Machine Learning Based Password Strength Analysis

Sony Kuriakose, G Krishna Teja, Sravan Duggi, A Harshel Srivatsava, Venkat Jonnalagadda

Abstract**: *Passwords, as the most used method of authentication because to its ease of implementation, allow attackers to get access to the accounts owned by others by means of cracking passwords. This is cause of the similar patterns that users use to create a password, like dictionary words, common phrases, person and location names, keyboard pattern, and so on. Multiple password cracking techniques had been introduced to predict the password offline or online, with the majority of records say the one with weak password or familiar password patterns being cracked. This suggested prototype implements numerous machine learning methods such as Decision Tree (DT), Nave Bayes (NB), Logistic Regression (LR), and Random Forest (RF) on a web application in real time to force users to choose a secure password. This results in the user's account being logged into if particularly the password strength from more than half of the algorithms is strong.***

Keywords**: *Passwords, password strength, password analysis, machine learning.***

I. INTRODUCTION

This Web Application is a program which runs on network-connected gateways and are becoming more important for completing various activities in corporate, scholastic, social and other contexts. The web apps are connected to back end directories that store a large amount of data, such as usernames and passwords, and would be used for online connected exchanges, data storage, access to social groups, and so on. Without regard to the importance of the web apps, they provide a route for programmers or hackers to barge through these data sources. Protecting the web information should always be a primary concern for web application developers. SQL Injection Attack (SQLI), Cross-Site Scripting (XSS), XML External Entities (XXE), Broken Authentication, and a few other attacks affect about 98 percent of web applications. Password-based authentication is the initial and most fundamental strategy in the world of cyber security to effectively defend web information. Its popularity stems from the ease with which it may be implemented, as it does not require any expensive software or specialized hardware. Even though this approach has numerous flaws, yet it is widely applied. Passwords are closely guarded sequence of letters which are used in verifying id or to get entree to a service. Passwords may be required by a typical PC client for a variety of reasons, including logging into accounts, moving services, window shopping, accessing applications, databases, institutions, and websites, and, also for reading the newspaper online. Password is important not just for logging in, also for many other advanced service-providing system like Kerberos. Clients must use different passwords for multiple frameworks or systems for a number of reasons, including obvious security concerns, making it more difficult to keep note of in head and secure the users password. User typically would come up with a password that is basic, not hard to remember, and comfortable to recall, totally ignoring the password's strength. Many optional confirmation components (e.g., multi-layered validation) has been proposed in recent years, but passwords stubbornly remain and replicate with each new web system. This is mostly due to the fact that each of these choices has different disadvantages when compared to passwords. Researchers have gone through several phases and methodologies to truly understand password security, including algorithms, Markov-based, probabilistic context-free language and other algorithms that use identifying information (PII). The info leak incidents that has caused so much concern [1] around the cyber security field attracted these techniques. The strength of a password is a measure of its resistance to predicting and other types of password intrusions such as brute-force and dictionary-attacks. The span and intricacy of a password are considered the most critical attributes in keeping it secure, with length representing the total number of characters utilized and intricacy representing the practice of a variety of characters such as digits, upper-cases, lower-cases and also symbols. Although using a big and complicated password reduces danger of being cracked, security cannot be guaranteed. Any password can be hacked, although certain passwords take lesser time to crack and others take longer time. Many commercial password strength tools based on linguistic criteria have been developed in the field of password strength checking, such as Google Password Meter (GPM, 2008), Microsoft Password Checker (MPC, 2008), Password Meter (PM, 2008), and others. To add on to these methods, Decision Trees, En-filter, and other tools have been developed for testing password strength [2-4].

Manuscript received on 03 June 2022 | Revised Manuscript received on 08 June 2022 | Manuscript Accepted on 15 July 2022 | Manuscript published on 30 July 2022.

* Correspondence Author

Mrs. Sony Kuriakose, Assistant Professor, Department of Information Science and Engineering, New Horizon College of Engineering, Bangalore (Karnataka) India.

G Krishna Teja*, Department of Information Science and Engineering, New Horizon College of Engineering, Bangalore (Karnataka) India.

A Harshel Srivatsava, Department of Information Science and Engineering, New Horizon College of Engineering, Bangalore (Karnataka) India.

Sravan Duggi, Department of Information Science and Engineering, New Horizon College of Engineering, Bangalore (Karnataka) India.

Venkat Jonnalagadda, Department of Information Science and Engineering, New Horizon College of Engineering, Bangalore (Karnataka) India.

Hash Cat, John-the-Ripper, Pass-GAN, Tar-Guess I-IV, and other password guessing programmes have also been developed.

II. RELATED WORK

Several studies in the topic of password strength analysis and detection have been undertaken using alternative strategies, with the findings provided below. It was discovered in 2012 that password strength indicators efficiently increase password strength while also delivering very accurate password comments. However, according to a 2015 study, the great majority of people assume their credentials would safeguard them, which is not the case [5-6]. This is partly due to the false assertions made by the password strength metres. The randomness approach is a highly effective ad - hoc basis password strength metre. It employs heuristic criteria. Google, Yahoo, PayPal, Windows, and other digital PSMs employ the NIST password strength metre. This entropy-based method, however, has been proved to be ineffective When actual speculating blocking is taken into account, however, this entropy-based technique only delivers an inadequate approximation of password strength. A more convincing assessment for secret phrase strength is "guessability," that recognises with absolute security and quantifies the time complication necessary for a cryptanalysis calculation to retrieve a record [7-8]. There are two types of guessability: physical and digital guessing. Further trawling guessing and targeted guessing are two different forms of guessing that differ in whether or not they use user data.

III. METHODOLOGY

The developed model's main goal is to use numerous machine learning algorithms to evaluate password strength in a real-time web application.

- The first stage focuses on gathering a suitable dataset that contains a large number of passwords that are used to test password strength. We have gathered a dataset for this phase that includes passwords of three different strengths: weak, medium, and strong.
- The second stage emphasizes on preprocessing techniques and extracting features from the dataset using train test split. This is the stage where the dataset is labelled. After then, 70% of the dataset is used to train the models (a.k.a. Training Part).
- The third stage concentrates on testing and assessing the suggested model using the 30% of the dataset that we segregated from the actual gathered dataset (a.k.a. Testing Part).
- The last stage focuses on putting the proposed model on the web application that we created for the project to analyze the password strength in real time. The proposed model's main goal is to force users to choose a strong password.

A. Dataset

Define Collecting a meaningful dataset including a large number of passwords of weak, medium, and strong strength is the most crucial aspect that determines the strength of passwords using machine learning. The key contribution of this research is a manually produced labelled dataset for the topic. The passwords in the dataset are labelled as 0, 1, and 2 for weak, medium, and strong passwords, respectively.

This phase generates the following types of passwords and they are weak passwords, medium passwords, and strong passwords. We collected these passwords in text (.txt) format and also used several pre-processing methods to convert them to a csv (.csv) file. The database has 800,000 passwords. Each of the three groups in the dataset has its own set of conditions:

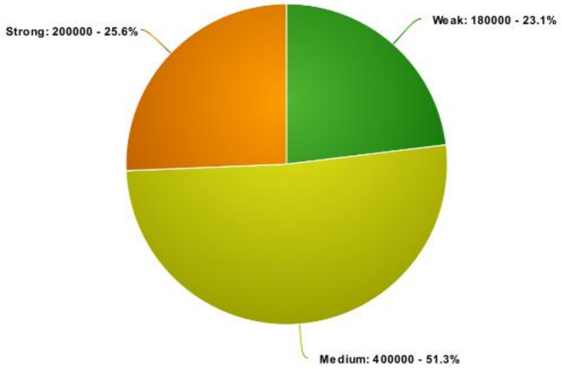


Fig. 1. Class Count

B. Activity Diagram

The activity diagrams show operations that involve options, recurrence, and synchronization. In UML, activity diagrams may be used to show the key business phase activities of network elements. The whole control flow is depicted in an activity diagram.

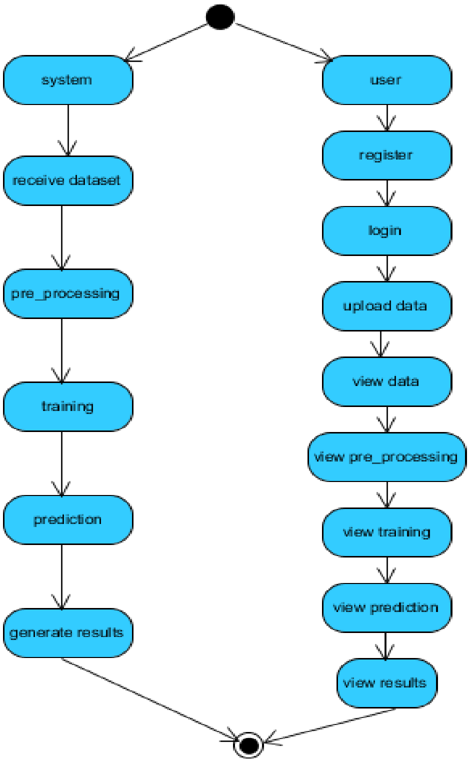


Fig. 2. Activity Diagram

C. Class Diagram

A class diagram is a type of nonlinear structural diagram in the uml that depicts the project's classes, characteristics, functions and connections amongst some of the classes. It specifies which class is in charge of info.

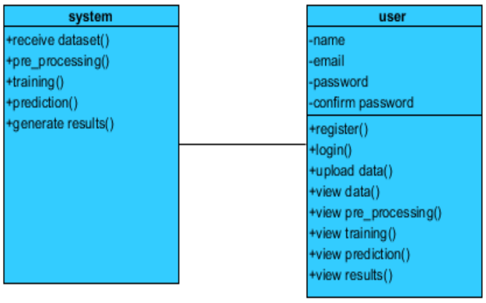


Fig. 3. Class Diagram

D. DFD Diagram

A Data Flow Diagram (DFD) is a typical visual representation of how data flows in via a system. A simple and brief DFD may graphically depict a significant percentage of the system's needs. This can be done by hand, by machine, or by combining the two. It shows how info enters and exits the network, and also how the data is changed and stored. The basic goal of a DFD is to determine a system's total nature and extent. It may be used as a medium of communication here between systems engineer and almost everyone involved in the organization, as well as a starting point for redesigning the system. Level 1 Diagram is listed below in Fig.5.

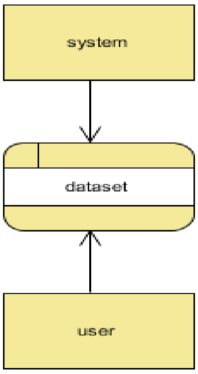


Fig. 4. DFD Components

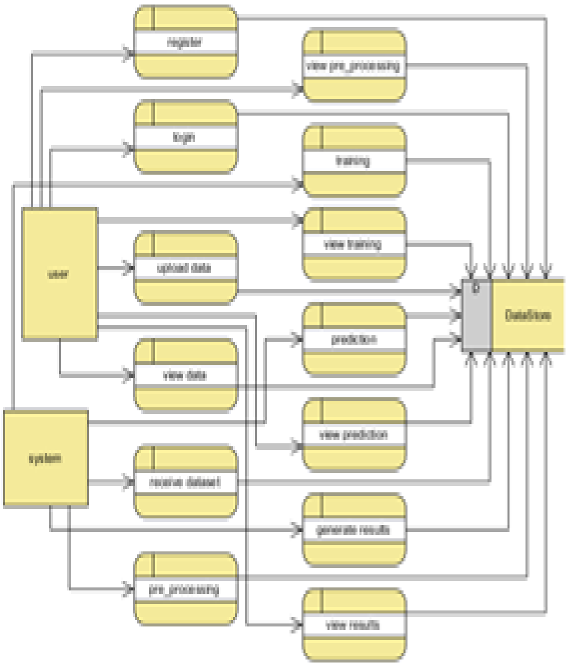


Fig. 5 Level 1 DFD Diagram

E. Training Models

The major part comprises training machine learning algorithms for password strength detection over a carefully gathered dataset and then deploying the models for real-time password strength analysis. However, there are some preprocessing procedures that must be completed on the dataset first. We merge the three text files containing weak, medium, and strong passwords and convert them to a csv (.csv) file. The missing values is then managed, we would have to replace it with the means, but we chose to drop to and rearrange it for robustness' reasons. We then employed pipelines for Decision Tree (DT), Nave Bayes (NB), Logistic Regression (LR) and Random Forest (RF) were the machine learning methods used in the suggested study. The trained models would be stored as joblib files (.joblib), which would be served while the suggested model was implemented in real time on a web - based application.

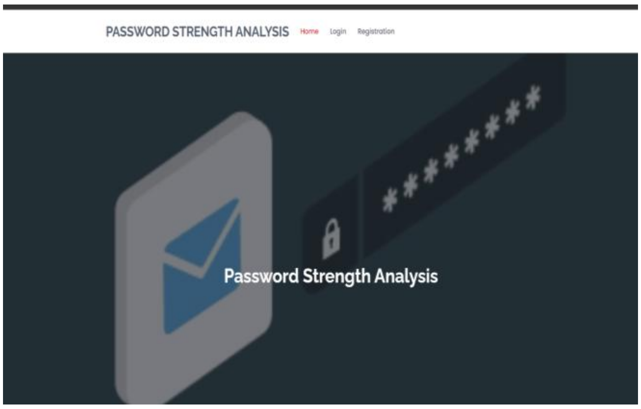
IV. RESULT AND DISCUSSION

A. Testing

As According to our tests, our proposed system is capable of detecting and analyzing password strength. On the testing data, we used the algorithms to assess the performance of our suggested model. As According to our tests, our proposed system is capable of detecting and analysing password strength. On the testing data, we used the algorithms to assess the performance of our suggested model.

B. Real Time Analysis

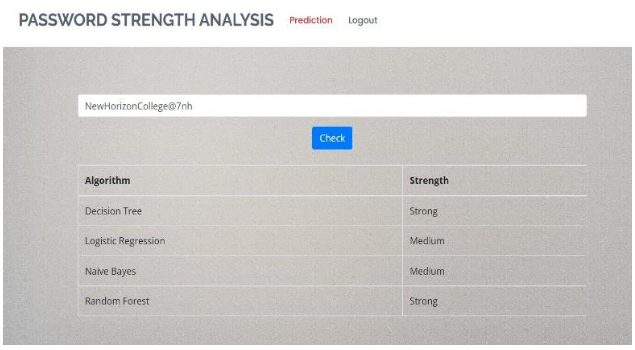
Multiple computer languages, including Python, HTML5, CSS3, and the Flask framework, are used to create the proposed system. HTML5 and CSS3 are used to create a webpage along with some Flask web framework. SQLyog is used to connect to the back-end database. The Python programming language is utilized to implement real-time password analysis.



The user cannot proceed to register their account until the strength of the password has been determined, and only then may the user access the account if the strength of two or more (more than 50%) algorithms has been determined to be strong. We used the model to guess several passwords and got the results shown below.

Password	DT	LR	NB	RF
Nhce@123	Weak	Medium	Strong	Weak
Nhce@#0123	Medium	Strong	Medium	Medium
NhcE07@123ok	Medium	Strong	Strong	Strong

The user is now constrained to come up with a strong password that is acceptable by the model in terms of its conditions. Thus it increases the complexity of the password the user desires to select for their account, therefore it does gets hard for the attackers to gain access of our accounts.



NewHorizonCollege@7nh

Algorithm	Strength
Decision Tree	Strong
Logistic Regression	Medium
Naive Bayes	Medium
Random Forest	Strong

V. CONCLUSION

The major takeaway from this project is that the user will able to scale up the strength of the password they decide to enter for their account according to the analysis provided in the application before creating the password. This will help the user to not stay on the vulnerable side of the cyber password attacks which is easily possible in case of weak or common passwords. Hence, this project will be executed using different algorithms such as Random forest, Naïve Bayes, Logistic Regression and Decision Tree to perform analysis of efficiency for password strength analysis in order to produce the best results for the user.

REFERENCES

1. Grassi P.A., Garcia M., Fenton J. NIST Special Publication 800–63–3 Digital Identity Guidelines National Institute of Standards and Technology, Los Altos, CA (2020)
2. Zhou Huan, Liu Qixu, Cui Xiang, Zhang Fangjiao. Research on Targeted Password Guessing Using Neural Networks. Journal of Cyber Security. 2018. 3(5): p. 25-37.
3. M. Weir, S. Aggarwal, B. d. Medeiros and B. Glodek, "Password Cracking Using Probabilistic Context-Free Grammars," 2009 30th IEEE Symposium on Security and Privacy, Berkeley, CA, 2009, pp. 391-405, doi: 10.1109/SP.2009.8. [CrossRef]
4. J. Blocki, B. Harsha and S. Zhou, "On the Economics of Offline Password Cracking," 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, 2018, pp. 853-871, doi: 10.1109/SP.2018.00009. [CrossRef]
5. Melicher, W., et al., Better passwords through science (and neural networks). ; 2017. 42(4).
6. A LeCun, Y., Y. Bengio, and G. Hinton, Deep learning. Nature, 2015. 521(7553): pp. 436-444. [CrossRef]
7. B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor, “How does your password measure up? The effect of strength meters on password creation,” in Proceedings of the 21st USENIX conference on Security symposium, ser. USENIX-SS ’12. USENIX Association, 2012, pp. 65–80.
8. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191. Workshop (NSPW ’05), pp. 67-72, Sept. 2005. [CrossRef]

AUTHORS PROFILE

Mrs. Sony Kuriakose is presently working as an Assistant Professor in the Department of Information Science and Engineering at New Horizon College of Engineering, Bangalore and she is a research scholar of Amrita University which is located in Bangalore, Karnataka. She has completed her B-Tech Information Technology in the year 2014 at Kerala University and has completed her M-Tech in Computer Science and Engineering in the year 2016 in Kerala University. She has guided and worked with this batch of students on a project named Machine Learning Based Password Strength Analysis’ which is mostly implemented using Python language and Machine Learning algorithms.

G Krishna Teja (1NH18IS033) is presently studying in the Department of Information Science and Engineering at New Horizon College of Engineering, Bangalore affiliated to Visvesvaraya Technological University, Bangalore and has worked on a project named Machine learning based Password Strength Analysis which uses Machine Learning algorithms and Python Language in the final semester of his college. He was born and brought up in Hyderabad, Telangana. He has finished his Intermediate at Page Junior College in the year 2018 which is located in Hyderabad, Telangana and completed his schooling at Meridian School, Banjara Hills in the year 2016 in the city Hyderabad, Telangana.

A Harshel Srivatsava (1NH18IS001) is presently studying in the Department of Information Science and Engineering at New Horizon College of Engineering, Bangalore affiliated to Visvesvaraya Technological University, Bangalore and has worked on a project named Machine learning based Password Strength Analysis which uses Machine Learning algorithms and Python Language in the final semester of his college. He was born and brought up in Tirupati, Andhra Pradesh. He has finished his Intermediate at Sri Chaitanya Junior College in the year 2018 in Tirupati, Andhra Pradesh and completed his schooling at Sri Chaitanya Techno School in the year 2016 in the city Tirupati, Andhra Pradesh.

Sravan Duggi (1NH18IS032) is presently studying in the Department of Information Science and Engineering at New Horizon College of Engineering, Bangalore affiliated to Visvesvaraya Technological University, Bangalore and has worked on a project named Machine learning based Password Strength Analysis which uses Machine Learning algorithms and Python Language in the final semester of his college. He was born and brought up in Nellore, Andhra Pradesh. He has finished his Intermediate at Narayana Junior College in the year 2018 in Nellore, Andhra Pradesh and completed his schooling at Roshini Ratnam English Medium High School in the year 2016 in the city Nellore, Andhra Pradesh.

Venkat Jonnalagadda (1NH18IS042) is presently studying in the Department of Information Science and Engineering at New Horizon College of Engineering, Bangalore affiliated to Visvesvaraya Technological University, Bangalore and has worked on a project named Machine learning based Password Strength Analysis which uses Machine Learning algorithms and Python Language in the final semester of his college. He was born and brought up in Nandigama, Andhra Pradesh. He has finished his Intermediate at Sri Chaitanya Junior College in the year 2018 in Vijayawada, Andhra Pradesh and completed his schooling at Apollo English Medium School in the year 2016 in the town Nandigama, Andhra Pradesh.