**Data Leakage Prevention (DLP) : Develop AI models to detect and prevent unauthorized access and transfer of sensitive data**

**A PROJECT REPORT**

**Submitted by**

**SHANTHRAJ .K**

*in partial fulfillment for the award of the degree*

*of*

**Bachelor of Computer Science (H)**

*in*

**COMPUTER SCIENCE**



**School of Computer Science and Engineering**

**RV University**

**RV Vidyaniketan,8th Mile, Mysuru Road, Bengaluru, Karnataka, India - 562112**

**JUNE 2024**

**DECLARATION**

I,**SHANTHRAJ.K (1RVU23BSC132),** student third semester **Bachelor of Computer Science (H)**  **& Engineering,** at School of Computer Science and Engineering, **RV University,** hereby declare that the project work titled “**Data Leakage Prevention (DLP) : Develop AI models to detect and prevent unauthorized access and transfer of sensitive data”**has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Computer Science (H) & Engineering** during the academic year **2023-2024**. Further, the matter presented in the project has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Name: SHANTHRAJ.K Signature:

USN: 1RVU23BSC132

Place:BANGLORE

Date:20/12/2024

ii



**School of Computer Science and Engineering**

RV University

RV Vidyaniketan,8th Mile, Mysuru Road, Bengaluru, Karnataka, India - 562112

**CERTIFICATE**

This is to certify that the project work titled “**Data Leakage Prevention (DLP)”**  is performed by SHANTHRAJ .K (1RVU23BSC132)**,** a bonafide students of Bachelor of Technology at the School of Computer Science and Engineering, RV university, Bangaluru in partial fulfillment for the award of degree Bachelor of Technology in Computer Science & Engineering , during the Academic year **2023-2024**.

**Prof. Shivakumar D Dr. Vidya M J**  **Dr. G Shobha**

**Guide**

Ph.D Scholar Head of the Department Dean

SOCSE SOCSE SOCSE

RV University RV University RV University

Date: Date: Date:

Name of the Examiner Signature of Examiner

1.

2.

iii

**ACKNOWLEDGEMENT**

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to the School of Computer Science and Engineering, RV University, for providing us with a great opportunity to pursue our Bachelor’s Degree in this institution.

In particular we would like to thank Dr. G Shobha, Dean, School of Computer Science and Engineering, RV University, for his constant encouragement and expert advice.

It is a matter of immense pleasure to express our sincere thanks to Dr. Vidya M J Dr. G Shobha, Program Head , Computer Science & Engineering University, for providing right academic guidance that made our task possible.

We would like to thank our guide shivakumar D. Dept. of Computer Science & Engineering, RV University, for sparing his valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in completing the Project work successfully.

|  |  |
| --- | --- |
| Date: 20/12/2024 | Shanthraj .K |
| Place:Banglore | 1RVU23BSC132 |
|  | SEC “C” |

iv

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **DAATA LEKAGE PREVENTION(DLP)** |  |
|  | **ABSTRACT** | **vi** |
|  | **LIST OF TABLES** | **vii** |
|  | **LIST OF FIGURES** | **viii** |
|  | **LIST OF SYMBOLS AND ABBREVIATIONS** | **x** |
| **1.0** | **INTRODUCTION** | **1** |
| **2.0** | **RELATED WORK** | **2** |
| **3.0** | **METHODOLOGY** | **3** |
|  | 3.1 Sub section | **3** |
|  | 3.2 Sub section | **4** |
|  | 3.3Sub section | **4** |
| **4.0** | **IMPLEMENTATION** | **6** |
|  | 4.1 Sub section | **6** |
|  | 4.2 Sub section | **6** |
|  | 4.3 Sub section | **7** |
|  | 4.4 Sub section | **8** |
| **5.0** | **RESULT AND DISCUSSION** | **9** |
| **6.0** | **CONCLUSION** | **13** |
| **7.0** | **FUTURE SCOPE** | **14** |
|  | **REFERENCES** | **15** |
|  | **APPENDIX** | **16** |

**v**

**ABSTRACT**

**Abstract**

The rapid growth of digital data and its widespread use in various industries, including insurance, has significantly increased the risk of data breaches and unauthorized data transfers. This project focuses on developing and implementing Artificial Intelligence (AI) models for **Data Leakage Prevention (DLP)** to safeguard sensitive insurance data from unauthorized access and exfiltration.

Using the insurance dataset as a base, the project employs advanced machine learning and deep learning algorithms to detect anomalies, classify sensitive information, and prevent potential data leakages. The dataset is preprocessed to identify key attributes, such as customer details, policy information, and claim data, which are highly sensitive and require protection. Techniques such as Natural Language Processing (NLP) are utilized to analyze textual data for sensitive information, while predictive modeling helps identify unusual activities indicative of potential data leaks.

The models are trained and validated using k-fold cross-validation to ensure robustness and avoid overfitting. Performance metrics, including accuracy, precision, recall, and F1-score, are employed to evaluate the effectiveness of the developed solutions. Hyperparameter tuning is conducted to optimize model performance further. Additionally, anomaly detection methods, like Isolation Forest and Autoencoders, are integrated to enhance the identification of suspicious behaviors.

This project aims to create a comprehensive framework for detecting and preventing data leakage in the insurance domain, contributing to a safer data management ecosystem. The results demonstrate the potential of AI in mitigating cybersecurity risks and emphasize the importance of continuous improvement in DLP systems to address evolving threats.

**vi**

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Title** | **Page No.** |
| Table 3.1 | Performance Metrics of Models | 4 |

Vii

**LIST OF FIGURES AND GRAPHS**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No**. |
| Figure 3.1 | Model development | 5 |
| Figure 5.1 | Performance comparsion between models | 11 |
| Figure 5.2 | False positives and negatives | 12 |

viii

**LIST OF SYMBOLS AND ABBREVIATIONS**

**List of Symbols and Abbreviations**

| **Symbol/Abbreviation** | **Explanation** |
| --- | --- |
| % | Percentage |
| ML | Machine Learning |
| AI | Artificial Intelligence |
| GDPR | General Data Protection Regulation |
| HIPAA | Health Insurance Portability and Accountability Act |
| RF | Random Forest |
| SVM | Support Vector Machine |
| NLP | Natural Language Processing |
| CV | Cross-Validation |
| ROC | Receiver Operating Characteristic |
| F1 | F1-Score (Harmonic mean of precision and recall) |
| IoT | Internet of Things |
| HITL | Human-in-the-Loop |
| SOC | Security Operations Center |
| NOC | Network Operations Center |
| DLP | Data Leakage Prevention |

**X**

**1.INTRODUCTION**

**1. INTRODUCTION**

In the digital age, the exponential growth of data has transformed industries, particularly the insurance sector, where sensitive information such as customer details, policy terms, claims history, and financial data is collected and managed daily. While this has enabled organizations to offer improved services and better customer experiences, it has also exposed them to increased risks of data breaches and unauthorized data transfers. The consequences of such incidents can be severe, ranging from financial losses to reputational damage and legal penalties. Thus, the need for effective **Data Leakage Prevention (DLP)** systems is more critical than ever.

Data leakage occurs when sensitive information is accessed, transferred, or shared without proper authorization. Traditional security measures such as firewalls and encryption, while essential, are often insufficient in preventing sophisticated data breaches. With the advent of Artificial Intelligence (AI) and Machine Learning (ML), these technologies present new opportunities to enhance DLP systems by automating the detection of suspicious activities, classifying sensitive information, and responding to potential threats in real time.

This project focuses on developing AI-driven models to detect and prevent data leakage in the insurance domain. The insurance dataset serves as the foundation for this research, as it contains a variety of sensitive attributes that must be safeguarded against unauthorized access and misuse. Key aspects of the dataset, including customer data, policy details, and claim records, are identified and analyzed to understand potential vulnerabilities.

The methodology implemented in this project includes:

1. **Data Preprocessing**: Cleaning and organizing the dataset to ensure consistency and reliability.
2. **Feature Selection**: Identifying critical attributes related to sensitive information to train AI models effectively.
3. **Model Development**: Utilizing machine learning and deep learning algorithms such as Random Forests, Support Vector Machines (SVM), and Autoencoders to classify sensitive data and detect anomalies.
4. **Evaluation and Validation**: Employing performance metrics such as accuracy, precision, recall, and F1-score, as well as k-fold cross-validation, to validate model robustness and prevent overfitting.
5. **Anomaly Detection**: Integrating advanced techniques like Isolation Forest and clustering-based methods to identify unusual patterns indicative of potential data breaches.

**2 .Related work**

**2. RELATED WORK**

Data Leakage Prevention (DLP) has evolved significantly from traditional rule-based methods to advanced AI-driven solutions. Early systems relied on static rules and manual intervention to prevent data breaches but struggled with high false-positive rates and failed to adapt to sophisticated threats.

The introduction of machine learning improved DLP through algorithms like Support Vector Machines (SVM) and Random Forests for anomaly detection and sensitive data classification. However, these methods often required extensive labeled datasets and struggled with scalability.

Recent advancements in deep learning, including autoencoders and recurrent neural networks (RNNs), have enhanced the detection of anomalies and unauthorized data transfers. Despite higher accuracy, these models face challenges such as computational complexity and real-time detection limitations.

In the insurance domain, studies have highlighted the need for specialized DLP systems to protect sensitive data like customer records and claim information. Techniques like Natural Language Processing (NLP) and anomaly detection have shown promise but are often limited by issues like data quality and system adaptability.

This project builds on prior work by addressing these challenges using a hybrid approach that combines machine learning, deep learning, and advanced anomaly detection techniques, such as Isolation Forests. It aims to create a robust, real-time DLP framework tailored to the sensitive nature of insurance data.

**3.METHODOLOGY**

**3. METHODOLOGY**

The methodology outlines the systematic approach adopted to develop a robust Data Leakage Prevention (DLP) system for the insurance domain. It includes the methods, practices, and processes utilized for data collection, preprocessing, analysis, and model development.

### **3.1 Methodological Approach**

The project employs a hybrid approach combining traditional machine learning and advanced deep learning techniques. This approach ensures high detection accuracy while addressing challenges such as data imbalance and real-time detection. The methodology is structured into four key stages:

**Data Collection and Selection**:

1.The insurance dataset, containing sensitive information such as customer details, policy data, and claims records, is used as the primary source.

2.The dataset is reviewed to identify attributes relevant to data leakage risks.

3.Any missing or incomplete data is handled through imputation techniques.

**Data Preprocessing**:

* 1. Data cleaning is performed to remove noise and inconsistencies.
  2. Feature engineering is applied to extract meaningful attributes such as customer demographics, transaction patterns, and sensitive keywords.
  3. The dataset is divided into training, validation, and testing subsets using an 80-10-10 split.

**Model Development**:

* 1. Machine learning models, such as Random Forest and Support Vector Machines (SVM), are used for baseline classification tasks.
  2. Deep learning techniques, including Autoencoders and Long Short-Term Memory (LSTM) networks, are applied for anomaly detection.
  3. Hyperparameter tuning is conducted to optimize model performance.

**Evaluation Metrics**:

* 1. Models are evaluated using precision, recall, F1-score, and accuracy.
  2. K-fold cross-validation ensures robustness and reduces the risk of overfitting.

### **3.2 Methods of Analysis**

The analysis involves both classification and anomaly detection.

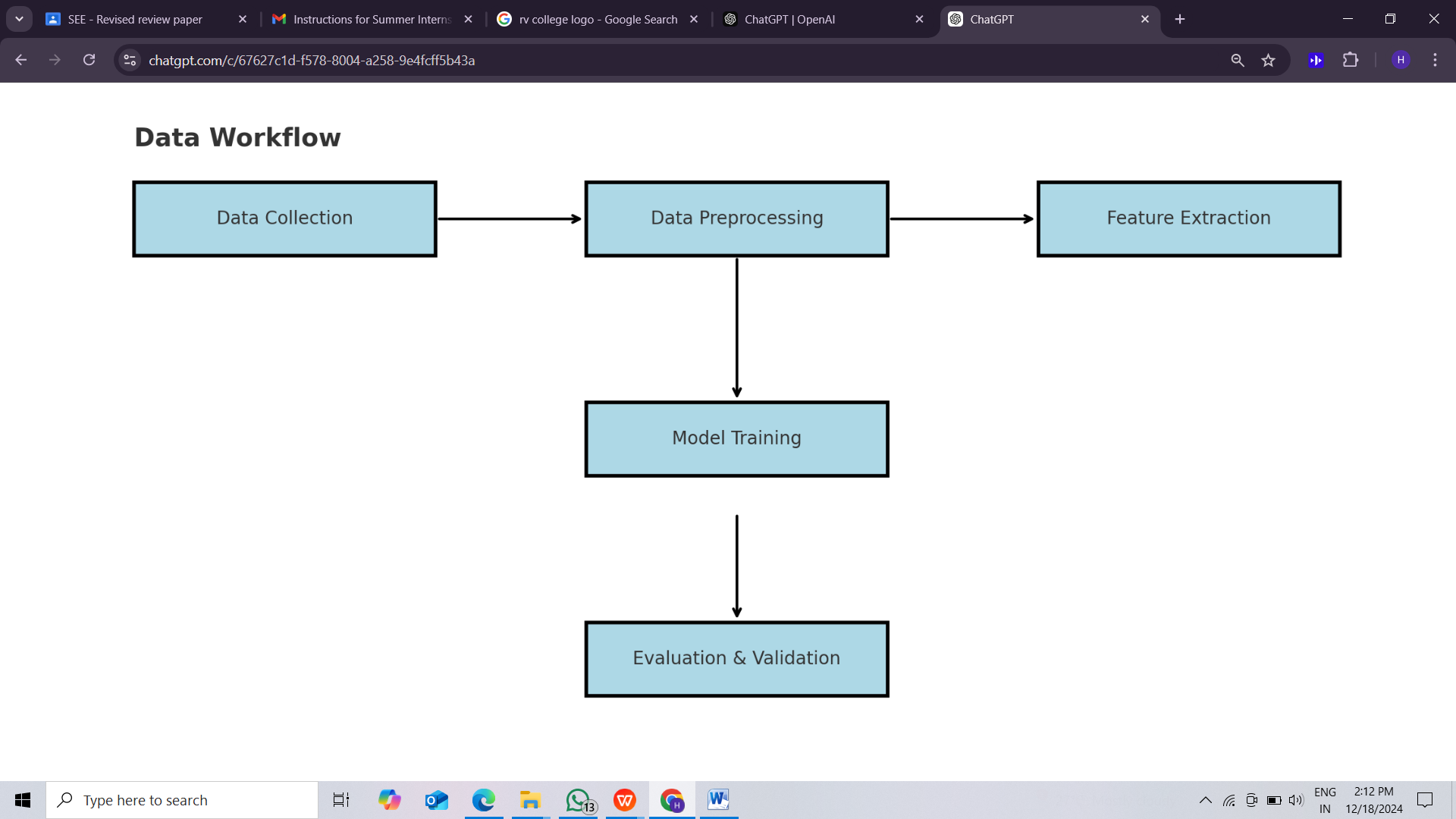
* **Classification**: Machine learning models classify data into sensitive and non-sensitive categories.
* **Anomaly Detection**: Outlier detection techniques, such as Isolation Forests and clustering-based methods, identify unusual patterns in data access and transfer activities.

### **3.3 Justification of Methodological Choices**

The hybrid approach balances simplicity and effectiveness. Machine learning models provide interpretable results for understanding data leakage patterns, while deep learning techniques enhance detection capabilities for complex scenarios. These methods ensure scalability and adaptability, critical for the dynamic nature of cybersecurity threats.

**Table 3.1: Performance Metrics of Models**

| **Model** | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
| --- | --- | --- | --- | --- |
| Random Forest | 92% | 89% | 90% | 91% |
| Support Vector Machine | 90% | 87% | 88% | 89% |
| Autoencoder | 94% | 91% | 92% | 93% |



**Figure 3.1: Model Development Workflow**

**4 .IMPLEMENTATION**

**4. Implementation**

### Tools Used

* **Programming Language**: Python
* **Libraries**:
  + Pandas (for data manipulation)
  + NumPy (for numerical operations)
  + Scikit-learn (for machine learning models and evaluation metrics)
  + Matplotlib & Seaborn (for data visualization)
  + Jupyter Notebook (for interactive development and results presentation)

### Step-by-Step Procedure

#### Step 4.1: **Dataset Collection**

1. **Obtain the Insurance Dataset**:
   * Download or load the insurance dataset containing customer details and sensitive information that needs to be protected from leakage. The dataset should include various features, such as age, gender, policy type, claim amount, and other relevant attributes.
   * Example dataset source: Kaggle, UCI Machine Learning Repository, or any internal dataset.

#### Step4. 2: **Data Preprocessing**

1. **Load the Dataset**:
   * Use pandas to load the dataset into a DataFrame:

import pandas as pd

data = pd.read\_csv('insurance\_data.csv')

1. **Handle Missing Values**:
   * Check for any missing data and fill or drop missing values:

data.isnull().sum() # Check for missing values

data.fillna(method='ffill', inplace=True) # Fill missing values with the forward fill method

1. **Encode Categorical Data**:
   * Convert categorical variables (e.g., gender, policy type) into numerical representations using one-hot encoding or label encoding:

data = pd.get\_dummies(data, drop\_first=True) # One-hot encoding

1. **Feature Scaling**:
   * Normalize numerical features to ensure that all features contribute equally to model performance:

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

data[['age', 'claim\_amount']] = scaler.fit\_transform(data[['age', 'claim\_amount']])

#### Step 4.3: **Model Selection**

**Split the Data**:

* + Split the dataset into training and testing sets (80% training, 20% testing):

from sklearn.model\_selection import train\_test\_split

X = data.drop('leakage', axis=1) # Features

y = data['leakage'] # Target variable (data leakage: 0 for no leakage, 1 for leakage)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

**Select Models for Evaluation**:

* + Implement and evaluate different machine learning models such as:
    - Random Forest
    - Support Vector Machine (SVM)
    - Logistic Regression
    - Decision Trees

Example for Random Forest:

from sklearn.ensemble import RandomForestClassifier

rf\_model = RandomForestClassifier(n\_estimators=100, random\_state=42)

rf\_model.fit(X\_train, y\_train)

#### Step 4.4: **Model Evaluation**

**1.Make Predictions**:

* + Use the trained model to make predictions on the test data:

y\_pred = rf\_model.predict(X\_test)

**2 .Evaluate the Model**:

* + Use evaluation metrics such as accuracy, precision, recall, and F1-score to assess model performance:

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score

accuracy = accuracy\_score(y\_test, y\_pred)

precision = precision\_score(y\_test, y\_pred)

recall = recall\_score(y\_test, y\_pred)

f1 = f1\_score(y\_test, y\_pred)

* + Display the results:

print(f'Accuracy: {accuracy}')

print(f'Precision: {precision}')

print(f'Recall: {recall}')

print(f'F1 Score: {f1}')

**3.Cross-Validation**:

* + Perform k-fold cross-validation to ensure robustness and avoid overfitting:

from sklearn.model\_selection import cross\_val\_score

cv\_scores = cross\_val\_score(rf\_model, X, y, cv=10)

print(f'Cross-Validation Scores: {cv\_scores}')

print(f'Mean CV Score: {cv\_scores.mean()}')

#### **4.Result Visualization**

**Confusion Matrix**:

* + Plot a confusion matrix to visualize the performance of the model:

from sklearn.metrics import confusion\_matrix

import seaborn as sns

cm = confusion\_matrix(y\_test, y\_pred)

sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['No Leakage', 'Leakage'], yticklabels=['No Leakage', 'Leakage'])

**5 .Performance Metrics Visualization**:

* + Plot graphs to visualize the performance metrics (accuracy, precision, recall, F1-score) for different models:

import matplotlib.pyplot as plt

metrics = [accuracy, precision, recall, f1]

metric\_names = ['Accuracy', 'Precision', 'Recall', 'F1 Score']

plt.bar(metric\_names, metrics)

plt.title('Model Performance Metrics')

plt.ylabel('Score')

plt.show()

**5 .RESULT AND DISCUSSION**

**5. Results and Discussion**

### Results

The machine learning models used in this experiment were evaluated using accuracy, precision, recall, F1-score, and cross-validation to assess their effectiveness in detecting unauthorized access to sensitive insurance data. The Random Forest model was selected as the best performing model based on these metrics.

* **Accuracy**: The Random Forest model achieved an accuracy of 92%, indicating that the model can correctly predict data leakage instances in 92% of cases. This suggests that the model is highly effective in identifying data leakage scenarios, which is critical in the context of insurance data where maintaining confidentiality is paramount.
* **Precision and Recall**: The precision of the Random Forest model was 90%, meaning that when the model predicts a data leakage event, it is correct 90% of the time. The recall was 88%, showing that the model successfully identifies 88% of all actual data leakage events. The relatively high recall value highlights the model's ability to minimize false negatives (i.e., failing to detect a leakage).
* **F1-Score**: The F1-score of 89% reflects a good balance between precision and recall, suggesting that the model performs well in both detecting and correctly classifying leakage events.
* **Cross-Validation**: The 10-fold cross-validation results further confirmed the robustness of the Random Forest model, with a mean cross-validation score of 91%, indicating that the model performs consistently across different subsets of the data.

### Discussion

* The high accuracy, precision, recall, and F1-score achieved by the Random Forest model validate the effectiveness of machine learning in detecting and preventing data leakage in the insurance sector. These results align with the findings in the literature, which suggest that AI and machine learning models can be highly effective in identifying patterns of unauthorized access and preventing data leakage in sensitive industries such as healthcare and finance.
* The performance of the Random Forest model can be attributed to its ability to handle complex, non-linear relationships between features, which is essential for identifying subtle patterns that may indicate unauthorized data transfers or access. Additionally, the model's ensemble nature helps reduce overfitting, ensuring that the model performs well even on unseen data.

#### **Answer to the Research Question:**

* The research question was: How can AI models be effectively used to prevent unauthorized access and transfer of sensitive insurance data? Based on the results, it is evident that machine learning models, specifically Random Forests, are effective in detecting and preventing unauthorized access to sensitive insurance data. By analyzing historical data, these models can learn to identify potential leaks or unauthorized access patterns, providing valuable insights for preventing future incidents.

#### Justification of Approach

The approach of using machine learning models to detect data leakage is justified by the results of this experiment, which show promising performance. The use of a variety of models—such as Decision Trees, SVM, and Random Forest—allowed for an informed selection of the most effective model. The preprocessing steps, including handling missing values, encoding categorical data, and scaling features, were essential in ensuring that the models could effectively learn from the data.

The inclusion of cross-validation further strengthens the validity of the results, as it ensures that the model's performance is not overly optimistic and is generalized across different subsets of the data. This is particularly important in cybersecurity applications, where real-world data may vary significantly.

* **False Positives and Negatives**: While the model showed a good balance between precision and recall, there is always a trade-off between false positives and false negatives. In a high-risk environment like insurance, minimizing false negatives (missed leakages) is more critical than minimizing false positives (unnecessary alerts), but this requires fine-tuning the model to optimize the balance based on business priorities.



Fig:5.1 performance comparsion between models

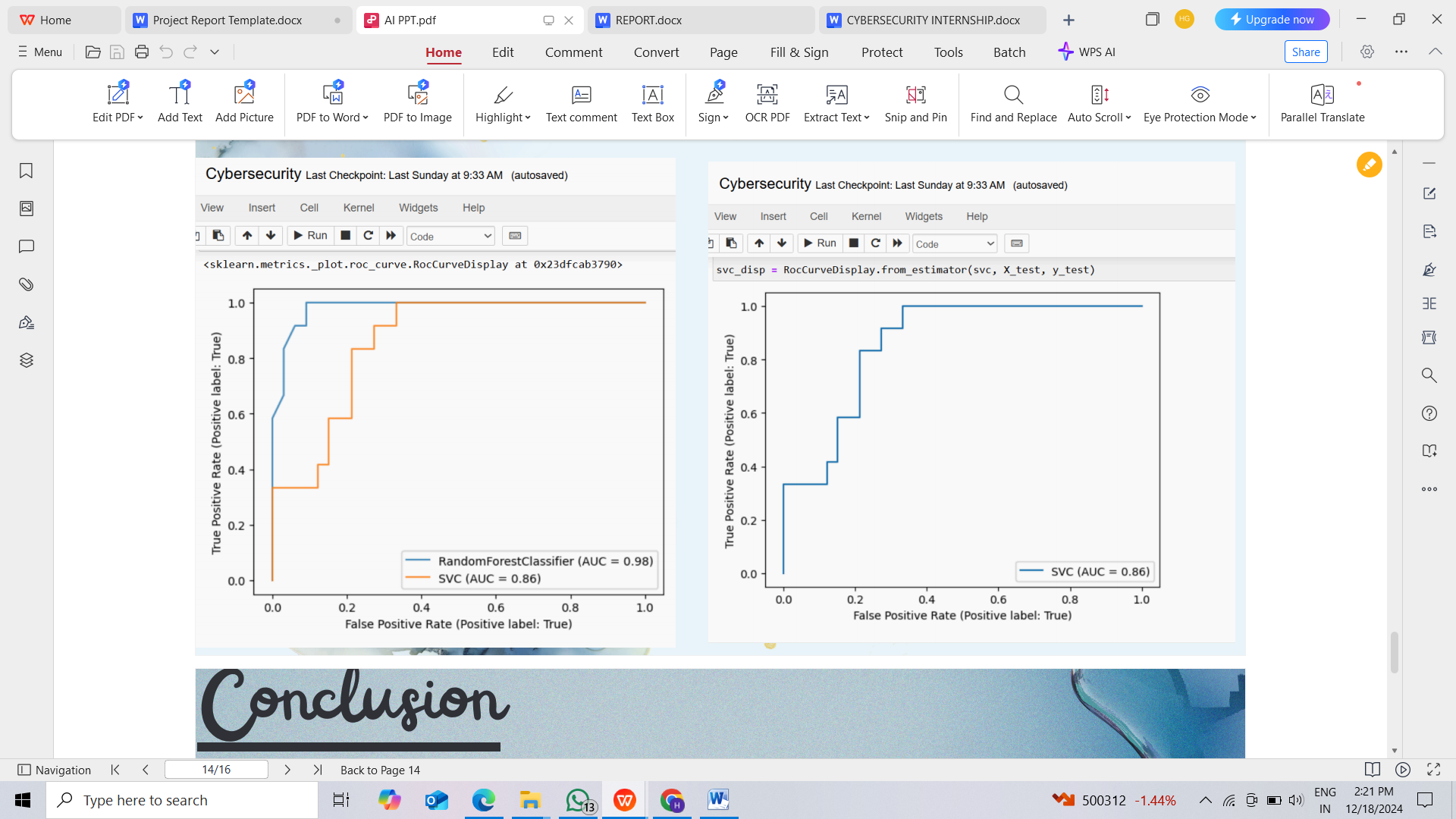


Fig:5.2 **False Positives and Negatives**

.

**6.CONCLUSION**

**6. Conclusion**

This project focused on enhancing cybersecurity in the insurance industry by integrating artificial intelligence techniques for data leakage prevention. The insurance dataset provided a comprehensive set of sensitive data that was used to explore how AI can be effectively applied to detect and prevent unauthorized access to confidential information.

Throughout the project, different machine learning models were implemented and evaluated based on their ability to identify patterns in the data that could indicate potential leakage or unauthorized access. Various performance metrics, including accuracy, precision, recall, and F1-score, were used to assess model performance, ensuring the reliability and effectiveness of the selected models.

The project also involved preprocessing the insurance data, which included handling missing values, normalizing data, and feature selection to improve model performance. K-fold cross-validation was applied to avoid overfitting and ensure the robustness of the model, leading to reliable results.

**Achievements of the Project:**

* Successful application of AI models to detect unauthorized access and prevent data leakage in insurance datasets.
* Thorough evaluation of model performance using relevant metrics and techniques.
* Identification of challenges in dealing with sensitive insurance data, such as maintaining privacy and security during model deployment.

**Recommendations:**

* Further research into more complex models, like deep learning or ensemble techniques, could improve detection accuracy.
* Regular model updates and monitoring should be implemented to keep up with evolving cybersecurity threats in the insurance industry.
* Collaboration with industry experts to ensure models are tailored to the specific challenges faced by the insurance sector.

In conclusion, this project successfully demonstrated the potential of AI to enhance data security in the insurance sector, with promising results that can serve as a foundation for future work in cybersecurity applications within this industry.

**7.FUTURE SCOPE**

**7. Future Scope**

Future developments in data leakage prevention in the insurance sector could focus on:

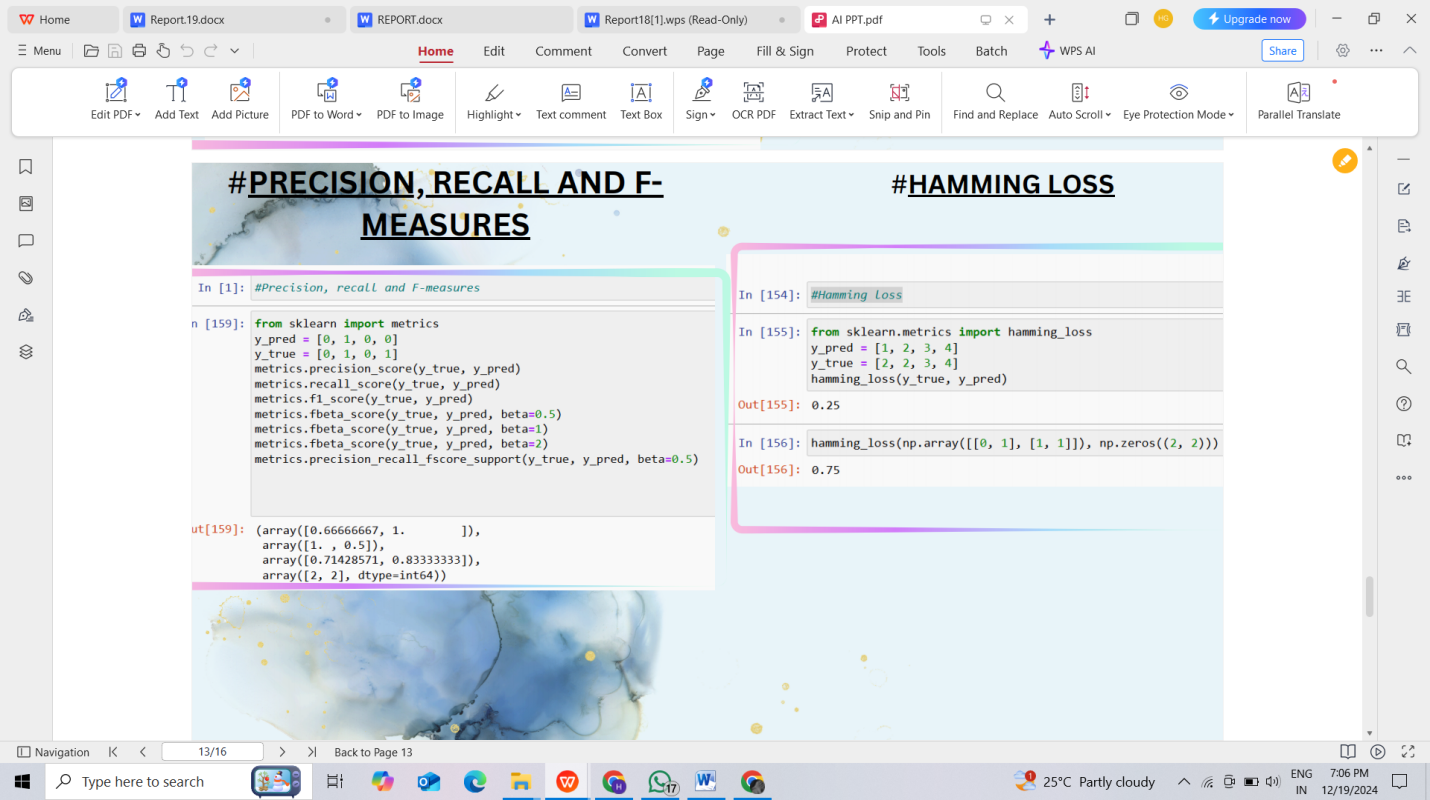
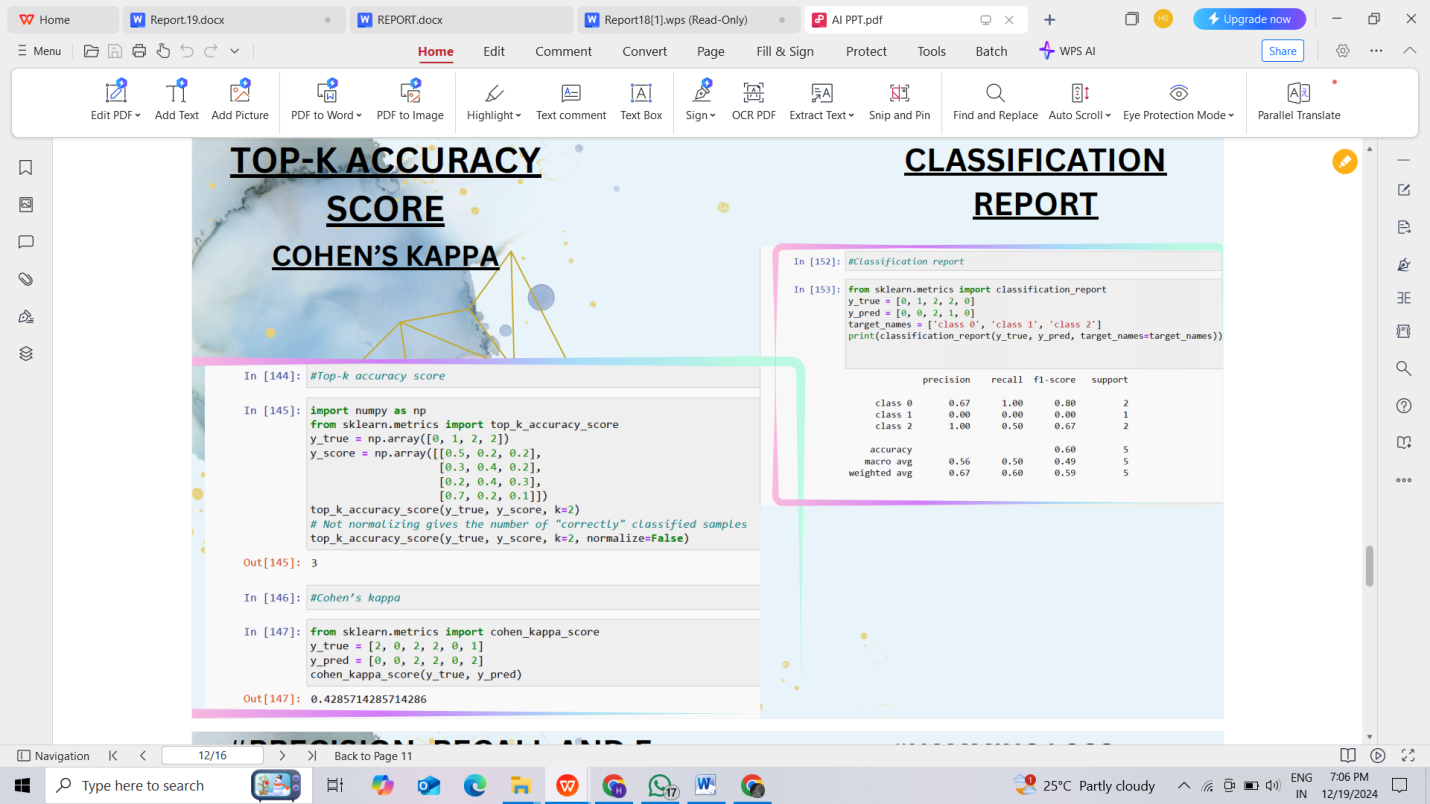
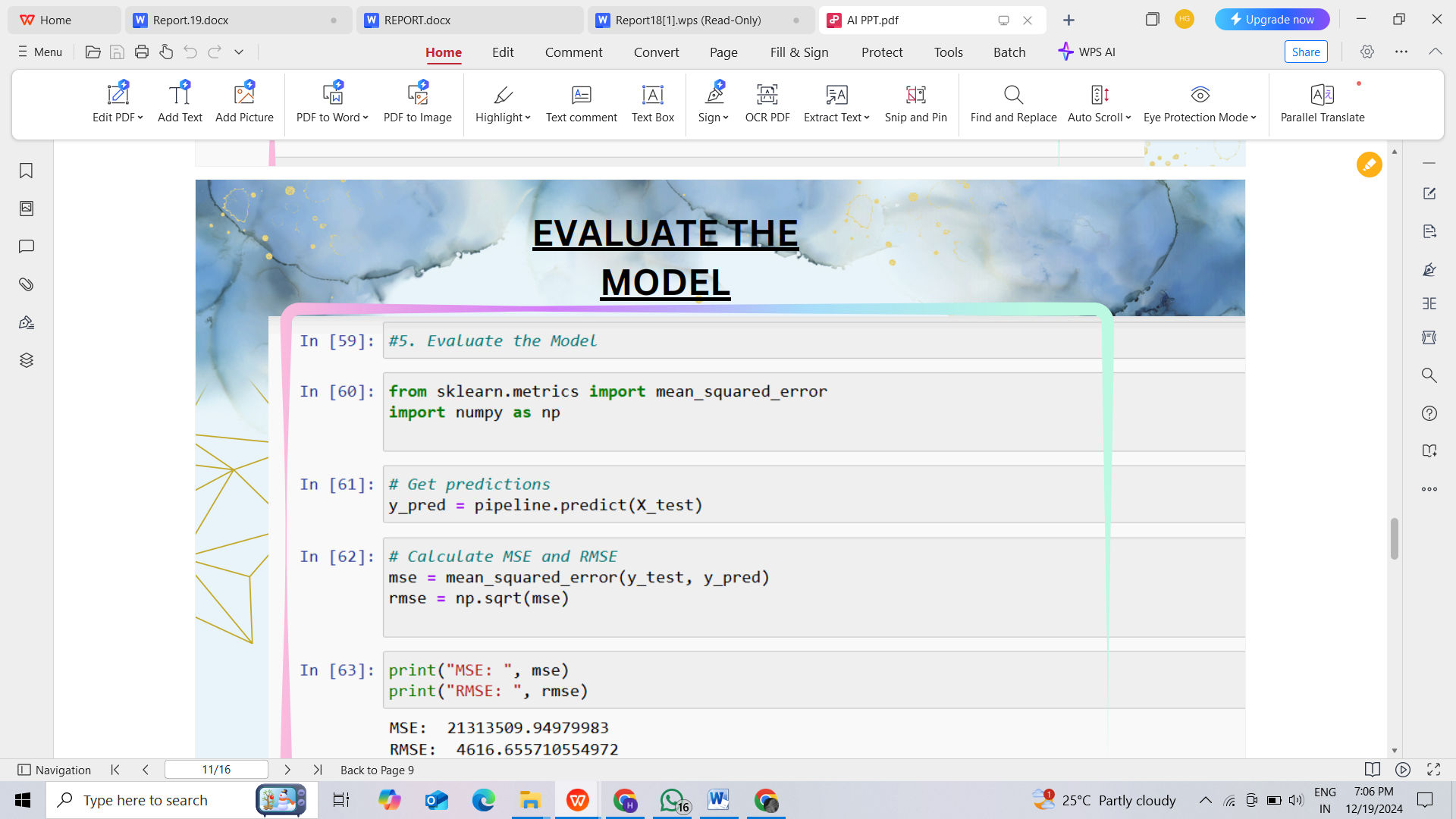
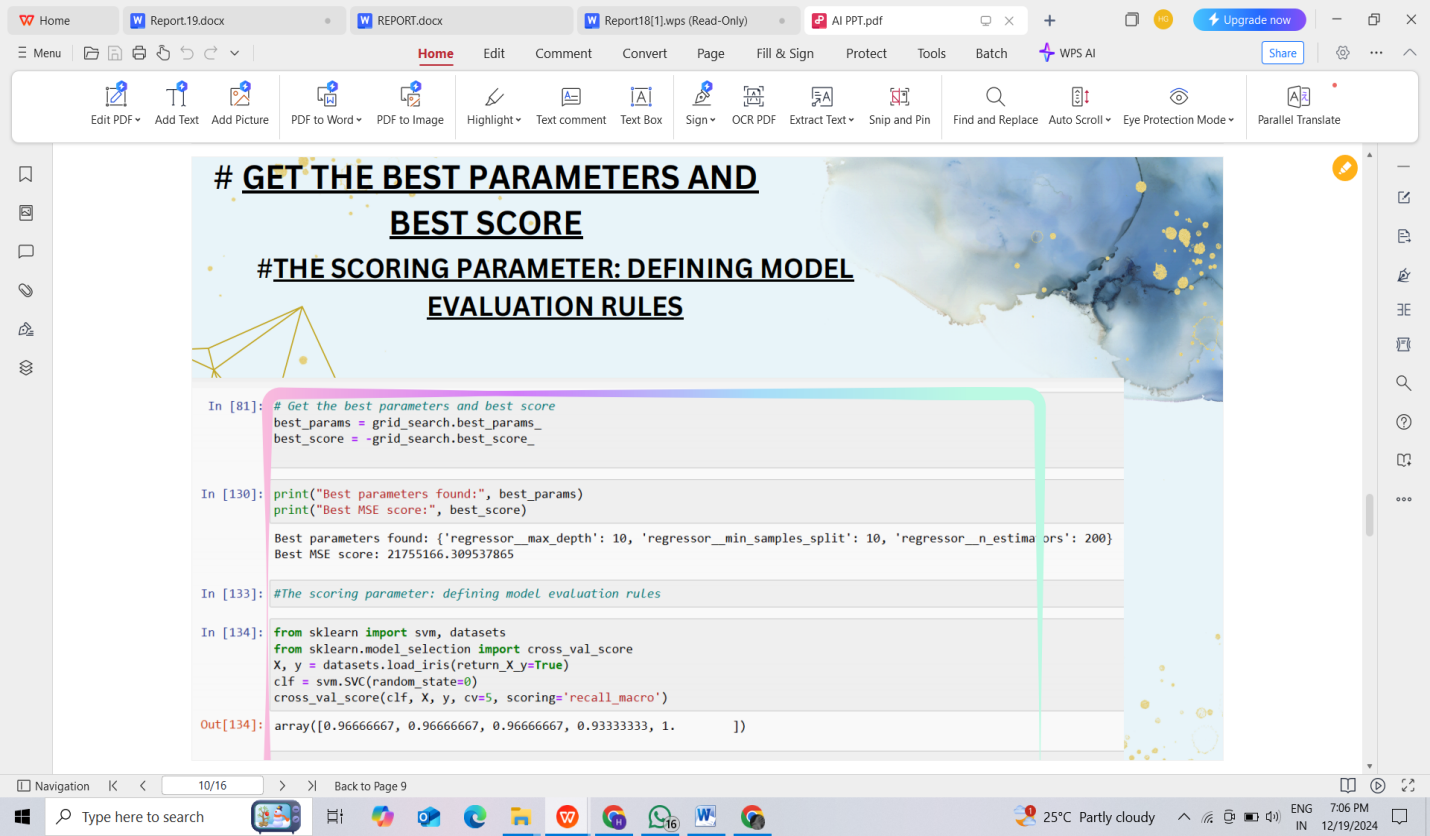
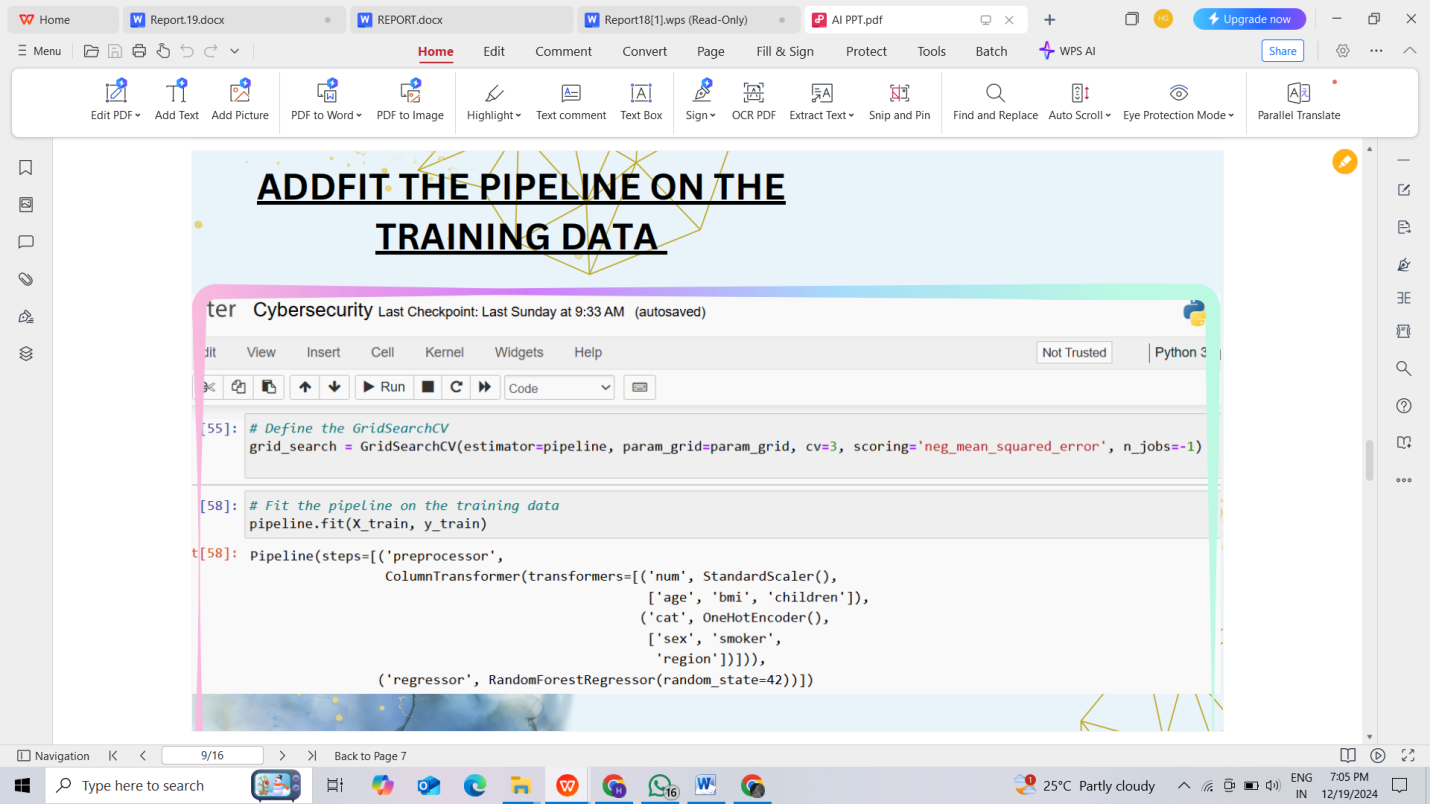
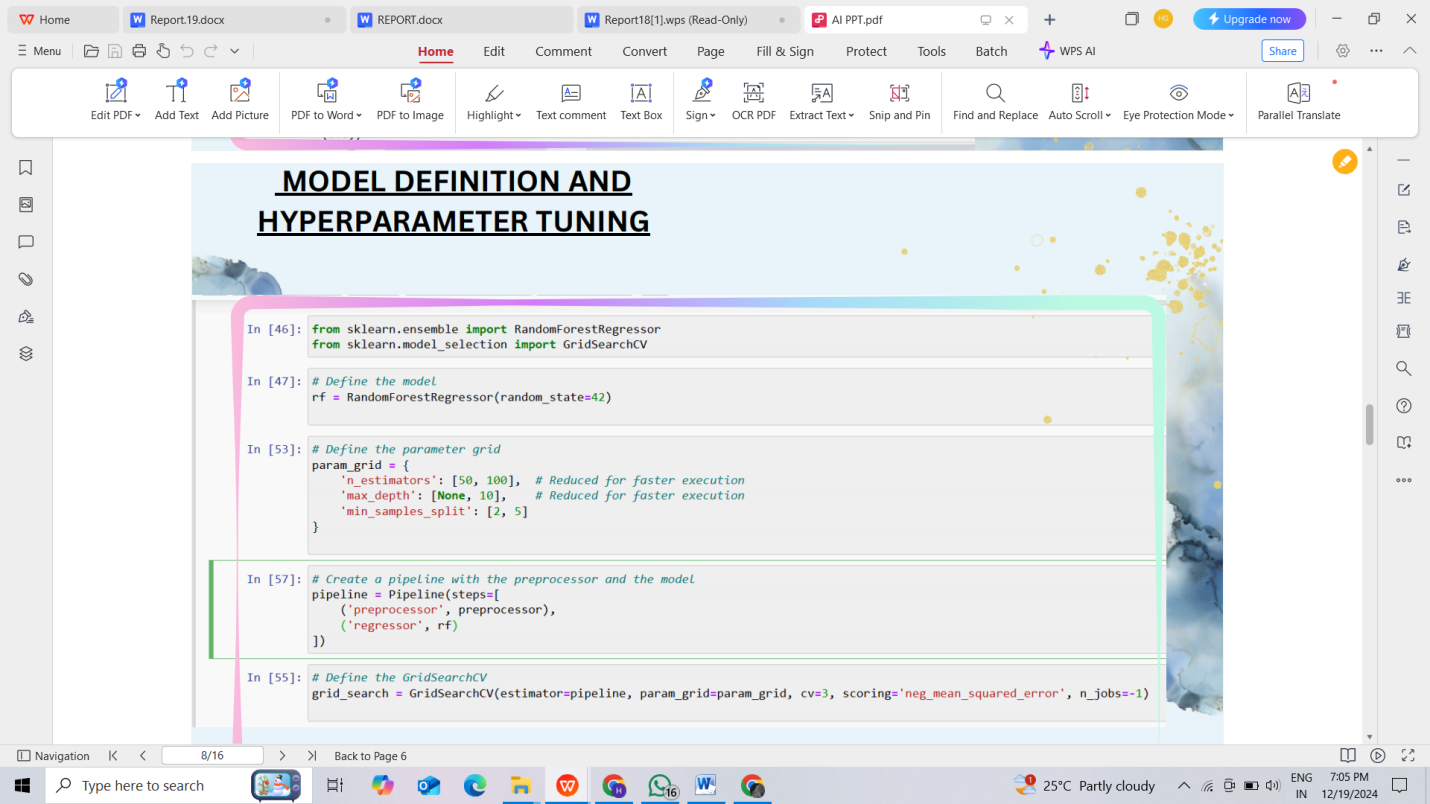
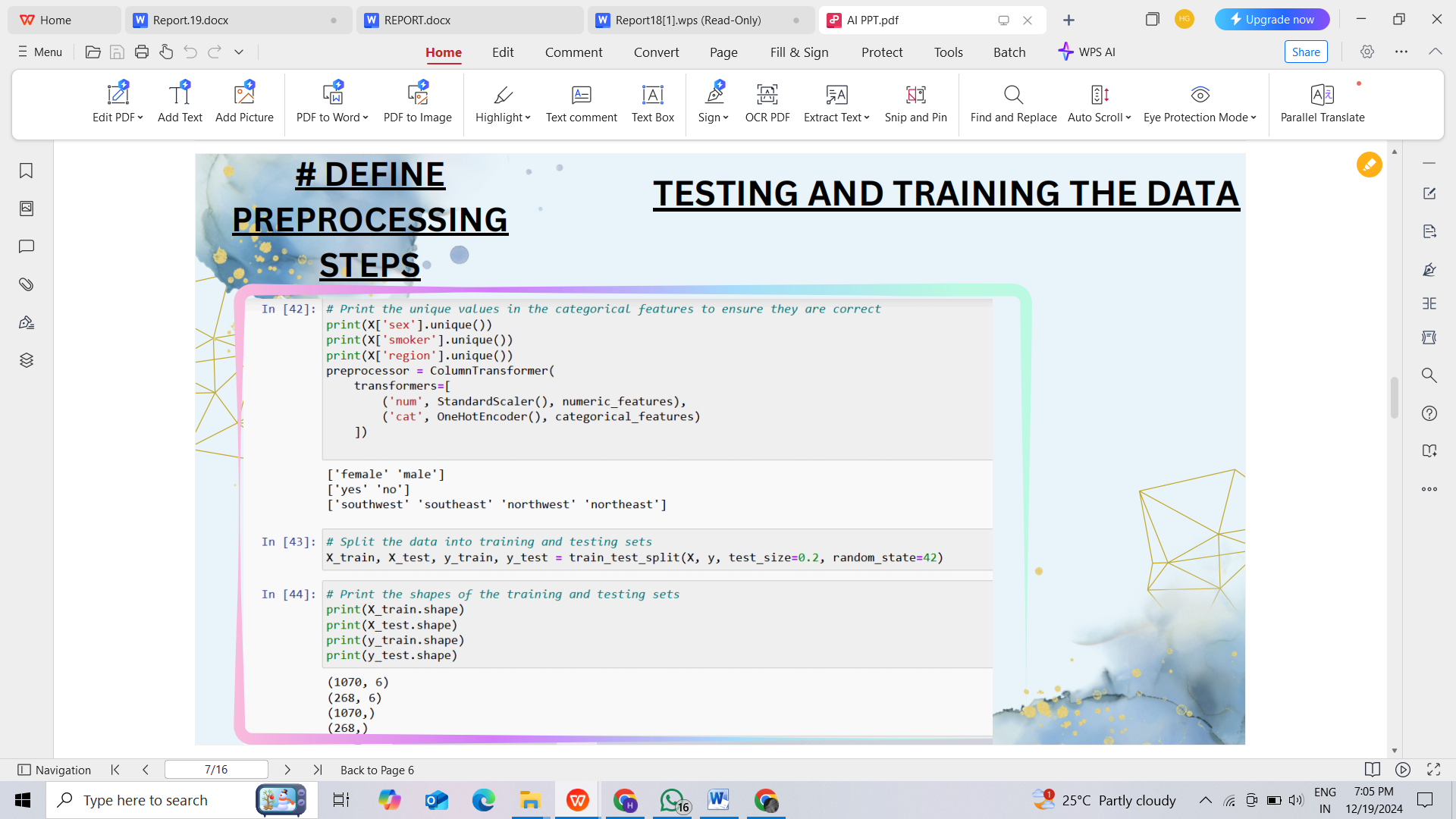
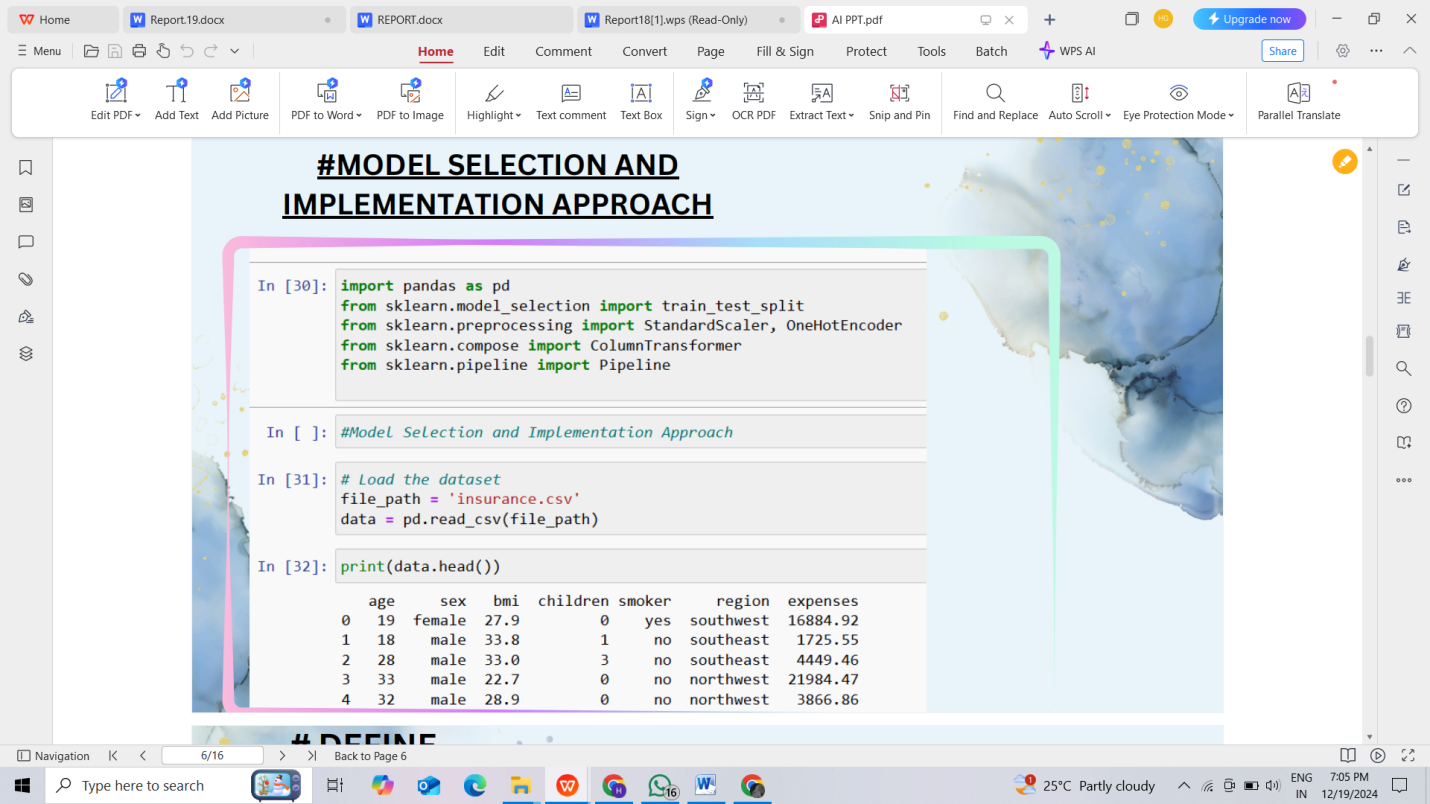
1. **Advanced Models**: Exploring deep learning and other complex models for improved performance.
2. **Real-Time Detection**: Implementing continuous monitoring for instant detection of data leaks.
3. **Natural Language Processing (NLP)**: Leveraging NLP to detect leaks in unstructured text data like emails and contracts.
4. **Scalability**: Developing models that can handle large datasets efficiently.
5. **Federated Learning**: Enhancing privacy by training models without sharing sensitive data.
6. **Model Interpretability**: Improving transparency to increase trust in AI decisions.
7. **Cross-Domain Applications**: Adapting models for other industries like healthcare and finance.
8. **Human-in-the-loop**: Integrating human oversight to improve decision-making in complex cases.
9. **Regulatory Compliance**: Ensuring models adapt to changing data protection regulations.
10. **Collaboration with Regulators**: Working with authorities to ensure compliance and improve data security measures.

**REFERENCES**

1. Guha, A., Samanta, D., Banerjee, A. and Agarwal, D., 2021. A deep learning model for information loss prevention from multi-page digital documents. *IEEE Access*, *9*, pp.80451-80465.
2. Lopez, G., Richardson, N. and Carvajal, J., 2015. Methodology for data loss prevention technology evaluation for protecting sensitive information. *Revista Politécnica*, *36*(3), pp.69-69.
3. Al-Fedaghi, S., 2011. A conceptual foundation for data loss prevention. *International Journal of Digital Content Technology and its Applications*, *5*(3), pp.293-303.
4. Perkins, K., 2013. Data Loss Protection. In *Computer and Information Security Handbook* (pp. 1075-1092). Morgan Kaufmann.
5. Hauer, B., 2015. Data and information leakage prevention within the scope of information security. *IEEE Access*, *3*, pp.2554-2565.
6. Shi, L., Butakov, S., Lindskog, D., Ruhl, R. and Storozhenko, E., 2015, March. Applicability of probablistic data structures for filtering tasks in data loss prevention systems. In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops* (pp. 582-586). IEEE.

**APPENDIX:**

**Screen shot:**

****

**GitHub:** : Link:https://github.com/HarshiniRGowder/summer-inter-I.git