MAHARASHTRA STATE BOARD OF TECHNICAL EDUCATION, MUMBAI

GOVERNMENT POLYTECHNIC KARAD

THIRD YEAR DIPLOMA COMPUTER ENGINEERING (I-SCHEME)

PART [B]

MICRO-PROJECT REPORT

"MITIGATING PHISHING ATTACKS USING MACHINE LEARNING TECHNIQUES"

UNDER THE SUBJECT

Emerging Trends in Computer and Information Technology (22618)

SUBMITTED BY

Sr.no	Roll No	Enrollment No	Name of Team Member
1.	2251	2100100053	Pratik Pramod Shejwal
2.	2254	2100100056	Shravani Bharat Mahajan
3.	2260	2100100063	Riya Sunil Kharade

UNDER THE GUIDANCE

Mrs. M.A. Birnale

(DEPARTMENT OF COMPUTER ENGINEERING)

2023-24

Certificate of completion

Of Micro-project Assessment at the end of Semester

This is to certify that,

Sr.no	Roll No	Enrollment No	Name of Team Member
1.	2251	2100100053	Pratik Pramod Shejwal
2.	2254	2100100056	Shravani Bharat Mahajan
3.	2260	2100100063	Riya Sunil Kharade

Has successfully completed "MITIGATING PHISHING ATTACKS USING MACHINE LEARNING TECHNIQUES" Micro-project of the Sixth semester Diploma in Computer Engineering of subject Emerging Trends in Computer and Information Technology (22618) from Government Polytechnic Karad. Institute with Institute code (0010).

Mrs. M.A.Birnale	**Prof. S. B. Patil**	**Dr. K. M. Bagwan**
(Project Guide)	(Head of the department.)	(Head of Institute)

Seal of the institute

ACKNOWLEDGEMENT

We take this opportunity to thank all those who have directly and indirectly inspired, directed and assisted us towards successfully completion of this project report.

We express our sincere thanks to Dr. K. M. Bagwan Principal of Government Polytechnic, Karad and the Head of Department Prof. Patil S.B, for having us allowed to submit this report as a part of our academic learning.

We express our sincere thanks to Mrs. M.A. Birnale Lecturer in Computer Engineering, Govt. Polytechnic, Karad for encouragement throughout the project report and guideline in designing and working out this project. We are also grateful to team of project.

Place: Government Polytechnic Karad

Date: 30-03-2024

Yours Sincerely,

2251-Pratik Pramod Shejwal

2254-Shravani Bharat Mahajan

2260-Riya Sunil Kharade

➢ RATIONALE

Phishing tricks online aim to deceive people into sharing private information, and current detection methods often lag behind hackers' innovations. Machine learning offers a solution by training computers to swiftly identify these traps, outsmarting cyber crooks. With machine learning, we stay proactive against threats, safeguarding personal data from falling into the wrong hands. It's like having a digital watchdog, keeping us one step ahead in the cybersecurity game. This approach strengthens our defenses, ensuring our online safety and peace of mind.

➢ AIM AND BENEFITS

◉ AIM:

1) **Spotting Phishing Tricks:**
   Teach computers to quickly identify phishing emails by analyzing sender, content, and language cues.

2) **Quick Protection:**
   Implement real-time phishing attack detection systems using machine learning for immediate response.

◉ BENEFITS:

1) **Enhanced Security:**
   Machine learning improves email security by swiftly identifying and blocking phishing attempts, protecting personal information.

2) **Cost Savings:**
   Early detection of phishing attacks reduces financial losses and expenses associated with cybersecurity breaches, ensuring financial stability.

➢ COURSE OUTCOMES ACHIEVED

CO A] Describe Artificial Intelligence, Machine learning and deep learning.

CO F] Describe Network, Operating System and applications vulnerabilities.

➢ LITERATIVE REVIEW

1. We Referred Book “A MACHINE LEARNING APPROACH TO PHISHING DETECTION AND DEFENSE” by Oluwatobi Ayodejo Akanbi and his team.
2. We Referred Following Link:
    a) https://en.wikipedia.org/wiki/Machine_learning
    b) https://www.geeksforgeeks.org/machine-learning/
    c) https://www.phishprotection.com/phishing-awareness/machine-learning-helps-fighting-phishing-attacks

➢ ACTUAL METHODOLOGY FOLLOWED

1. Discussion about given topic.
2. Selection of good leader and distribution of responsibilities.
3. Collection of information using different resources, Analysis of given information.
4. Presentation of given report.
5. Completion and submission of given tasks.

➢ ACTUAL RESOURCES USED

No	Name of the Resource	Specifications	Quantity	Remark
1	Computer System	At least 4GB RAM, i5 Processor	1	Required
2	Software	Microsoft word	1	Required

➢ OUTPUT OF MICROPROJECT:

In our microproject titled "Mitigating Phishing Attacks Using Machine Learning Techniques," we've explained simple concepts about machine learning and phishing. We talked about how phishing tricks people online and why it's a big problem. We described how phishing attacks happen and the harm they cause to people and companies. We stressed why it's crucial to stop phishing to protect online safety. Additionally, we discussed how machine learning methods can help fight against phishing, making it easier to understand and use in real situations.

➢ INTRODUCTION TO PHISHING ATTACKS

In today's digital world, there's a sneaky threat called phishing. Phishing is when someone tricks you into sharing personal info like passwords or credit card numbers by pretending to be someone trustworthy. They might use fake emails, texts, or websites to do this. But we have a superhero called machine learning.

Machine learning is like teaching computers to spot patterns and learn from them, just like humans do. So, it can analyze lots of data to catch sneaky phishing attempts that humans might miss. these smart computer programs can look at different parts of an email, text, or website, like who sent it or how it's written, to figure out if it's real or fake.

The best part is, they keep getting better at it! They learn from each new phishing trick and get better at stopping them. so, by using machine learning, we can make it harder for phishers to trick us and keep our personal info safe online.

➢ TYPES OF PHISHING ATTACKS

1. **Email Phishing:**

   Cybercriminals send fake emails from trusted places like banks, trying to trick people into giving away sensitive info or clicking on bad links.

2. **Spear Phishing:**

   Bad guys target specific people with personalized messages, using info from social media to make the emails seem real.

3. **Clone Phishing:**

   Crooks make fake copies of real emails, adding bad links or files. They send these to trick people into clicking on them.

4. **Whaling:**

   Hackers go after big shots like CEOs with fancy emails, hoping to steal important company info or money.

5. **SMiShing (SMS Phishing):**

   Scammers send sneaky texts with bad links or asks for personal info, hoping to trick people into giving it to them.

➢ IMPACTS OF PHISHING ATTACKS

1. **Financial Loss:**

   Phishing makes people lose money by tricking them into unauthorized transactions or stealing their identity, causing big financial problems.

2. **Identity Theft:**

   Phishing steals personal information to pretend to be someone else, leading to fake activities and messed-up identities.

3. **Data Breaches:**

   Phishing exposes important info, causing big problems like reputation damage, fines, and more risks for everyone involved.

4. **Reputational Damage:**

   Phishing makes companies look bad in the news, hurting their reputation and making people trust them less.

5. **Emotional Distress:**

   Phishing hurts people emotionally by invading their privacy and making them anxious and upset about being tricked.

6. **Trust Issues:**

   Phishing erodes trust between individuals and organizations, making people hesitant to share personal information or engage in online transactions.

7. **Productivity Loss:**

   Phishing attacks disrupt normal business operations, leading to wasted time and resources in resolving security breaches and restoring systems.

➢ NEED TO STOP PHISHING ATTACKS

Phishing is a big problem that we need to stop because it tricks people into giving away important stuff like passwords or credit card numbers. This happens when bad guys pretend to be someone trustworthy, like a bank or a company, in emails or messages. It's super important to stop phishing because it can make people lose money or even ruin a company's reputation. Also, it makes it hard for us to trust emails or websites, which is not good for online shopping or doing things on the internet. Sometimes, phishing can lead to even worse stuff like viruses or hackers getting into our computers, which can cause a lot of damage. So, stopping phishing is really important to keep our personal info safe, trust the internet, and make sure things run smoothly online.

➢ Challenges in Phishing Detection

Detecting phishing attacks can be tough because hackers keep inventing new tricks. Phishing emails often look real, making them tricky to spot. Hackers also use fake websites or links that seem real, making it hard to know what's safe. They switch tactics fast, so security systems must stay updated. And with so many emails sent every day, some phishing attempts sneak by unnoticed. Lastly, hackers are always improving their techniques, making it a constant challenge for security experts to stay one step ahead. give in simple

➢ Introduction to Machine Learning

Machine learning is like teaching computers to think and learn like humans. It's a way for computers to learn from data and make decisions without being explicitly programmed. Imagine if you could teach your computer to recognize faces in photos or predict which movies you might like to watch. That's what machine learning does! It's used in lots of things we use every day, like voice assistants, recommendation systems, and even self-driving cars. Instead of following strict rules, machine learning algorithms can learn from examples and improve over time. So, it's like giving computers the ability to learn from experience, just like we do.

➢ Deep Learning Techniques for Phishing Detection

1. **Long Short-Term Memory (LSTM) Networks:**

   LSTM networks are a type of RNN that can capture long-range dependencies in sequential data, enabling them to identify subtle patterns or anomalies in email content or user interactions indicative of phishing attacks.

2. **Deep Autoencoders:**

   These can compare normal and phishing content, like email text or website features, to spot differences and catch phishing attempts.

3. **Generative Adversarial Networks (GANs):**

   Can create fake phishing examples to help train anti-phishing models better.

4. **Capsule Networks:**

   These can understand the relationships between different parts of emails or websites to find phishing signs based on how things are arranged.

5. **Attention Mechanisms:**

   They focus on the important parts of data, like specific words or images, to help catch phishing attempts more accurately.

➢ Feature Selection Methods:

"It is a process of automatically or manually selecting the subset of most appropriate and relevant features to be used in model building."
Feature selection involves identifying the most relevant features from a dataset to improve model performance by reducing dimensionality and eliminating irrelevant or redundant information. Common methods include evaluating feature importance, statistical tests, and iterative techniques such as recursive feature elimination or wrapper methods.

Below are some benefits of using feature selection in machine learning:

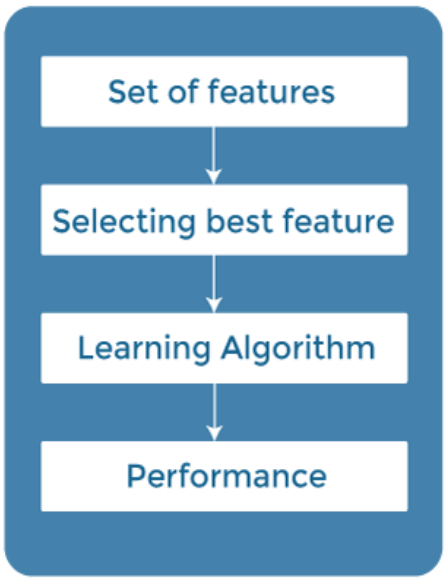
1. It helps in avoiding the curse of dimensionality.
2. It helps in the simplification of the model so that it can be easily interpreted by the researchers.
3. It reduces the training time.
4. It reduces overfitting hence enhance the generalization.

1. Filter Methods:

In Filter Method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step.

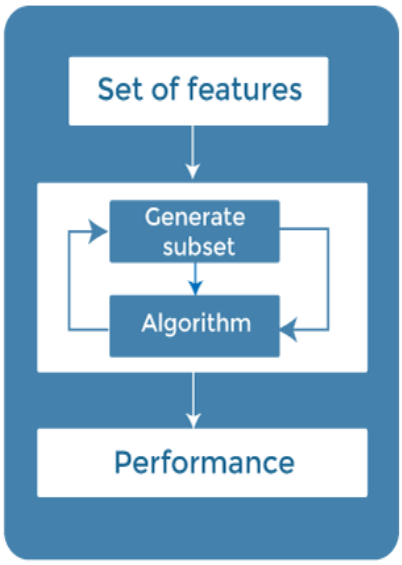
The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.

The advantage of using filter methods is that it needs low computational time and does not overfit the data.



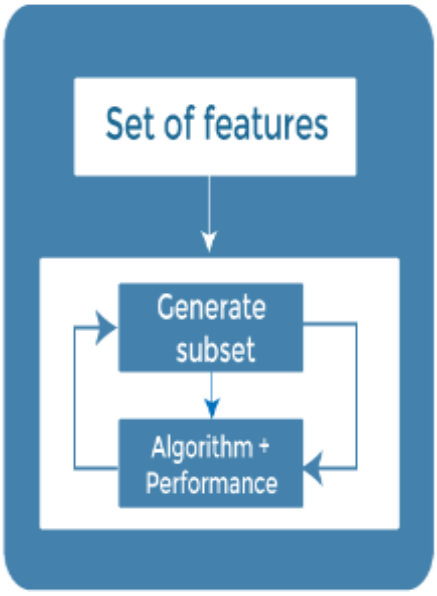
2. **Wrapper Methods:**

In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations. It trains the algorithm by using the subset of features iteratively. By considering various feature subsets and their corresponding model performances, wrapper methods aim to find the optimal set of features that maximizes the model's predictive accuracy.



3. **Embedded Methods:**

Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter method but more accurate than the filter method. These methods are also iterative, which evaluates each iteration, and optimally finds the most important features that contribute the most to training in a particular iteration.



➢ Evaluation Metrics for Phishing Detection

Some Example of Evaluation Metrics are :

1. **Accuracy:**

   Accuracy measures the proportion of correctly classified instances (both phishing and legitimate), providing an overall assessment of model performance but may be misleading in imbalanced datasets.

2. **Precision:**

   Precision calculates the ratio of correctly classified phishing instances to the total number of instances classified as phishing, indicating the model's ability to avoid false positives (misclassifying legitimate emails as phishing).

3. **Recall (Sensitivity):**

   Recall measures the ratio of correctly classified phishing instances to the total number of actual phishing instances, indicating the model's ability to detect all phishing emails (minimizing false negatives).

4. **F1 Score:**

   F1 score is the harmonic mean of precision and recall, providing a balanced measure of model performance that considers both false positives and false negatives, suitable for imbalanced datasets.

5. **Receiver Operating Characteristic (ROC) Curve:**

   The ROC curve plots the true positive rate (TPR or recall) against the false positive rate (FPR) at various classification thresholds, -providing insights into the trade-off between sensitivity and specificity.

6. **Area Under the Curve (AUC):**

   AUC represents the overall performance of a binary classification model by calculating the area under the ROC curve, with higher values indicating better discrimination between phishing and legitimate emails.

7. **Confusion Matrix:**

   The confusion matrix summarizes model predictions by tabulating true positives, false positives, true negatives, and false negatives, providing a detailed breakdown of classification outcomes for further analysis.

➢ Real-world Datasets for Phishing Detection

Machine learning datasets are collections of data used to train and test machine learning models. They're like a library of examples that help the model learn patterns and make predictions. Creating a good dataset involves understanding what kind of data the model needs and making sure it's accurate and representative of real-world scenarios. The more high-quality data the model has, the better it can learn and perform. But it's also important to manage the size of the dataset so it doesn't overwhelm the computer's processing power.

- ### Creating a dataset for machine learning (ML) step:

1 . Training Data Collection: Gather relevant data for training your model.

2 . **Data Storage:** Keep your collected data in a safe and organized place.

3 . Data Preparation: Clean and format your data to make it usable for training.

4. Quality Control: Check your data for accuracy, outliers, and biases.

5. Rescaling: Adjust the scale of your data to ensure fair comparison between features.

- ### Various datasets available for training and testing machine learning models for phishing detection

1. **Phishing Websites Dataset (UCI):**
   Contains features extracted from URLs of phishing and legitimate websites, aiding in training phishing detection models.

2. **Phishing Email Corpus (Enron):**
   Consists of Enron emails, helping develop and test email phishing detection algorithms.

3. **Websites Phishing URL Dataset (Kaggle):**
   Contains URLs labeled as phishing or legitimate, with features like domain age, useful for building website-based phishing detection models.

4. **Microsoft Malware Prediction Dataset:**
   Telemetry data from Microsoft Defender Antivirus for developing models to detect phishing and other malicious activities.

➢ Transfer Learning in Phishing Detection

Transfer learning involves leveraging knowledge gained from solving one problem and applying it to a different but related problem. By initializing a new model with pre-trained weights from a model trained on a similar task, transfer learning accelerates training and improves performance, especially when data for the new task is limited. This approach saves time and computational resources while enabling effective adaptation to new tasks with varying degrees of similarity to the original task.

- **Steps to Use Transfer Learning:**

1 Training a Model to Reuse it:

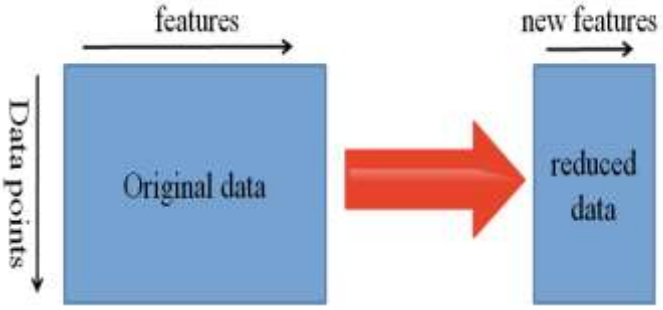
If you lack data for Task A, train a deep neural network on a related Task B with plenty of data. Then, adapt this model for Task A by either using the entire model or just specific layers, depending on how similar the inputs are.

2 Using a Pre-Trained Model:

Utilize existing pre-trained models and decide how many layers to reuse and retrain based on your specific task. Many pre-trained models are available, including those in Keras, which can be fine-tuned for your specific needs.

3 Extraction of Features:

Use deep learning to automatically identify important features through representation learning. This method often produces better results than manually designing features, but domain knowledge is still essential for selecting relevant features.



4 Extraction of Features in Neural Networks:

Neural networks inherently learn which features are important, reducing the need for manual feature engineering. Representation learning algorithms can efficiently identify significant feature combinations, even for complex tasks, minimizing the need for manual feature selection.

➢ Ethical Considerations in Phishing Detection Using Machine Learning

1. **Privacy Preservation:**

   Phishing detection systems must ensure the privacy and confidentiality of user data, minimizing the risk of unauthorized access or misuse of sensitive information during the detection process.

2. **Bias and Fairness:**

   Phishing detection models should be free from bias and ensure fairness in their decisions, avoiding discrimination based on factors such as race, gender, or socio-economic status in classifying emails or websites as phishing.

3. **Transparency and Explainability:**

   Phishing detection algorithms should be transparent and explainable, allowing users to understand how decisions are made and enabling accountability for false positives or false negatives.

4. **Data Security:**

   Phishing detection systems must prioritize data security measures to protect against data breaches or unauthorized access, safeguarding sensitive information collected for training or evaluation purposes.

5. **User Consent and Control:**

   Users should have the right to consent to the use of their data for phishing detection purposes and maintain control over how their information is collected, stored, and utilized by detection systems.

6. **Accountability and Liability:**

   Organizations developing phishing detection systems must be accountable for the performance and outcomes of their algorithms, assuming liability for any harm caused by false positives or failures to detect phishing attacks.

7. **Impact on User Experience:**

   Phishing detection mechanisms should minimize disruption to user experience while effectively protecting users from phishing threats, balancing security needs with usability considerations.

8. **Continual Monitoring and Evaluation:**

   Phishing detection systems should undergo continual monitoring and evaluation to assess their effectiveness, detect and mitigate biases or ethical concerns, and ensure compliance with evolving ethical standards and regulations.

➢ Case Studies of Successful Phishing Detection Systems

1. **Microsoft Defender for Office 365:**

   This system uses advanced machine learning to scan email content and catch phishing attempts, automatically isolating suspicious messages to protect users.

2. **PhishMe (Cofense):**

   Now called Cofense, this platform trains users to recognize phishing attacks by analyzing behavior with machine learning, helping organizations educate employees to detect and handle phishing threats.

3. **Proofpoint Email Protection:**

   Proofpoint's system employs machine learning to scan email content and sender behavior, quickly spotting and blocking phishing emails in real-time while providing training to improve user awareness.

4. **Symantec Email Security:**

   Symantec's solution uses machine learning to detect and block phishing emails, utilizing threat intelligence and behavioral analysis to identify evolving phishing tactics and protect organizations.

5. **Barracuda Sentinel:**

   Barracuda Sentinel uses AI and machine learning to identify and prevent spear phishing attacks, analyzing email patterns and content to detect impersonation and fraudulent activities, keeping organizations safe from targeted phishing threats.

6. **Google's Safe Browsing:**

   Google's system uses machine learning to spot phishing websites, showing warnings in browsers to prevent users from accessing harmful sites.

➢ APPLICATIONS OF MICRO-PROJECT

1. **Social Media Monitoring:**

   Utilize machine learning to scan social media platforms for phishing scams and malicious content.

2. **Website URL Analysis:**

   Teach a machine learning model to analyze website URLs for signs of phishing.

3. **User Behaviour Analysis:**

   Train a model to analyze user behaviour patterns, such as clicking on links or downloading attachments.

4. **Endpoint Security Solutions:**

   Integrate machine learning algorithms into endpoint security solutions to protect devices from phishing threats.

➢ SKILLS DEVELOPED OF THIS MICRO-PROJECT

1. **Cybersecurity Awareness**:

   understand how machine learning can help in fighting against them.

2. **Problem-solving and Critical Thinking:**

   Learners will improve their ability to solve real-world problems related to phishing by critically analyzing detection methods.

3. **Machine Learning Fundamentals:**

   Participants will gain a strong understanding of machine learning algorithms.

4. **Project Management:**

   You'll gain experience in project planning, execution, testing, data collection etc.

5. **Communication Skill:**

   ability to convey information clearly and effectively to other.

➢ CONCLUSION

Machine learning helps spot phishing emails quickly and accurately, protecting us from scams. It saves time and money by automating the process and catching tricky emails that could fool us. But we still need to learn about phishing and use strong passwords to stay safe online. Remember, combining machine learning with other security measures gives us the best protection against scams.

➢ REFERENCES

1. We Referred Book "A MACHINE LEARNING APPROACH TO PHISHING DETECTION AND DEFENSE" by Oluwatobi Ayodejo Akanbi and his team

2. We Referred Following Link:

    a. https://en.wikipedia.org/wiki/Machine_learning
    b. https://www.geeksforgeeks.org/machine-learning/
    c. https://www.phishprotection.com/phishing-awareness/machine-learning-helps-fighting-phishing-attacks