KARPAGAM
COLLEGE OF ENGINEERING
Rediscover | Refine | Redefine
Accredited by NAAC with 'A+' grade
Autonomous | Affiliated to Anna University
(An ISO 9001:2015 and ISO 14001:2015 Certified Institution)

Artificial Intelligence
Data Science

Hack-AI-Thon

PHISHING WEBSITE DETECTION

THEME: SECURITY

KARPAGAM
COLLEGE OF ENGINEERING
Rediscover | Refine | Redefine
Accredited by NAAC with 'A+' grade
Autonomous | Affiliated to Anna University
(An ISO 9001:2015 and ISO 14001:2015 Certified Institution)

Artificial Intelligence
Data Science

Hack-AI-Thon

GANGESH BASKER
[Gangesh Basker]

NAGA SUDHARSHAN K
[Naga Sudharshan K]

TEAM GEEKS:

Gangesh Basker
Syed Altaf SA
Naga Sudharshan K
Syed Arshad SA

SYED ALTAF S A
[Syed Altaf SA]

SYED ARSHAD S A
[Syed Arshad SA]

INTRODUCTION

- *These days internet fraudulence is tremendously huge, the websites we access are trying to acquire our personal information or data in an illegitimate way.*
- *Since there are too many websites at hand, the risk of getting your devices hacked and being scammed is substantial.*
- *Phishing websites are known for criminally acquiring passwords, private information, credit card details, locations, and any such undesirable data.*
- *Often times website appears to replicate the original site, but turns out to be a phishing site.*
- *This being the case, it is a time of immense need for an effective solution to detect and provide timely acknowledgment of whether the websites we access are safe to use.*

ABSTRACT

Hack-AI-Thon

- *Much of what we do is managed online in today's increasingly technological world. This surge in online engagement has resulted in a tremendous spike in cybercrime. Phishing has been the most powerful and harmful of all cyber-attacks. Phishing has been a critical security problem resulting in significant losses for businesses and customers. Phishing attempts are becoming more common because of inadequate identification techniques and protective methods.*

- *This project presents an enhanced approach for phishing site detection, leveraging advanced machine learning techniques. Various ML algorithms like Decision Tree, Random Forest, Multilayer Perceptrons, XGBoost, Autoencoder Neural Networks and Support Vector Machines have been compared.*

- *The proposed methodology encompasses a multi-layered framework that combines website content analysis, URL-based features, and behavioural analysis to identify potential phishing sites accurately.*

LITERATURE REVIEW

Hack-AI-Thon

- *Our model unites seven main features that enhance security.*
- *The below-mentioned category of features is extracted from the URL data:*
    - *Address Bar-based features*
        - *In this category, 9 features are extracted.*
    - *Domain-based Features*
        - *In this category, 4 features are extracted.*
    - *HTML & Javascript-based Features*
        - *In this category, 4 features are extracted.*
- *So, all together 17 features are extracted from the 10,000 URL dataset and are stored in '5.urldata.csv' file in the Data Files folder.*
- *Before starting the ML model training, the data is split into 80-20, i.e., 8000 training samples & 2000 testing samples. From the dataset, it is clear that this is a supervised machine-learning model. This data set comes under a classification problem, as the input URL is classified as phishing (1) or legitimate (0).*
- *Various machine learning models are compared based on their train and test accuracy.*

TECHNICAL PROBLEM SOLVED

- *The technical problem addressed is the development of an enhanced approach for detecting phishing sites with improved accuracy and adaptability. Traditional phishing detection methods often rely on simplistic features and fail to effectively detect sophisticated attacks. They struggle to keep pace with the evolving tactics employed by cybercriminals, making it challenging to accurately identify and differentiate between legitimate websites and malicious ones.*

- *The developed approach addresses the technical challenge of adapting to evolving phishing techniques. By leveraging machine learning algorithms , the models are trained on large-scale datasets to capture the changing patterns and characteristics of phishing sites. This allows the approach to stay up-to-date with emerging threats and provide robust and effective detection capabilities.*

- *By addressing these technical challenges, the proposed approach offers an advanced solution for the accurate and efficient detection of phishing sites. It improves the security of online users, safeguards their sensitive information, and contributes to the development of a safer digital environment.*

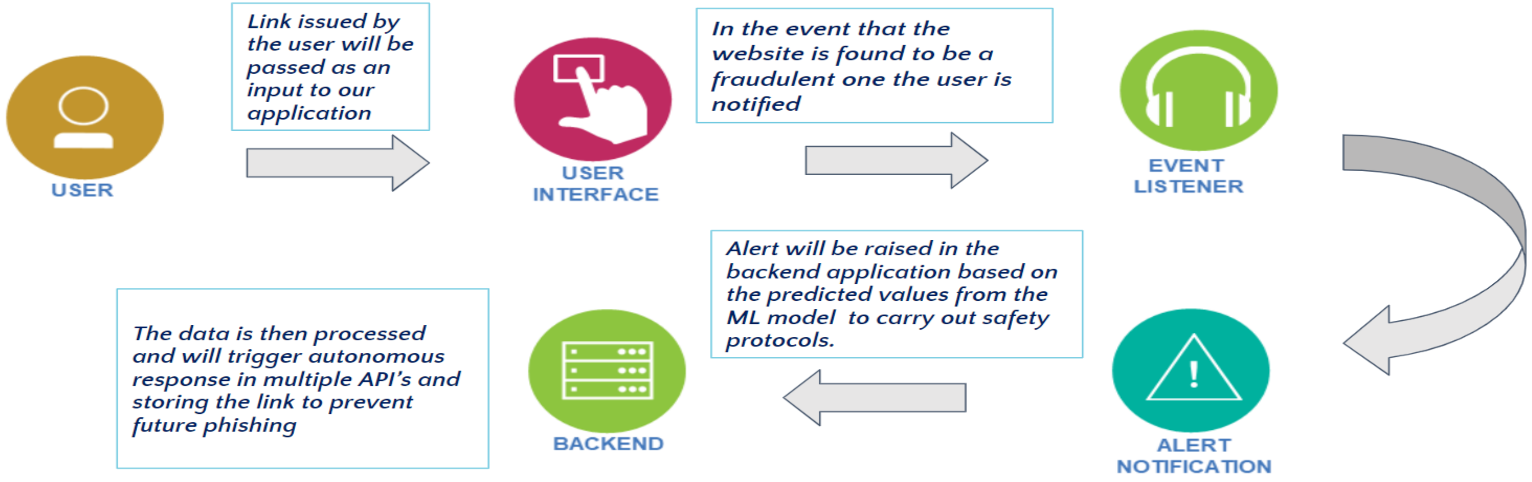
TECH STACK

- Python
- Scikit-learn
- Pandas
- Numpy
- Tensorflow
- Jupyter Notebook
- Flutter

UNIQUENESS OF THE MODEL

- Integration of multiple techniques and features:
  - *Its emphasis on textual content analysis, the inclusion of URL-based features, and the utilization of behavioural analysis. This comprehensive approach enhances the model's accuracy, adaptability, and real-time detection capabilities, providing robust protection for online users.*
- Robust Machine Learning Algorithms:
  - *The model employs advanced machine learning algorithms, such as Decision Trees, Random Forest, Support Vector Machines (SVM), or neural networks, trained on large-scale datasets. By utilizing these robust algorithms, the model can effectively learn patterns and classify websites accurately, adapting to the evolving landscape of phishing techniques.*
- Focus on Adaptability and Real-Time Detection:
  - *The model addresses the challenge of adaptability by training on diverse datasets and capturing the changing patterns of phishing sites. It aims to stay up-to-date with emerging threats and provides a defense mechanism that can adapt to evolving phishing tactics. Additionally, the model emphasizes real-time detection, enabling timely identification and mitigation of phishing attacks.*

BLOCK DIAGRAM



USER

Link issued by the user will be passed as an input to our application

USER INTERFACE

In the event that the website is found to be a fraudulent one the user is notified

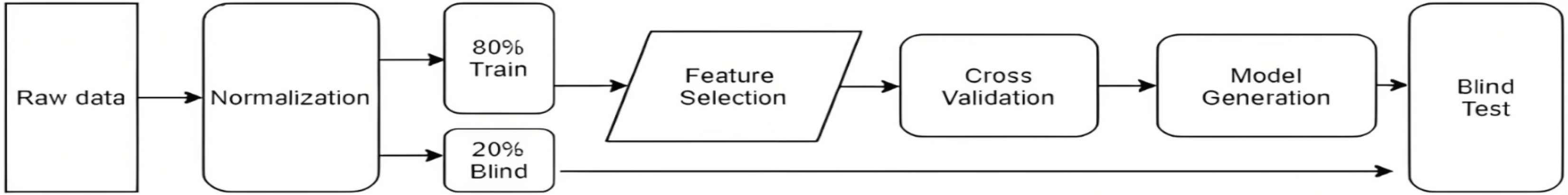
EVENT LISTENER

ALERT NOTIFICATION

Alert will be raised in the backend application based on the predicted values from the ML model to carry out safety protocols.

BACKEND

The data is then processed and will trigger autonomous response in multiple API's and storing the link to prevent future phishing



Raw data

Normalization

80% Train

20% Blind

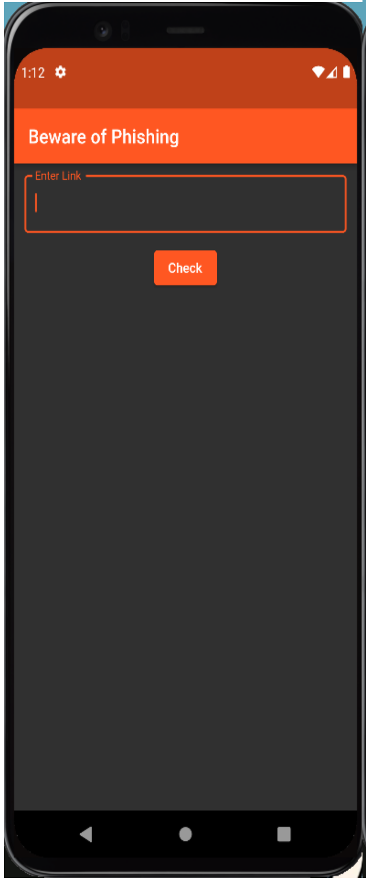
Feature Selection

Cross Validation

Model Generation

Blind Test

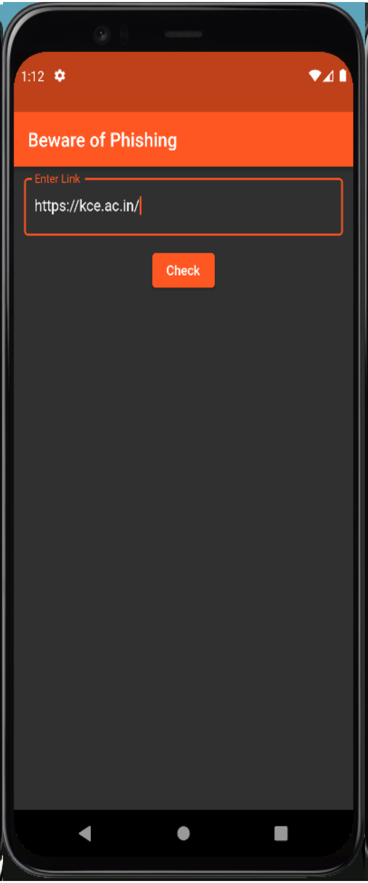
STRUCTURE OF APPLICATION


1:12

Beware of Phishing

Enter Link

Check

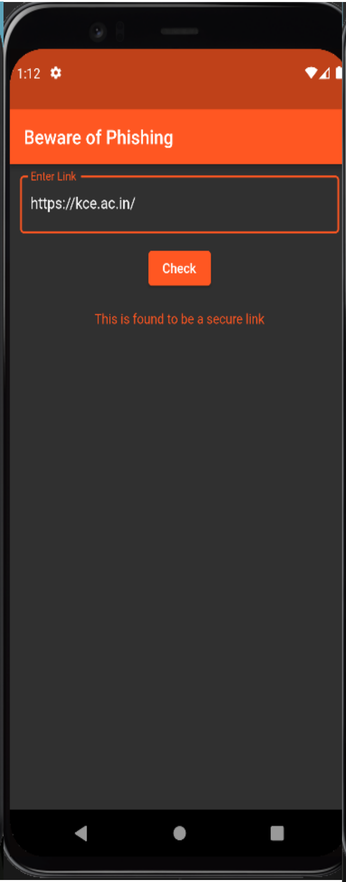

1:12

Beware of Phishing

Enter Link

https://kce.ac.in/

Check


1:12

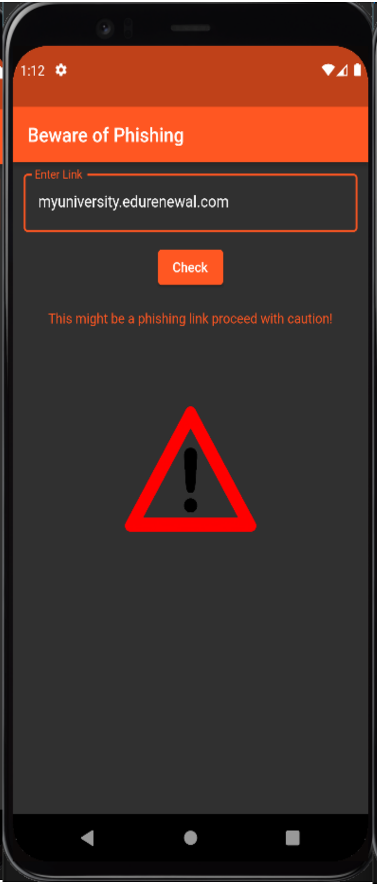
Beware of Phishing

Enter Link

https://kce.ac.in/

Check

This is found to be a secure link


1:12

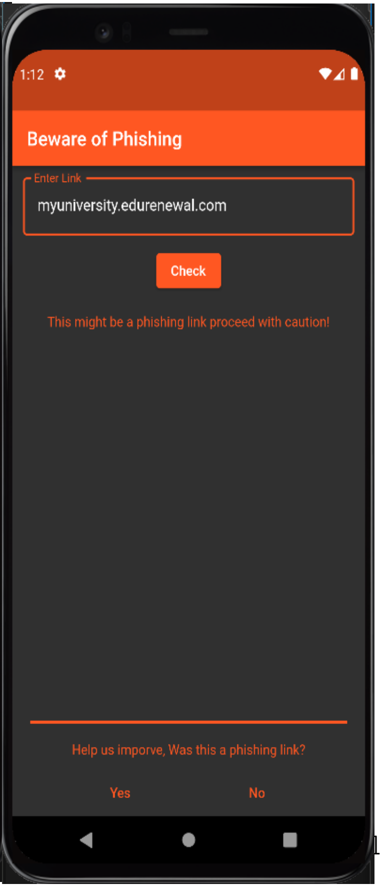
Beware of Phishing

Enter Link

myuniversity.edurenewal.com

Check

This might be a phishing link proceed with caution!


1:12

Beware of Phishing

Enter Link

myuniversity.edurenewal.com

Check

This might be a phishing link proceed with caution!

Help us imporve, Was this a phishing link?

Yes No

STRUCTURE OF APPLICATION

- *We have built a companion app to aid with our model. Which is built with flutter a cross-platform app development framework by Google which goes hand in hand with the model to help ensure the safety of the users.*

- *How the front end differs from existing solutions is by prioritising Mobile first approach. The user will be given an option to open any and all links from various platforms Through the app with which the links will be tested for genuineness and other fraudulent behaviours.*

- *Backend : Our developed application sends the URL received from the user end and transfers the input to our ML model using a request module.*
- *A well trained ML model i.e., XGBoost detects (predicts) whether the website with the given URL is safe to use.*

UNIQUE SELLING POINT

Hack-AI-Thon

- ADAPTABILITY :
  - *This is a cross platform app that works on all platforms natively*
- NETWORK EFFECT :
  - *Furthermore as we encounter new links we can forever improve on the accuracy by getting real time feedback from user*
  - *Our database is continuously growing because furthermore as we encounter new phishing links we can forever improve on the accuracy by growing the database.*
  - *Hybrid real time adaptive learning*
- WEB VIEW :
  - *It enables the user the preview the website before actually providing his saved credentials*
  - *User Interaction*
- APP INTERACTION :
  - *The major differentiating factor of our product is its mobile first approach*
  - *The app allows the user to first open the link within the app instead of a browser and verify its legitimacy before proceeding further, only when the app categorises it as a safe one ,block if the link is found to be a bad actor*
  - *Which creates a psychological effect on user as to proceed with caution*

FEASIBILITY

- *Phishing attacks account for 36% of all US data breaches and 83% of all companies experience a phishing attack each year.*

- *There was a 345% increase in unique phishing sites between 2020 and 2021 and there were 300,497 phishing attacks reported to the FBI in 2022.*

- *Each phishing attack costs corporations $4.91 million, on average .*

- *Existing Solutions :*
  - *Website*
  - *Browser extensions*
- *Our Scaled up unique Solutions :*
  - *Web view*
  - *Multi platform application*
  - *App first interaction*

SCALABILITY

- *Our main goal is to enhance user's web safety by preventing phishing attacks caused by phishing websites. Since our solution solves a technical problem that addresses the development of an enhanced approach for detecting phishing sites with improved accuracy and adaptability, implementing or getting a complete product out of it is less complicated.*

- *Therefore implementing this as a separate system or as an integrated mechanism in every android and Ios will be a primary demand soon.*

- *As the essentiality of the product is massive, the growth for products associated with web safety will have the potential to withstand its position.*

- *The further evolution of this will include Market Expansion, Strategic partnerships, continuous research and development and much more that intensify user's web safety by avoiding phishing attacks in any possible way.*

BUSINESS MODEL:

TYPE :

- *Mobile Application*
- *Software system*

APPLICATIONS :

- *Smart Phones*
- *Antivirus Software and applications*

- *The business model for this project is to provide a comprehensive phishing site detection solution to various entities, such as businesses, organizations, and individuals, who require robust protection against phishing attacks.*

- Product Offering:
  - *The core offering would be the developed phishing site detection solution. This includes the software or platform that integrates advanced machine learning algorithms, feature engineering techniques, and behavioral analysis for accurate and real-time detection of phishing sites.*

- Licensing or Subscription Model :
  - *The business model can involve licensing the detection solution to organizations or offering it as a subscription-based service. This would give customers access to the detection system, regular updates, and technical support.*

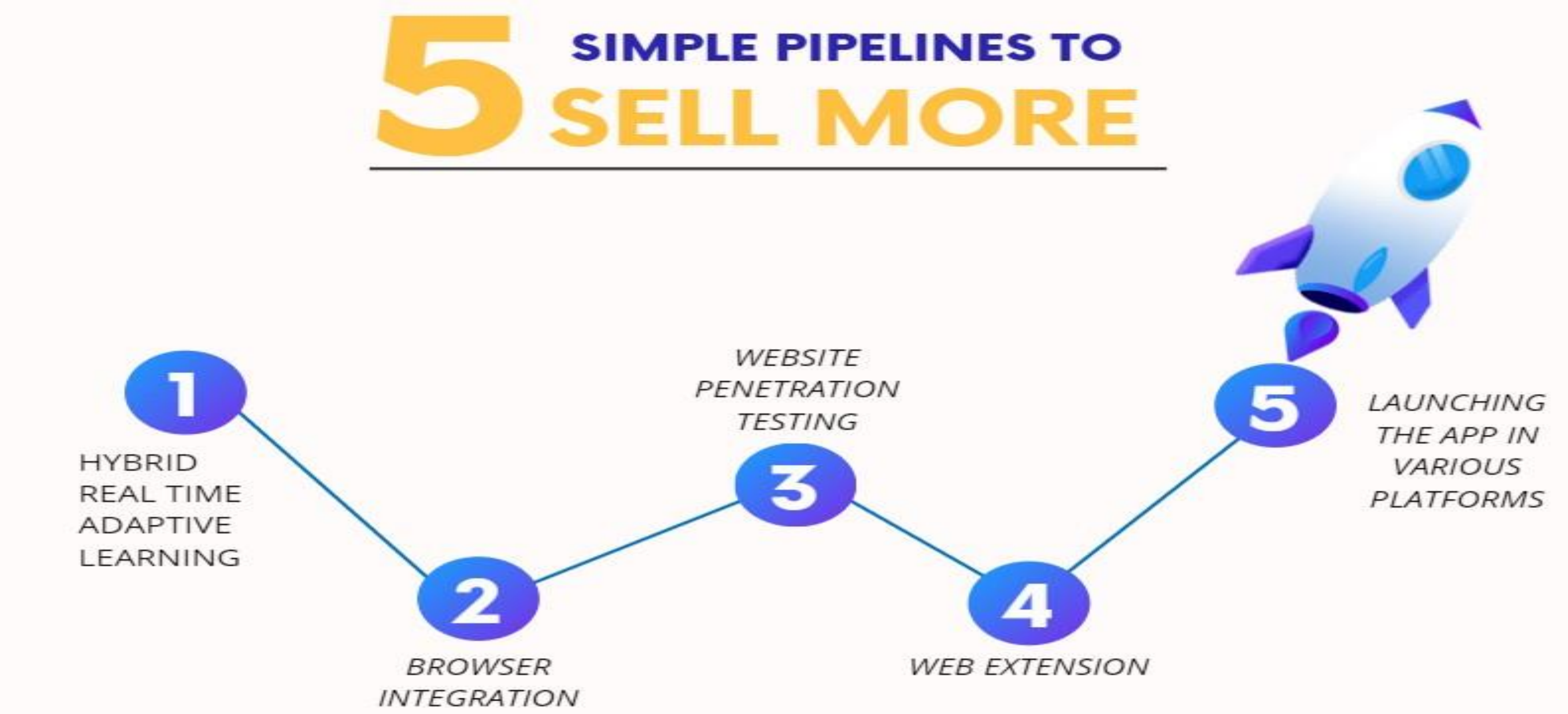
BUSINESS MODEL:

- Customization and Integration:
  - *The business model could include customization options to cater to the specific requirements of different organizations. This might involve tailoring the detection system to integrate seamlessly with existing cybersecurity infrastructure, ensuring smooth implementation and compatibility.*
- Partnerships and Collaborations:
  - *Collaboration with cybersecurity firms, IT service providers, or industry-specific organizations can enhance the business model. This could involve partnerships to integrate the detection solution into existing security suites or to collaborate on joint marketing efforts to reach a broader customer base.*
- Data Analytics and Insights:
  - *The business model could include data analytics and reporting capabilities, providing customers with insights into detected phishing attempts, trends, and potential vulnerabilities, enabling proactive measures and informed decision-making to enhance overall cybersecurity strategies.*
- Continuous Research and Development:
  - *To stay at the forefront of phishing detection, constant research and development efforts should be incorporated into the business model. This would involve staying updated with emerging threats, exploring new machine-learning techniques, and refining the detection system to adapt to evolving attack methods.*

REVIEWING TESTIMONIALS

- *Recently, a person we know had a massive financial loss due to a phishing scam attack.*
- *The existing phishing detection solutions primarily focus on detecting a mail-based phishing attack and majorly work on desktop systems.*
- *These days, prey for phishing websites are predominantly mobile users via messaging apps and other mobile platforms.*
- *There are not much existing models that provide phishing detection services for mobile users.*
- *This project is in its final stage of development, which is undergoing its testing phase by a limited number of consumers.*
- *Henceforth, there are not many testimonials available to be showcased as of now.*

PIPELINE



5 SIMPLE PIPELINES TO SELL MORE

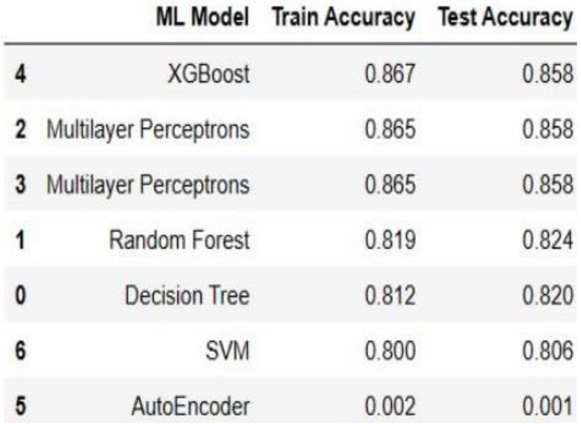
1. HYBRID REAL TIME ADAPTIVE LEARNING
2. BROWSER INTEGRATION
3. WEBSITE PENETRATION TESTING
4. WEB EXTENSION
5. LAUNCHING THE APP IN VARIOUS PLATFORMS

CONCLUSION

- *From the obtained results of the above models, XGBoost Classifier has the highest model performance of 86.4%. So the model is saved.*

- *As of now, We are working towards creating a browser extension for this project.*

- *The machine learning models shown here can be easily served as REST API endpoints which can further be used with add-ons to detect phishing websites in real time.*



	ML Model	Train Accuracy	Test Accuracy
4	XGBoost	0.867	0.858
2	Multilayer Perceptrons	0.865	0.858
3	Multilayer Perceptrons	0.865	0.858
1	Random Forest	0.819	0.824
0	Decision Tree	0.812	0.820
6	SVM	0.800	0.806
5	AutoEncoder	0.002	0.001

Github Demo video and Blog Post link

- Github : Repository Link
- Demo video : Demonstration link
- Blog Post : Blog Post

THANK YOU