything with one tool/system.

A) Data Hygiene

Data hygiene, also known as data quality, refers to the process of maintaining clean, accurate, and consistent data. Machine learning can play a significant role in enhancing data hygiene solutions .Data hygiene is a critical aspect of data management, ensuring that datasets are accurate, consistent, and reliable. Over the years, various techniques and approaches have been developed to address data hygiene issues. This literature review provides an overview of key data hygiene techniques as discussed in relevant research papers. The various techniques used in Data hygiene solutions are described below [1].

1) Data Quality Tools: These are specialized software tools designed to assess, monitor, and improve the quality of data. They offer features like data profiling, data cleansing, data validation, and data enrichment. Examples include Informatica Data Quality, Talend Data Quality, Trifacta, and IBM InfoSphere QualityStage.

ISSN (online): 2581-3048
Volume 7, Issue 11, pp 42-49, November-2023
https://doi.org/10.47001/IRJIET/2023.711007

2) Machine Learning and AI: Machine learning algorithms are used to detect anomalies, outliers, and patterns in data that could indicate data quality issues. AI-powered models can identify discrepancies and help predict missing or erroneous data.
3) Natural Language Processing (NLP): NLP is used to process and understand unstructured text data. It can assist in text normalization, entity recognition, sentiment analysis, and extracting metadata from textual documents.
4) Automated Data Cleansing: Automated data cleansing tools use algorithms to identify and correct errors in data. These tools can handle tasks like removing duplicate records, standardizing formats, and validating data against predefined rules.
5) Blockchain and Distributed Ledgers: Blockchain technology is sometimes used to create an immutable and tamper-proof record of data transactions. This can enhance data integrity and transparency.

We developed data hygiene solution phase using Machine learning and NLP.[2] Here are some common existing system problems and how machine learning can help address them.[3]

1) Data Duplication: Problem: Duplicate records can lead to inaccurate analytics, wasted storage, and confusion. Solution with Machine Learning: Machine learning models can be trained to identify duplicate records by analyzing various data attributes and similarity metrics. These models can help consolidate duplicate records and prevent their entry.

2) Inconsistent Data: Problem: Inconsistent formatting, misspellings, and different units can lead to confusion and errors during analysis. Solution with Machine Learning: Machine learning algorithms, such as natural language processing (NLP) models, can help standardize and normalize data by identifying patterns, correcting misspellings, and converting units.

3) Missing Data: Problem: Incomplete data can lead to biased or incomplete analyses and inaccurate results. Solution with Machine Learning: Machine learning techniques like imputation can predict missing values based on the patterns and relationships within the existing data. This helps fill in gaps and maintain the integrity of the dataset.

4) Outliers: Problem: Outliers can skew statistical analyses and modeling results. Solution with Machine Learning: Machine learning algorithms can automatically detect outliers by analyzing data distributions and identifying data points that deviate significantly from the norm. These outliers can then be reviewed for accuracy or treated appropriately.

5) Data Validation: Problem: Incorrect or invalid data can infiltrate the system due to human errors or malicious intent. Solution with Machine Learning: Machine learning models can be trained to validate incoming data by comparing it to historical patterns, flagging anomalies, and ensuring that data adheres to predefined rules or constraints.

B) Data Classification

In the era of information explosion, the proliferation of digital content has led to an unprecedented need for efficient and accurate methods of data classification. Text data and image data classification involve the process of categorizing and labeling data into confidential or non-confidential categories, spanning from natural language processing to computer vision. This research paper aims to delve into the intricate landscape of text and image data classification methodologies, utilizing four machine learning algorithms. In this system used random forest, support vector machine, logistic regression, and decision tree for data classification. This literature review provides an overview of key data classification as discussed in relevant research papers.

1) Optical Character Recognition (OCR) technology: This technology used to convert images to text, which analyzed and classified using NLP and machine learning algorithms [4].

2) Natural Language Processing (NLP) techniques: These techniques used to extract features from the text data, identify patterns, and classify the data into different categories based on its content.[5]

3) Machine Learning algorithms: These algorithms used to train the classification model, identify patterns in the data, and classify the data into different categories based on its content and level of risk.

i. Logistic Regression: Logistic Regression (LR) is a well-known statistical classification method for modeling dichotomous (binary) data. [6]
ii. Support Vector Machines: Support Vector Machines are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. [7]
iii. Random Forest: Random Forest developed by Leo Bierman is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. [8]
iv. Decision Trees: Decision Trees embody a supervised classification approach. The idea came from the ordinary tree structure which is made-up of a root and nodes. [9]

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 11, pp 42-49, November-2023

https://doi.org/10.47001/IRJIET/2023.711007

4) Python programming language: Python is a popular programming language for data science and machine learning applications. It has many libraries and frameworks that can be used for developing data classification tool, such as TensorFlow, Scikit-learn, and NLTK.

5) MySQL: MySQL is a popular open-source relational database management system (RDBMS) that serves as a server for web-based applications, especially those that run on websites and in cloud environments.

Overall, the literature reviewed that machine learning algorithms, NLP techniques, and OCR technology effective tools for data classification. The research project aimed to develop a new data classification tool that utilizes OCR technology and evaluates the effectiveness of different ML algorithms and NLP techniques in classifying and processing text data based on their risk levels.

C) Data Highlight & Hidden

To protect the confidentiality and security of data, it can be "hidden" within other data using various approaches and procedures. The subject of computer science has seen the development and study of a number of different data-hiding strategies. Embedding secret information in seemingly innocuous cover media is a popular practice known as steganography. It is possible to alter the cover medium without damaging the steganographic message thanks to steganography systems that seek out duplicate data [10]. Steganographic systems are classified by their capacity, or the amount of data they can conceal, and their security, or the degree to which an eavesdropper will be unable to decipher the data they conceal [10]The information that is disguised using steganography is often vulnerable, despite the fact that the goal of the technique is high security and capacity [10].Watermarking is another method, with its primary focus being on attaining a high level of robustness [10], making it impossible to remove a watermark without affecting the quality of the data object. Digital media are often watermarked as a means of protecting its ownership and authenticity. When providing a database to external parties, association rule concealing can be utilized to protect the privacy of potentially sensitive association rules [11].The difficulty lies in keeping the database and the results of association rule mining usable while concealing the knowledge derived from sensitive association rules [12]. Steganography is the study of strategies for concealing data in non-obvious places; digital photographs are particularly well-suited for this purpose because of their extensive use on the web and enormous file sizes [13]. Steganography involves hiding information in plain sight by replacing unused or underutilized data in common digital files including images, audio, and text [12]. Data hiding methods try to strike a balance between secrecy and practicality by keeping the hidden data safe while keeping the cover media operational. Computer science, cyber security, and data mining are just a few of the fields that could benefit from these methods

D) Data Encryption

In data management systems, encryption and decryption are typically applied at different levels, including data at rest (stored data), data in transit (during communication or transmission), and data in use (during processing or analysis). Various encryption algorithms and techniques, such as symmetric encryption, asymmetric encryption, hashing, and digital signatures, are employed to achieve different security objectives. By implementing encryption and decryption mechanisms within data management systems, organizations can mitigate the risks associated with data breaches, unauthorized access, and data theft. These techniques provide an additional layer of protection, ensuring the confidentiality, privacy, and integrity of sensitive data, thereby instilling trust and confidence in the data management processes. There are some machine learning technologies that used for the encryption and decryption.

k-Nearest Neighbors (k-NN): The k-Nearest Neighbors algorithm is a non-parametric classification algorithm widely used in machine learning. In the context of encryption and decryption, k-NN can be applied for key generation, substitution ciphers, stream ciphers, and image encryption. k-NN involves finding the k closest neighbors to a given data point based on a distance metric, which can be utilized in various encryption and decryption processes.[15]

Recurrent Neural Networks (RNNs): Recurrent Neural Networks are a type of neural network architecture designed to process sequential data by incorporating feedback connections. RNNs have been explored for encryption and decryption tasks, particularly in sequence encryption and decryption as well as language-based encryption and decryption techniques. RNNs can learn patterns and dependencies in sequential data, making them suitable for tasks involving encryption and decryption of sequential information.[16]

Ensemble Methods: Ensemble Methods involve combining multiple individual models to improve overall performance and generalization. In the context of encryption and decryption, ensemble methods can be applied to enhance the accuracy of encryption or decryption processes, such as key recovery and cryptanalysis. Ensemble methods can leverage the diversity of multiple models to address the limitations of individual models, leading to improved security and robustness in encryption and decryption tasks.[17]

International Research Journal of Innovations in Engineering and Technology (IRJIET)

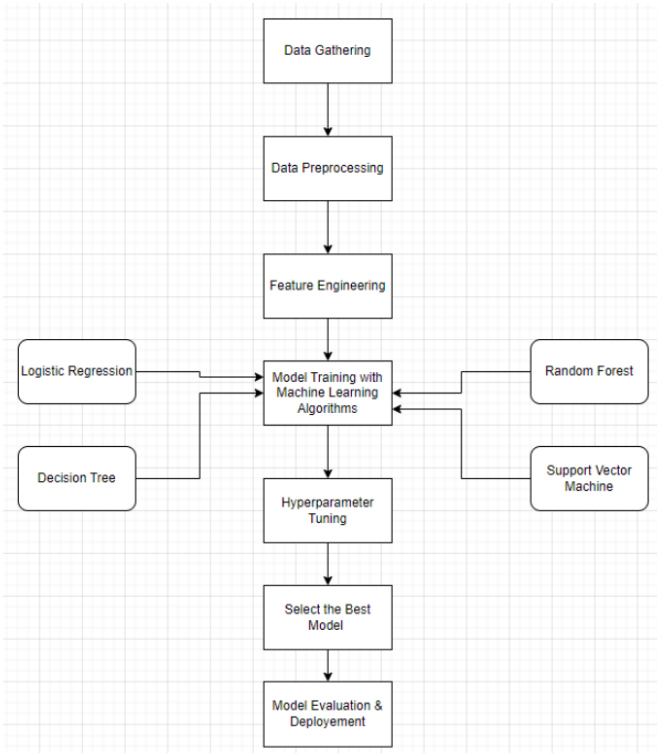
ISSN (online): 2581-3048

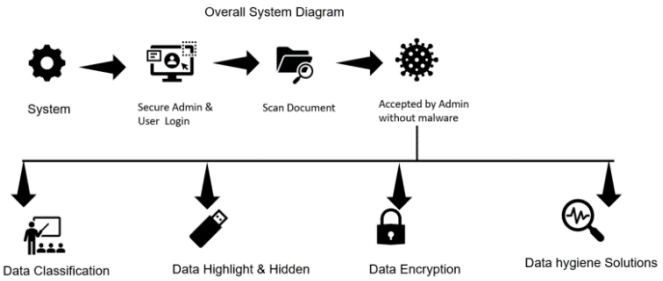
Volume 7, Issue 11, pp 42-49, November-2023

https://doi.org/10.47001/IRJIET/2023.711007

Clustering Techniques: Clustering techniques aim to group similar data points together based on certain criteria or similarity measures. In the context of encryption and decryption, clustering techniques can be used to partition data or identify clusters that share similar properties for secure data storage and retrieval. Clustering techniques can aid in data management, organization, and secure encryption and decryption processes based on the similarity of data points.[18]

III. METHODOLOGY

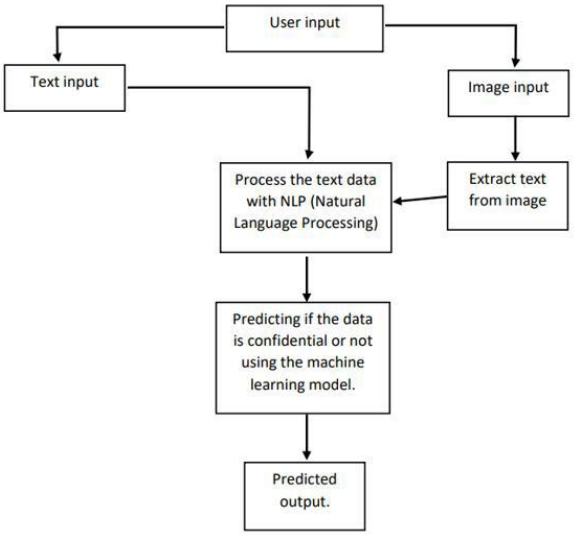




In methodology, we can log in as an admin or as a user. After logging in as a user and entering the data, it will be analyzed by a malware detection system and if there is no malware, the admin panel will approve the following data classification, data highlight and hidden, data encryption and data hygiene solutions services can be included. After that we have divided the whole approach into four phases and the description of each processes are clearly defined using the related algorithms and procedures.

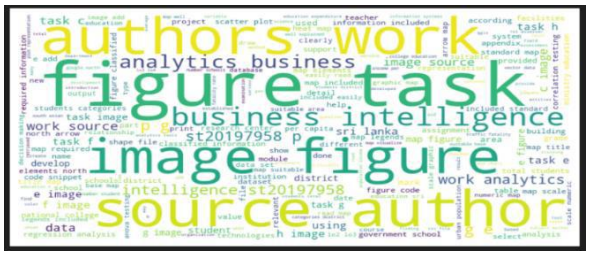
The technique utilized to train the machine learning model is what makes this unique. The model is most accurately trained via automation, which is the least time-consuming of the four training techniques, which are logistic regression, random forest, decision trees, and support vector machines.

Phase 1 – Data Classification



Automatically identifying and data classification based on patterns and trends in the data is distinct from existing systems. To extract text from images, OCR (Optical Character Recognition) can be employed. This involves combining machine learning and image processing techniques (NLP). After predicted if the text data is confidential or nonconfidential using four machine learning algorithms. In here used logistic regression, Support vector machine, random forest, and decision tree machine learning algorithms. Predicted most suitable output using four machine learning algorithms together. Predicted output was confidential or non-confidential.

In this context, displayed word-cloud reading all pdf and images. Word-cloud is collection or cluster of word depicted in different sizes.



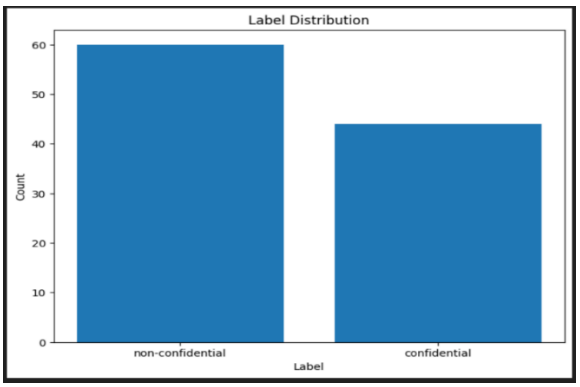
In this context, data is labeled as confidential and nonconfidential. Notably, labeled can even be classified within image files if required.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

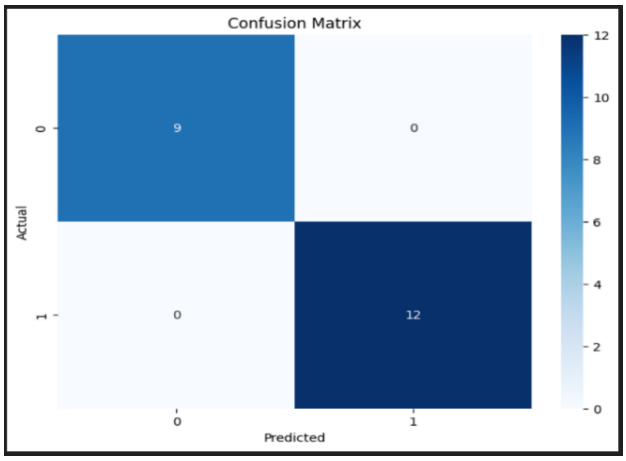
Volume 7, Issue 11, pp 42-49, November-2023

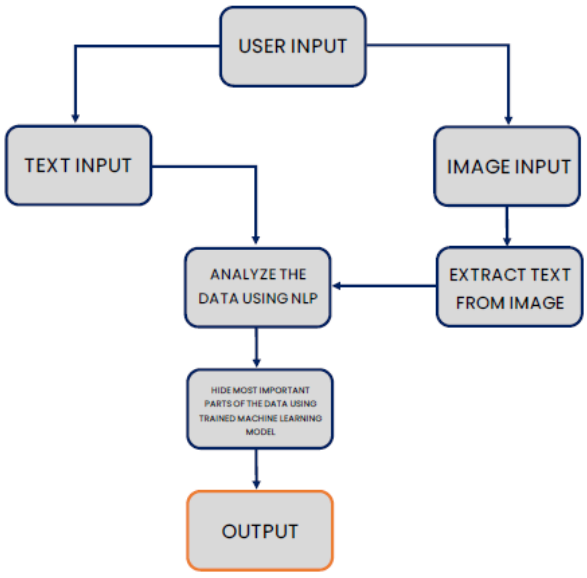
https://doi.org/10.47001/IRJIET/2023.711007



In this context, displayed confusion matrix reading all text data. Confusion matrix represents the prediction summary in matrix form. It shows how many predictions are correct and incorrect per class. It helps in understanding the classes that are being confused by model as other class.

Phase 2 – Data highlight and hidden





Methodology for Data Hiding using Text and Image Input

1. Input Data Collection:

- Allow users to input data in two formats: text input and image input.
- For text input, users can directly enter the data as text.
- For image input, extract text from images using Optical Character Recognition (OCR) techniques.

2. Data Analysis using NLP:

- Perform Natural Language Processing (NLP) techniques to analyze the collected data.
- NLP techniques may include tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, etc.
- These techniques help in understanding the structure and meaning of the text data.

3. Identify Important Data:

- Based on the analysis from the previous step, identify the most important parts of the data.
- This can be done using various criteria such as keywords, sentiment scores, named entities, etc.
- The identified important data will be used for data hiding.

4. Training a Machine Learning Model:

- Use the identified important data as the training dataset for a machine learning model.
- The machine learning model should be chosen based on the specific requirements of data hiding.
- This model will learn to hide the important parts of the data based on the input data and the desired output.

5. Data Hiding Process:

- Apply the trained machine learning model to the input data.
- The model will hide the most important parts of the data based on its learned patterns.
- This process can involve various techniques such as data obfuscation, encryption, steganography, etc.
- The output of this process will be the hidden data, which can be stored or transmitted securely.

6. Output Evaluation:

- Evaluate the effectiveness of the data hiding process.
- This can be done by comparing the original input data with the hidden data.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

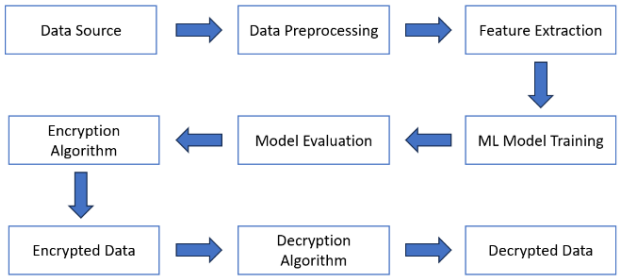
Volume 7, Issue 11, pp 42-49, November-2023

https://doi.org/10.47001/IRJIET/2023.711007

- Measure the accuracy of the hidden data in terms of the ability to protect the most important parts of the data.

The highlight phase also happens in this way.

Phase 3 -Data encryption



Briefly explain above steps;

Data source: Select a suitable data source for the encryption and decryption system.

Data preprocessing: Preprocess the data by cleaning it, handling missing values, and transforming it into a suitable format for further processing.

Feature extraction: Extract relevant features from the preprocessed data to capture important information for encryption and decryption.

ML model training: Train a machine learning model using the extracted features and a labeled training dataset.

Model evaluation: Evaluate the performance of the trained model using evaluation metrics and test datasets to assess its effectiveness.

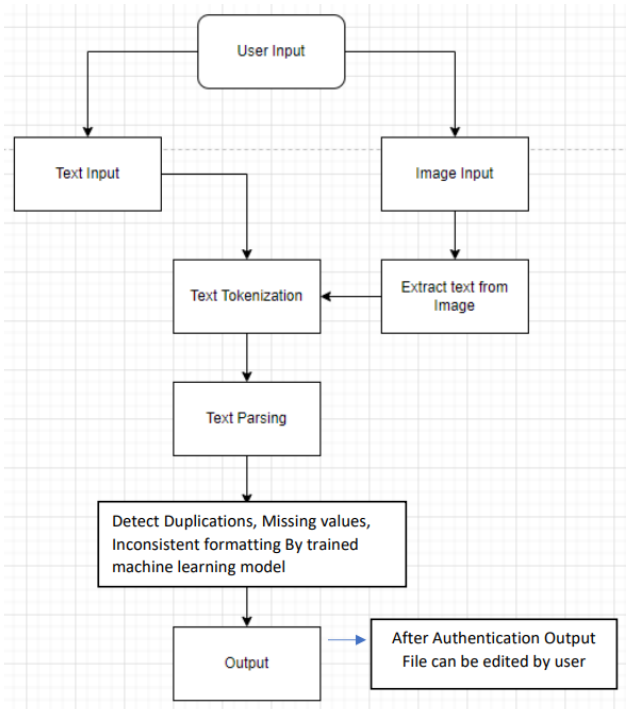
Encryption algorithm: Use AES and DES encryption algorithm based on either existing encryption techniques.

Encrypted data: I applied the encryption algorithm to the input data, transforming it into an encrypted form to protect its confidentiality.

Decryption algorithm: Use the same decryption algorithm that can reverse the encryption process.

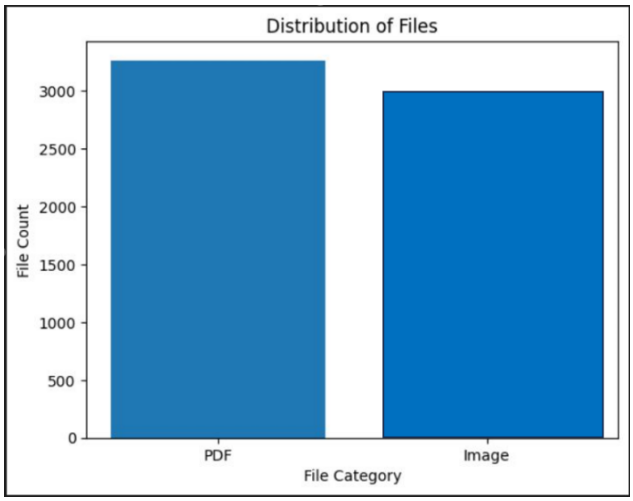
Decrypted data: Apply the decryption algorithm to the encrypted data, converting it back to its original form for further analysis or usage.

Phase 4 – Data Hygiene Solution



- Using automatically identify and flag data hygiene issues based on patterns and trends in the data , Different from existing systems.
- To extract texts from images we can use OCR (Optical Character Recognition). It combines machine learning and image processing. We can utilize deep learning techniques like CNNs for picture processing.

Here we can detect data duplication, detect missing values and inconsistent formatting. When finding duplicates, several files can be entered at the same time, and the special thing is that if necessary, we can identify duplications even in image files.



International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 11, pp 42-49, November-2023

https://doi.org/10.47001/IRJIET/2023.711007

IV. RESULTS

Data hygiene results

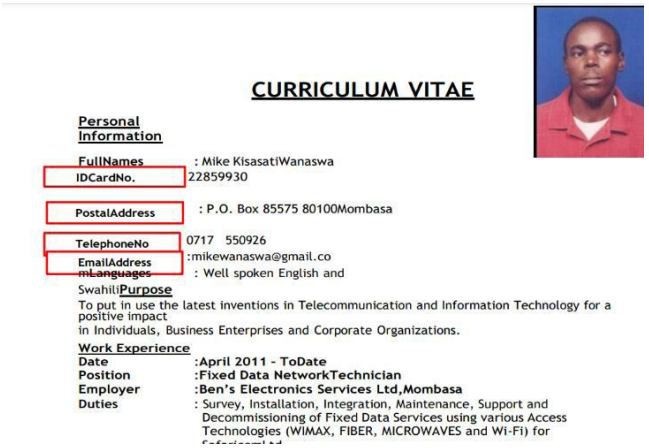
The below explanation of the results has done by the us, for the demonstration of the values after the hygiene phase. Here we can see an inconsistent formatting detection note and we can see what are that lines in that file.

```
Prediction: ['fix_formatting']
```

```
from sklearn.metrics import confusion_matrix
```

Data classification results

The following explanation of the results was prepared by us to demonstrate the values after the classification. In this section, predicted as confidential or non-confidential.



CURRICULUM VITAE

Personal Information

FullNames : Mike KisasatiWanaswa
IDCardNo. 22859930
PostalAddress : P.O. Box 85575 80100Mombasa
TelephoneNo 0717 550926
EmailAddress :mikewanaswa@gmail.co
mLanguages : Well spoken English and Swahili**Purpose**
To put in use the latest inventions in Telecommunication and Information Technology for a positive impact
in Individuals, Business Enterprises and Corporate Organizations.

Work Experience
Date :April 2011 - ToDate
Position :Fixed Data NetworkTechnician
Employer :Ben's Electronics Services Ltd,Mombasa
Duties : Survey, Installation, Integration, Maintenance, Support and Decommissioning of Fixed Data Services using various Access Technologies (WIMAX, FIBER, MICROWAVES and Wi-Fi) for

V. CONCLUSION

In this paper, authors have factually evaluated the results of data classification, data highlight and hidden, data hygiene solutions and encryption processed by machine learning and NLP techniques. Here, the authors' attention has been drawn to the fact that the OCR technology of this system can be used as an image input and the security of the data. When receiving all the outputs, one must provide his credentials once again. Thus, protection from man-of-the-middle attacks has been compromised. Many facilities have been provided in the same system and more attention has been paid to the convenience of the users.

REFERENCES

[1] F. Ridzuan and W. M. N. Wan Zainon, "A review on data cleansing methods for big data," Procedia Comput. Sci., vol. 161, pp. 731–738, 2019.

[2] J.-S. Hwang, S.-D. Mun, T.-J. Kim, G.-W. Oh, Y.-S. Sim, and S. J. Chang, "Development of data cleaning and integration algorithm for asset management of power system," Energies, vol. 15, no. 5, p. 1616, 2022.

[3] J. Clements, "Challenges involved in data cleansing & current approaches," Managed Outsource Solutions, 24-Aug-2022. [Online]. Available: https://www.managedoutsource.com/blog/data-cleansing-challenges-current-approaches/. [Accessed: 17-Aug-2023].

[4] I.P. Preethi P. Bhathh, "OPTICAL CHARACTER RECOGNITION USING DEEP LEARNING – A TECHNICAL REVIEW," India, 2018.

[5] A.K.K.K.S.S. Diksha Khurana, "Natural Language Processing: State of The Art, Current Trends and Challenges," India, 2022.

[6] A.B. Musa, "Logistic Regression Classification for Uncertain Data," China, 2014.

[7] D. K. Shivasthava, "DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE," India, 2010.

[8] R. K. N. A. I. M. Jehad Ali, "Random Forests and Decision Trees," Pakistan, 2012.

[9] B. T. Jijo, "Classification Based on Decision Tree Algorithm for Machine Learning," Iraq, 2021.

[10] Honeyman, "Hide and seek: an introduction to steganography," IEEE Secur. Privacy, vol. Volume: 1, no. IEEE, pp. 32 - 44, may-june 2003.

[11] K. Pathak, "A Survey of Data Hiding Techniques," International Journal of Computer Applications (0975 - 8887), vol. Volume 182, no. Foundation of Computer Science, No.28, November 2018.

[12] V. M. Manikandan, "A Reversible Data Hiding Scheme through Encryption using Rotated Stream Cipher," Computer Science, p. 22(2), 2021-04-15.

[13] V. M. Manikandan, "A Novel Reversible Data Hiding Scheme that Provides Image Encryption," Journal of Image and Graphics, vol. 6(1), pp. 64-68, 2018.

[14] Researchgate.net. [Online]. Available: https://www.researchgate.net/publication/355905228_Secure_KNN_Classification_Scheme_Based_on_Homomorphic_Encryption_for_Cyberspace. [Accessed: 19-Aug-2023].

[15] Y. Tong and I. Tien, "Time-series prediction in nodal networks using recurrent neural networks and a pairwise-gated recurrent unit approach," ASCE ASME J. Risk Uncertain. Eng. Syst. A Civ. Eng., vol. 8, no. 2, 2022.

[16] A.Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," J. King Saud Univ. - Comput. Inf. Sci., vol. 35, no. 2, pp. 757–774, 2023.

[17] Researchgate.net. [Online]. Available: https://www.researchgate.net/publication/265077297_

ISSN (online): 2581-3048
Volume 7, Issue 11, pp 42-49, November-2023
https://doi.org/10.47001/IRJIET/2023.711007

RESEARCH_PAPER_ON_CLUSTER_TECHNIQUES_OF_DATA_VARIATIONS. [Accessed: 19-Aug-2023].

Citation of this Article:

R.M.P.S.Rajapaksha, S.A.A.T.S.Sooriyamali, L.D.C.Jayarathna, H.S.D De Silva, Ms.Chethana Liyanapathirana, Dr.Lakmal Rupasinghe, ""DOCU SAFE" Secure Data Management System Using Machine Learning" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 11, pp 42-49, November 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.711007

****

