

# INDIAN INSTITUTE OF INFORMATION TECHNOLOGY , LUCKNOW

# REPORT ANALYSIS

# NAME :Pradeep Dubey, Aman Kumar, Apoorva, Anshi Gupta

# BRANCH : Msc. Data Science

# SUBJECT : Natural Language Processing

# Cybersecurity Analysis and Prediction Application using CVE Dataset

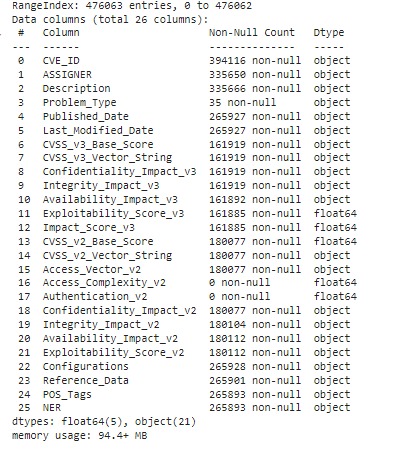
## Project Overview:

This project focuses on analyzing and predicting cybersecurity vulnerabilities using data from the CVE (Common Vulnerabilities and Exposures) database. The application extracts, processes, and models vulnerability data, aiming to assess and predict potential impacts, scores, and other relevant metrics. Leveraging machine learning, NLP, and visualization tools, the project enhances vulnerability assessment efficiency and decision-making for cybersecurity professionals.

## Project Workflow:

### 1. Data Extraction

- Process: Data was extracted from the NIST website using Selenium, a Python library for web scraping and automation. The data was initially available in JSON format.  
- Transformation: Relevant entities were parsed from JSON, converted into Excel format, and preprocessed for analysis.  
- Natural Language Processing (NLP): We used word tokenization followed by Part-of-Speech (POS) tagging with the Viterbi algorithm, dependency parsing, and Named Entity Recognition (NER) to identify and categorize important keywords and entities within each entry's description.



### 2. Data Cleaning

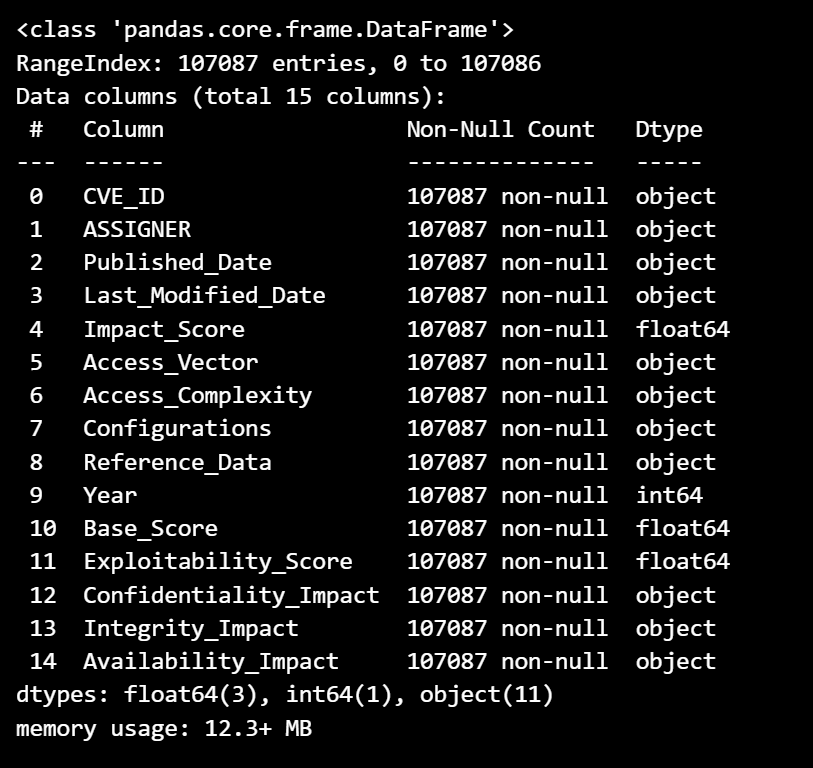
- Null Values Handling: The dataset contained numerous missing values. We removed these or replaced them based on specific rules (e.g., combining categorical data by selecting the latest version).  
- Parameter Consolidation: Different versions of similar parameters were harmonized. Numerical parameters were aggregated using averages, while categorical parameters were reduced to their latest version.  
- Result: After this cleaning process, the dataset expanded to 107,087 rows—a growth of around 30,000 entries due to the combination of similar parameters.

### 3. Description Embeddings with Sentence Transformer

- Embedding Generation: We used a pre-trained Sentence Transformer model to generate embeddings for each vulnerability description. This process created a feature matrix with a dimension of 107,087 x 767.  
- Performance Considerations: Due to the large dataset size, this step was computationally intensive, requiring significant processing time.

### 4. Data Merging

We merged the generated embeddings with the rest of the processed CVE data, resulting in a comprehensive dataset combining both textual and numerical features. This hybrid dataset enables more accurate and nuanced model predictions.



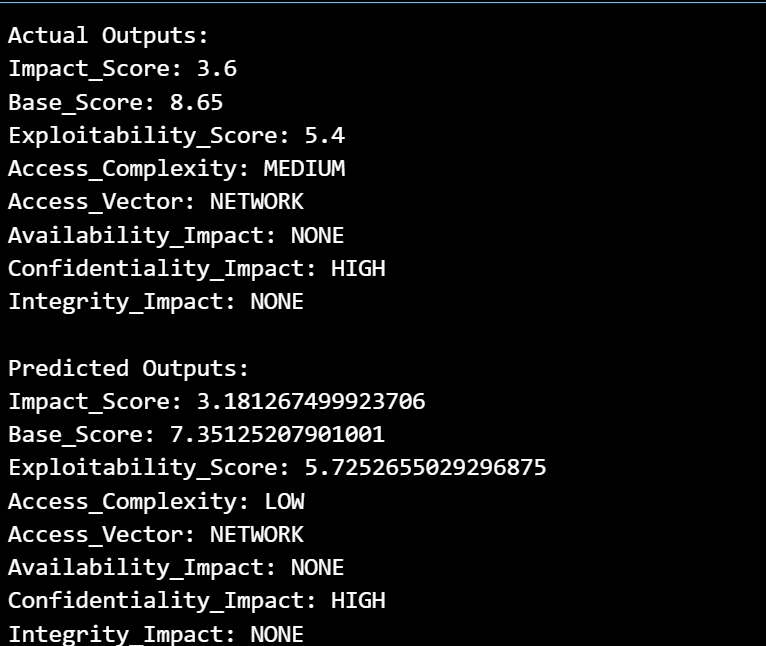
### 5. Model Building and Prediction

- Model Selection: We experimented with several ensemble models known for their efficiency on large datasets, including:  
 - XGBoost  
 - LightGBM  
 - CatBoost  
- Evaluation and Selection: Using GPUs to accelerate training, we selected the model with the highest prediction accuracy for various parameters like impact, scores, and vulnerability classification. The models take description embeddings as inputs to predict risk-related outcomes.

# Model Evaluation Metrics for different parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Description | Purpose | XGBoost GPU Metrics | LightGBM GPU Metrics / Other Model Metrics |
| Impact Score | Measures the overall effect of a vulnerability on confidentiality, integrity, and availability. | Assesses severity and prioritizes vulnerabilities for remediation. | MSE: 1.4874, MAE: 0.9358, R²: 0.4681 | MSE: 1.5338, MAE: 0.9703, R²: 0.4515 (LightGBM) |
| Base Score | Represents the fundamental risk level of a vulnerability, which remains constant over time. | Provides a standardized measure of vulnerability severity. | MSE: 1.4874, MAE: 0.9358, R²: 0.4681 | MSE: 1.5338, MAE: 0.9703, R²: 0.4515 (LightGBM) |
| Exploitability Score | Indicates how easily a vulnerability can be exploited. | Helps prioritize based on the difficulty of exploitation. | MSE: 1.1731, MAE: 0.8116, R²: 0.4526 | MSE: 1.2120, MAE: 0.8494, R²: 0.4344 (LightGBM) |
| Access Complexity | Describes the difficulty level of exploiting the vulnerability after initial access is gained. | Higher complexity reduces the likelihood of exploitation, impacting overall risk. | Accuracy: 0.8501, MSE: 0.1671, MAE: 0.1557 | Accuracy: 0.8517, MSE: 0.1670, MAE: 0.1545 (CatBoost, Best Model) |
| Access Vector | Refers to how the vulnerability can be accessed (e.g., local or network). | Remote access vectors generally indicate higher risk due to greater exposure. | Accuracy: 0.9133, MSE: 0.1291, MAE: 0.1008 | Accuracy: 0.9153, MSE: 0.1271, MAE: 0.0988 (CatBoost, Best Model) |
| Availability Impact | Measures the potential disruption of service or access to resources due to a vulnerability. | Important in environments where uptime is critical, impacting prioritization. | Accuracy: 0.8713, MSE: 0.4592, MAE: 0.2389 | Accuracy: 0.8729, MSE: 0.4511, MAE: 0.2351 (CatBoost, Best Model) |
| Confidentiality Impact | Evaluates the extent to which unauthorized disclosure of data is possible. | Crucial for systems handling sensitive information; higher impact requires prompt attention. | Accuracy: 0.8372, MSE: 0.4241, MAE: 0.2499 | Accuracy: 0.8388, MSE: 0.4199, MAE: 0.2475 (CatBoost, Best Model) |
| Integrity Impact | Assesses the potential for an attacker to alter or corrupt data within a system. | Essential for data-critical applications, with high impact indicating a risk to data accuracy or trust. | Accuracy: 0.8418, MSE: 0.4629, MAE: 0.2598 | Accuracy: 0.8434, MSE: 0.4506, MAE: 0.2546 (CatBoost, Best Model) |

**For the evaluation purpose we have compared our predicted output with the actual output for single entry(description).**



### 6. RAG based implementation

-Predicting similar vulnerabilities based on given description: This tool helps in identifying and analyzing similar vulnerabilities based on textual descriptions, which can be useful in cybersecurity assessments and understanding related vulnerabilities. It uses NLP embeddings and cosine similarity to perform this similarity search, allowing users to explore CVE records based on description context and relevant impact metrics.

### 7. Application Development with Streamlit

- Integration of Models: All models were integrated into a single user-friendly application using Streamlit, which allows users to perform predictions and data analysis.  
- Features:  
 - Prediction Module: Predicts impacts, scores, and other metrics based on the vulnerability descriptions.  
 - Data Analysis Module: Visualizes CVE data across various time ranges.  
- Data Visualization: The application supports multiple types of analysis, including:  
 - Categorical vs. Categorical Analysis  
 - Numerical vs. Numerical Analysis  
 - Categorical vs. Numerical Analysis

## Technical Details and Insights:

- POS Tagging and Dependency Parsing: These steps helped to enhance the relevance of extracted entities, ensuring the quality of inputs for subsequent analysis.  
- Named Entity Recognition (NER): This step was crucial for identifying key entities, such as product names, companies, and specific software versions affected by vulnerabilities.

- Ensemble Models with GPU Acceleration: Given the large dataset, using models optimized for GPU processing like XGBoost, LightGBM, and CatBoost enabled faster model training and more robust predictions.  
- Embedding Utilization: Embeddings from the Sentence Transformer allowed the models to incorporate nuanced information from vulnerability descriptions, which improved the prediction of impact levels and risk scores.

- Date Range Filtering: Users can filter CVE records by date, helping them focus on vulnerabilities within specific timeframes for trend analysis.  
- Interactivity: The application’s interactive visualizations facilitate in-depth analysis, allowing users to identify patterns in vulnerability types and their impact across software versions or product lines.

## Conclusion:

This project has demonstrated the potential of leveraging the CVE dataset to improve vulnerability assessment using advanced NLP and machine learning techniques. The application allows cybersecurity professionals to analyze and predict vulnerability impacts efficiently, thus supporting proactive security measures.

THANK YOU