Nilai Penting dan Penjelasan Fitur Pada Pengayaan Pendeteksian Web Phishing Menggunakan Pendekatan Explainable Artificial Intelligence = The Feature Importance and Explanations in Enhancing Web Phishing Detection Using an Explainable Artificial Intelligence Approach

Abdullah Fajar, author

Abstrak

Secara umum praktik praktik web phishing semakin kompleks dan sulit dideteksi, sehingga diperlukan pendekatan yang lebih maju untuk menghadapinya. Selain itu kesadaran dan kepercayaan pengguna terhadap sistem deteksi phishing perlu ditingkatkan dalam kerangka pemahaman tentang bagaimana sistem pendeteksian bekerja dan alasan di balik hasil yang diberikan. Penelitian penggunaan metode pendeteksian phishing menggunakan Machine Learning sifat pemrosesannya masih berupa Black-box yaitu proses pengambilan keputusannya tidak diketahui. Penelitian ini bertujuan membantu memberikan penjelasan pemahaman terhadap hasil pendeteksian phishing. Pendekatan XAI (eXplainable Artificial Intelligence) untuk pendeteksian phishing memberikan tambahan penjelasan fitur-fitur yang diekstraksi dari web berupa URL dan struktur HTML yang berkontribusi dalam hasil pendeteksian phishing. Selain memberikan kejelasan kontribusi juga dapat digunakan untuk menilai model menggunakan metrik kinerja, dan efek kausalitas dari fitur-fitur tersebut.

Penelitian menggunakan kombinasi algoritma yang termasuk dalam model White-box seperti algoritma Explainable Boosting Machine (EBM) dan model Black-box populer yaitu Random Forest, XGBoost dan CatBoost. Algoritma tersebut akan diukur dan dibandingkan kinerjanya pada sekumpulan dataset Phishing bersumber dari beberapa situs. Selain diukur kinerjanya, model yang dihasilkan diuraikan atau dijelaskan menggunakan metode XAI yaitu SHAP (Shapley Additives exPlanation). Adapun penjelasan hasil XAI divisualisasikan menggunakan Plot Nilai Penting Fitur dan Efek Kasualitas Model. Setiap Plot yang dihasilkan dianalisis dan dirangkum untuk mendapatkan temuan atau jawaban atas pertanyaan riset yang relevan.

Hasil yang didapatkan dari penelitian ini menunjukkan beberapa temuan penting bahwa ada korelasi dimensi dataset dengan metrik kinerja. Dari metrik penjelas, plot Nilai Penting Fitur memberikan beberapa gambaran kemampuan model terhadap dataset baru dan juga memberikan rekomendasi fitur-fitur yang perlu digunakan kembali dalam model. Penjelasan hasil metode XAI bersifat global artinya merupakan gambaran utuh perilaku dan kontribusi fitur terhadap model. Secara umum algoritma Black-box XGBoost mempunyai kinerja yang baik. Dalam perspektif fitur dalam dataset phishing, Fitur-fitur seperti url_length, n_slash, n_dots, SSLfinal_State, dan URL_of_anchor konsisten menjadi fitur paling signifikan di berbagai model.

.....A more sophisticated strategy is required to combat web phishing tactics since they are generally growing more intricate and challenging to identify. Furthermore, in order to better comprehend how detection systems operate and the rationale behind the findings they produce, user awareness and trust in phishing detection systems must be raised. The processing nature of research on machine learning-based phishing detection techniques is still a "black box," meaning that the approach used to make decisions is unknown. The purpose of this study is to contribute to the understanding and explanation of phishing detection outcomes. Additional explanations of online elements like URLs and HTML structures that are

extracted and contribute to the phishing detection findings are provided by the XAI (eXplainable Artificial Intelligence) approach. It can be used to assess the model using performance metrics and the causal impacts of certain features, in addition to offering clarity on contributions.

In this study, the Explainable Boosting Machine (EBM) algorithm and other White-box algorithms are combined with well-known Black-box models including Random Forest, XGBoost, and CatBoost. The algorithms' performance will be evaluated and contrasted using a collection of phishing datasets from various websites. The resulting model is described or explained using the XAI method, namely SHAP (Shapley Additives exPlanation), in addition to its performance being measured. Model Causality Effects and Feature Importance Value Plots are used to illustrate the explanation of the XAI results. Every plot that is produced is examined and condensed to produce conclusions or responses to pertinent research questions. There is a correlation between the dataset's dimensions and performance measures, according to the study's conclusions, which highlight numerous significant findings. The Feature Importance plot derived from the explanatory metrics In addition to suggesting which characteristics should be kept in the model, features give some insight into how well the model performs with new datasets. The outcomes of the XAI approach are explained globally, which gives a thorough understanding of how features behave and contribute to the model. The Black-box XGBoost algorithm works well overall. Features like url_length, n_slash, n_dots, SSLfinal_State, and URL_of_anchor are consistently the most important features in the phishing dataset when viewed from the standpoint of features in different models.