

**School of Computer Science and Engineering**

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

**Final Review Report**

**Program :** Integrated Mtech

**Course :** Predictive analytics with case studies

**Slot :** F2

**Faculty :** Dr. Sajidha

**Component :**  J

**Title** **: Intrusion Detection using ML and Ensemble Learning**

**ACKNOWLEDGEMENT**

We wish to express our sincere thanks and a deep sense of gratitude to our project guide, Dr. Sajidha, for her consistent encouragement and valuable guidance pleasantly offered to us throughout the course of the project work. We also take this opportunity to thank all the faculty of the School for their support and the wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

**CONTENTS:**

* **Abstract**
* **Introduction**
* **Problem Statement & Objective**
* **Methodology**
* **Result Analysis**
* **Conclusion and Future Scope**
* **References**

**ABSTRACT:**

Cardiovascular Intrusion Detection Systems (IDS) are an essential tool for detecting potential threats and attacks in computer networks. Predictive analytics techniques can significantly improve the accuracy of IDS by identifying potential threats and attacks using historical data and identifying patterns. This presentation discusses the various predictive analytics techniques, such as clustering, classification, and anomaly detection, that can be used in IDS. It also explores the popular algorithms used in predictive analytics, such as K-Means Clustering, Decision Trees, and Support Vector Machines. While predictive analytics can improve the effectiveness of IDS, it also presents several challenges that need to be addressed, such as data quality, scalability, diversity of data, complexity of attacks, and human expertise. Organizations need to invest in the right infrastructure, technology, and human resources to overcome these challenges and improve the effectiveness of their IDS. The presentation concludes by highlighting the importance of understanding predictive analytics techniques and algorithms to choose the best solution for organizations' needs.

**INTRODUCTION:**

In today's era, data is considered one of the most valuable assets, and it is critical to protect it from potential cyber threats and attacks. Intrusion Detection System (IDS) is a security solution that helps in detecting potential threats and attacks in real-time. IDS monitors network traffic, system activities, and user behavior, and alerts security personnel about any suspicious activity.

With the advancement in technology and the rise in cyber-attacks, predictive analysis has become an essential tool for IDS. Predictive analysis involves using statistical techniques, Machine Learning algorithms, and Artificial Intelligence to analyze historical data and predict future events.

IDS is a critical component of cybersecurity, and predictive analysis can significantly improve its effectiveness. By analyzing historical data and predicting potential threats, IDS can become more proactive and effective in detecting potential threats and attacks. With this project, we aim to provide you with an understanding of how predictive analysis can enhance IDS and help organizations better protect their data and assets.

The purpose of this paper is to provide an overview of Intrusion Detection System under the predictive analysis domain. We will discuss how predictive analysis can help in improving the accuracy of IDS, various types of IDS, predictive analytics techniques, machine learning algorithms, challenges in implementing IDS under predictive analysis domain, and the future of IDS under predictive analysis domain.

**Problem Statement & Objectives:**

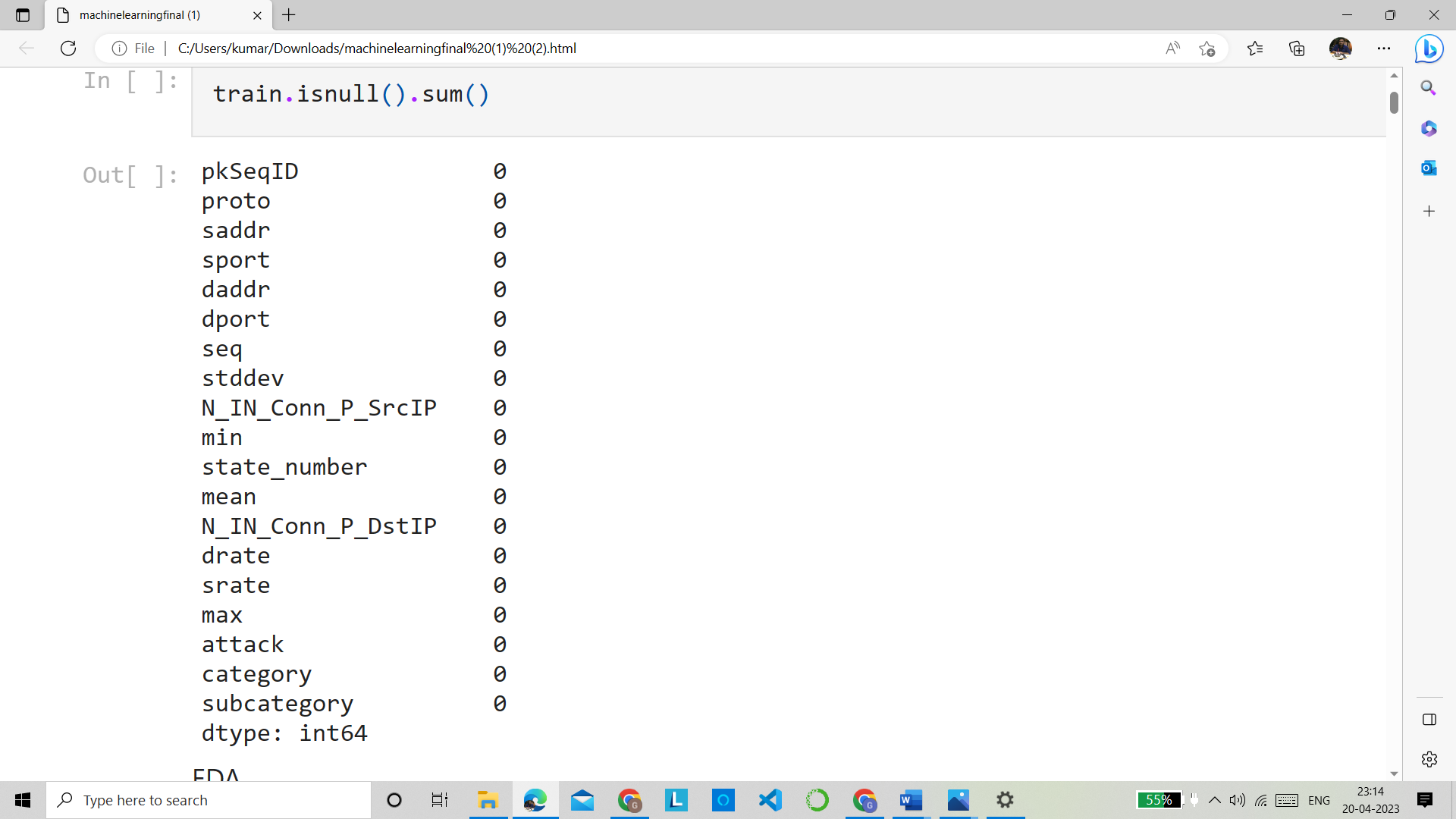
**Problem statement**- Despite significant advancements in network security, cyber attacks continue to pose serious threats to organizations and individuals alike. Traditional rule-based IDSs are limited in their ability to detect and prevent sophisticated attacks due to their reliance on predefined signatures or patterns. Moreover, the ever-evolving nature of cyber threats makes it challenging to keep rule-based IDSs updated with the latest attack techniques. ML-based IDSs, on the other hand, have the potential to overcome these limitations by leveraging data-driven approaches to learn and adapt to new attack patterns in real-time. However, there are still several challenges that need to be addressed in building effective ML-based IDSs, such as the need for large labeled datasets, the interpretability and explainability of ML models, and the potential for high false positives and false negatives. Ensemble learning, which combines multiple ML models to make more accurate and robust predictions, has emerged as a promising solution to address these challenges and improve the overall performance of IDSs. Thus, the problem statement for this research is to develop an intrusion detection system using ML and ensemble learning techniques that can effectively detect and prevent diverse cyber attacks, while addressing the challenges associated with data availability, interpretability, and accuracy of predictions.

**Objectives:**

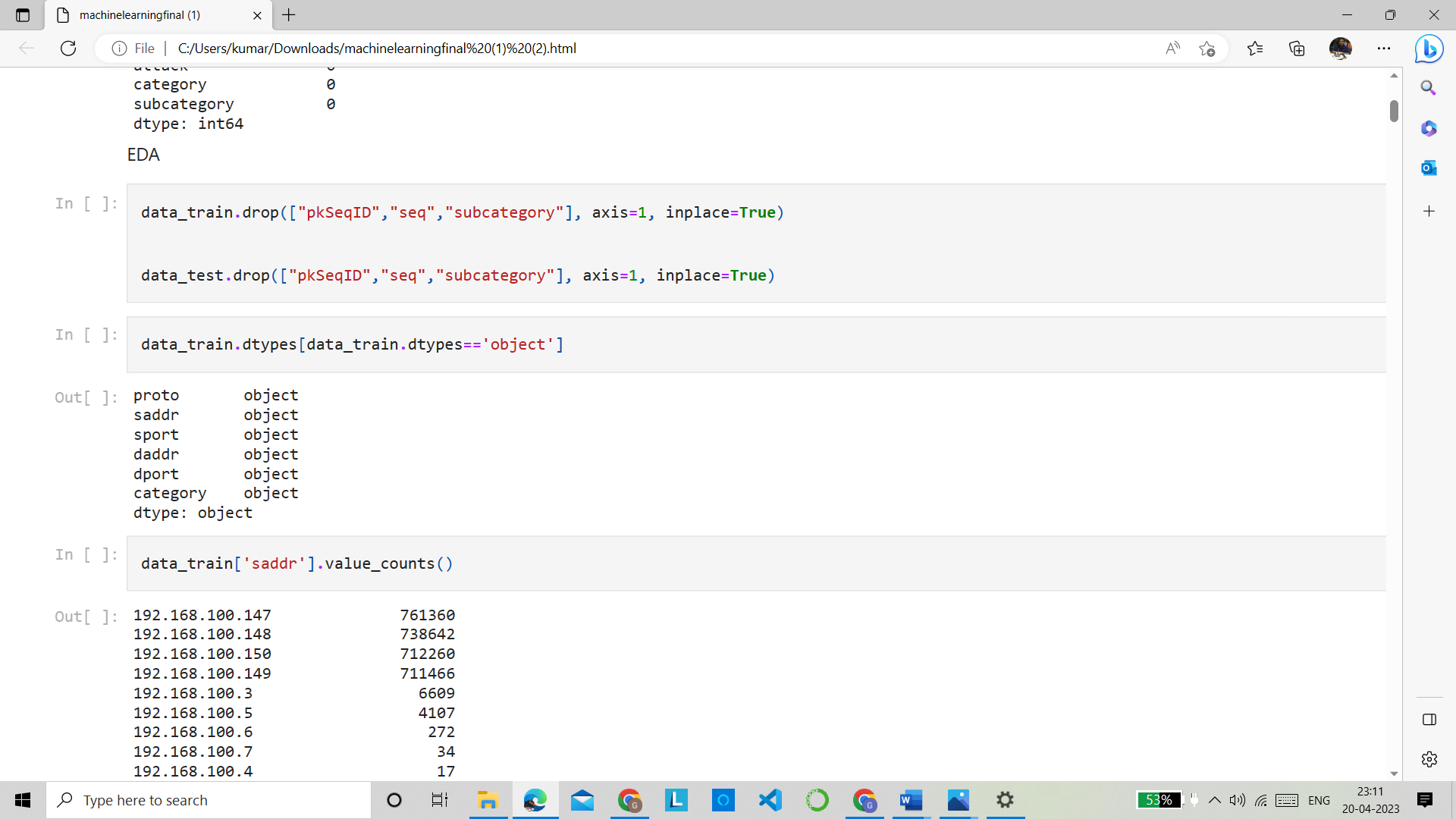
1. Develop ML-based IDS: Implement and train ML models, including logistic regression, SVM, and XGBoost, using labeled datasets to detect and classify network intrusions based on their characteristics.
2. Enhance Ensemble Learning: Utilize ensemble learning techniques, such as stacking or boosting, to combine the predictions of multiple ML models, including logistic regression, SVM, and XGBoost, to improve the overall accuracy and robustness of the IDS.
3. Explore Hybrid CNN+RNN Model: Investigate the integration of CNN and RNN architectures in a hybrid model for IDS, where CNN can capture spatial features from network traffic data and RNN can capture temporal dependencies for better intrusion detection performance.
4. Address Data Imbalance: Handle the issue of data imbalance commonly present in IDS datasets by employing techniques such as oversampling, undersampling, or using weighted loss functions to ensure that the ML models are not biased towards the majority class.
5. Evaluate Model Performance: Conduct comprehensive evaluations of the ML models, ensemble learning techniques, and hybrid CNN+RNN model using various performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve to assess their effectiveness in detecting and preventing network intrusions.
6. Enhance Interpretability: Investigate and implement techniques to enhance the interpretability and explainability of ML models, such as feature selection, feature importance analysis, and visualization of model decisions, to better understand the reasoning behind intrusion detection predictions and gain insights into potential attack patterns.
7. Compare and Analyze Results: Compare the performance of different ML models, ensemble learning techniques, and the hybrid CNN+RNN model, and analyze the strengths and weaknesses of each approach to identify the most effective approach for intrusion detection in the given scenario.
8. Contribute to the Field: Contribute to the field of network security by developing an effective IDS using ML and ensemble learning techniques, and potentially uncover novel insights or findings that can enhance the detection and prevention of network intrusions**.**

**Methodology:**

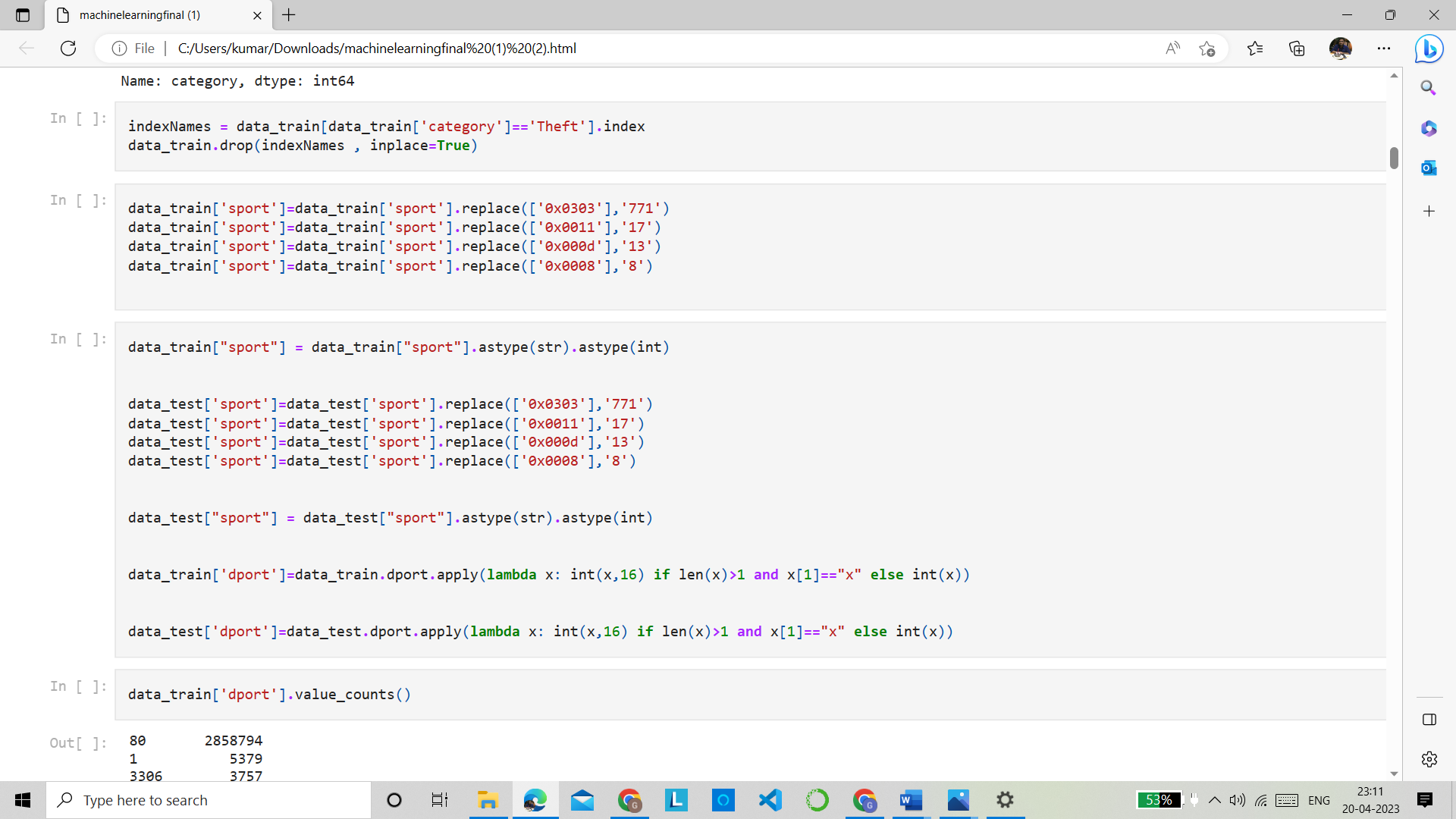
1. Data Collection: Collecting data from various sources, such as network traffic, system logs, and application logs, is the first step in the IDS methodology. The quality and quantity of data collected have a significant impact on the accuracy of predictive analytics.



1. Data Pre-processing: After collecting the data, it is pre-processed to remove noise and irrelevant information. Data pre-processing techniques such as data cleaning, normalization, and feature selection are applied to ensure that the data is of high quality and relevant to the analysis.

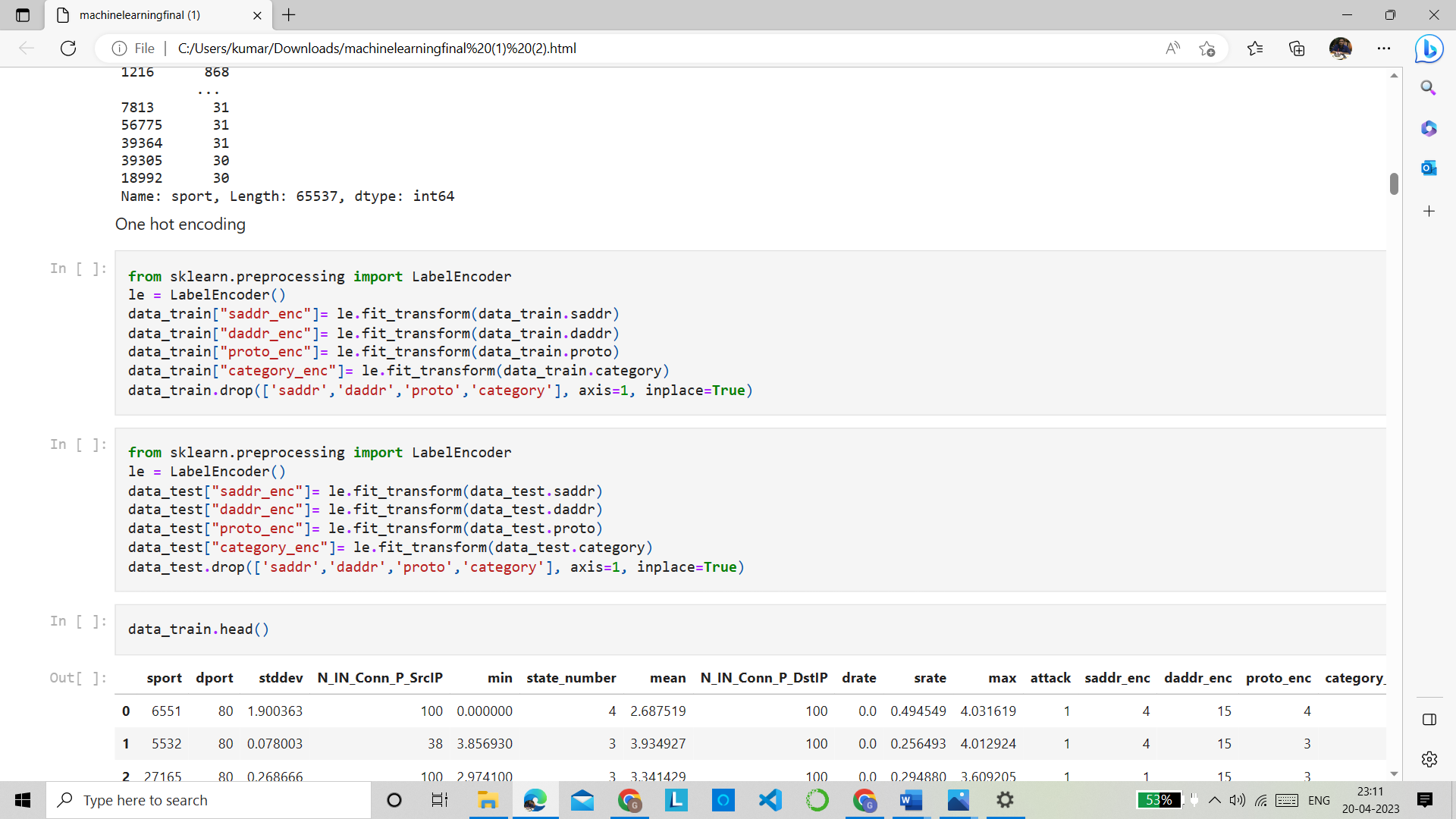


We are dropping the column which is not so useful for the prediction such as pkSeqID , seq and subcategory



In sport attribute there were some values in hexadecimal format so we converted them into the decimal format .

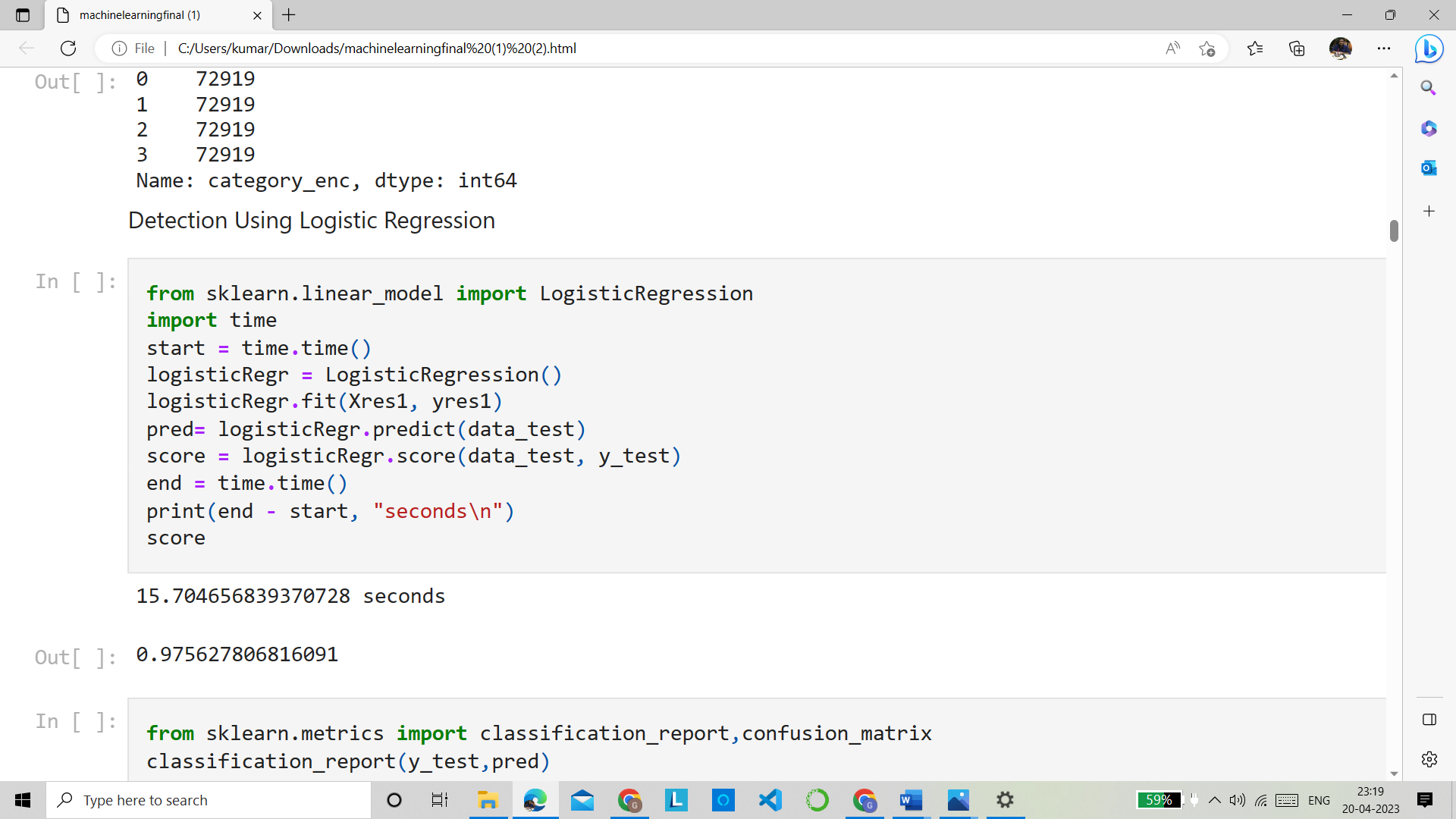
1. Feature Engineering: Feature engineering involves identifying the most relevant features or attributes that can contribute to the prediction of potential threats and attacks. This step is critical to improving the accuracy of predictive analytics in IDS.



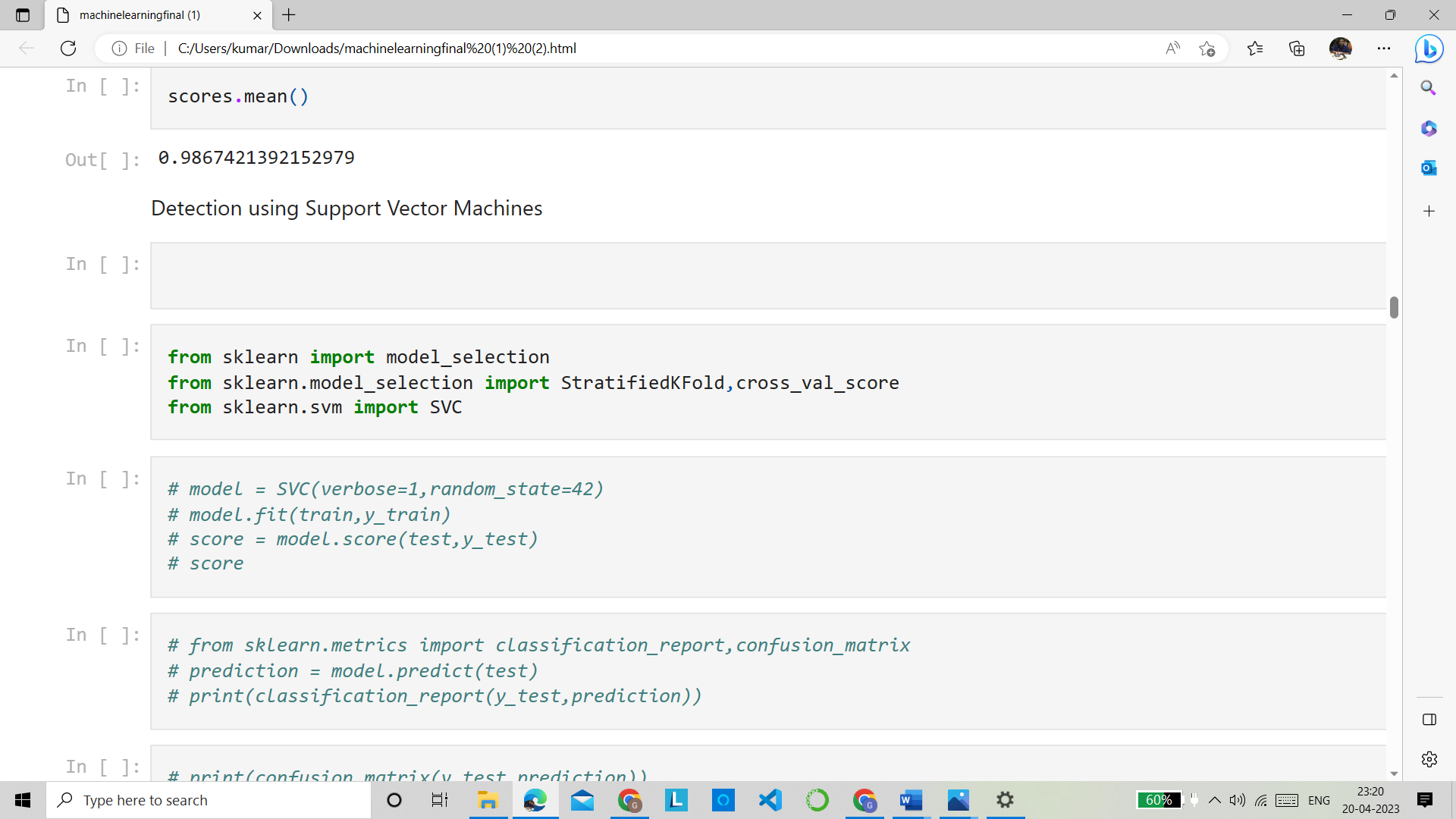
Using labelencoder we performed the one hot encoding for converting some of categorical features into numerical .

1. Model Building: Once the relevant features are identified, various predictive analytics techniques such as clustering, classification, and anomaly detection are applied to build a predictive model. The model is trained on historical data to identify patterns and predict potential threats and attacks

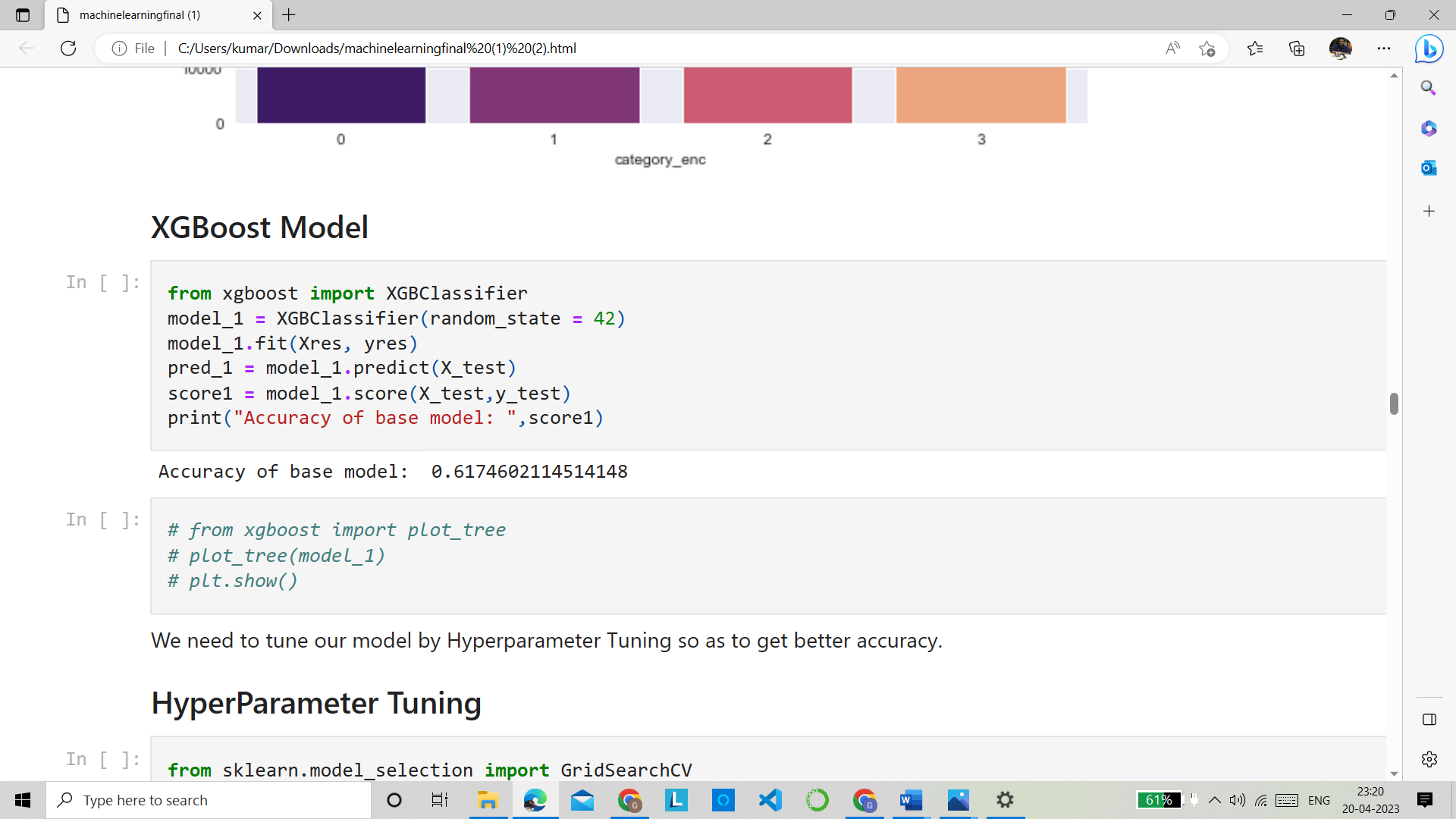
Logistic Regression



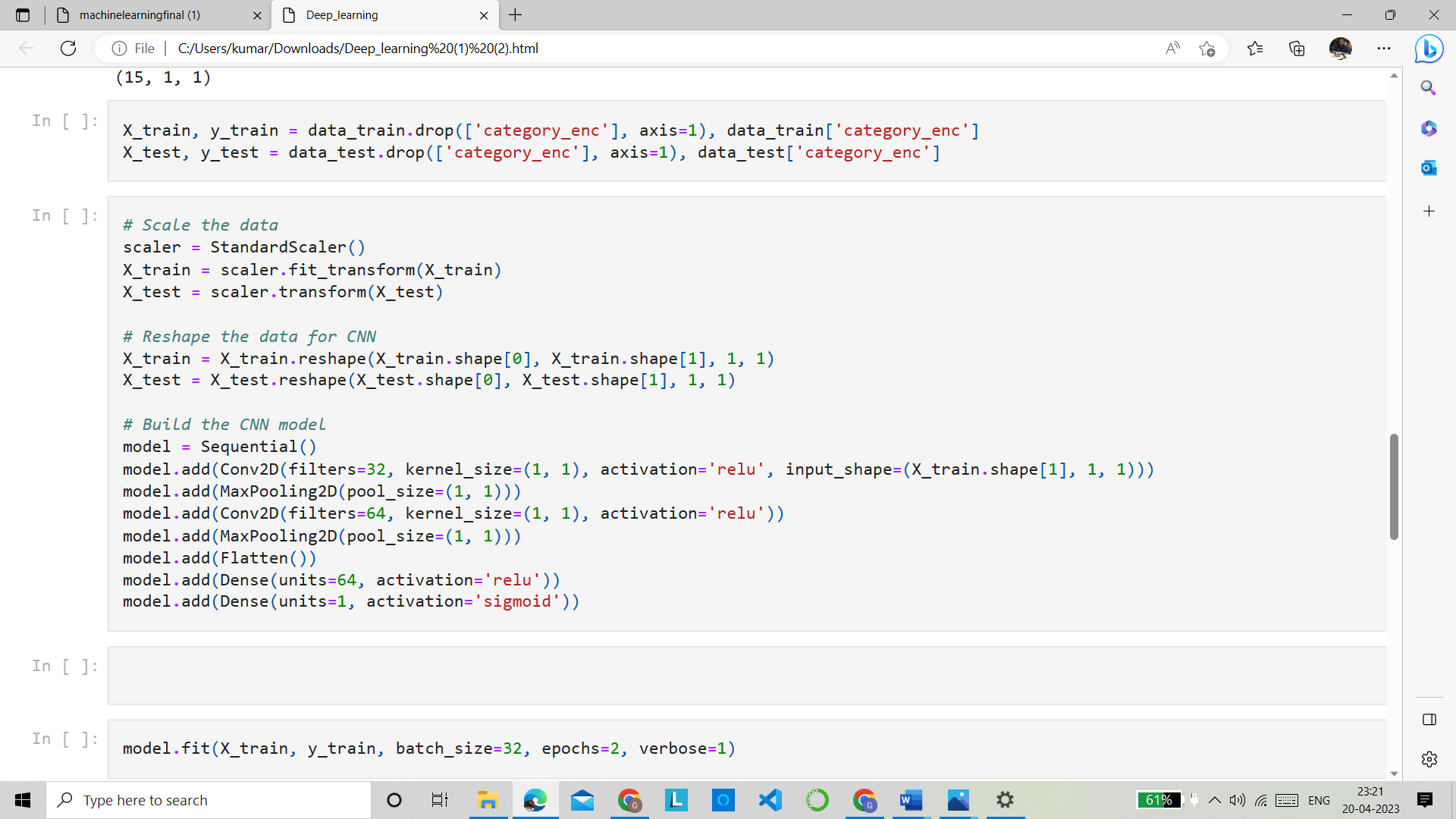
SVM



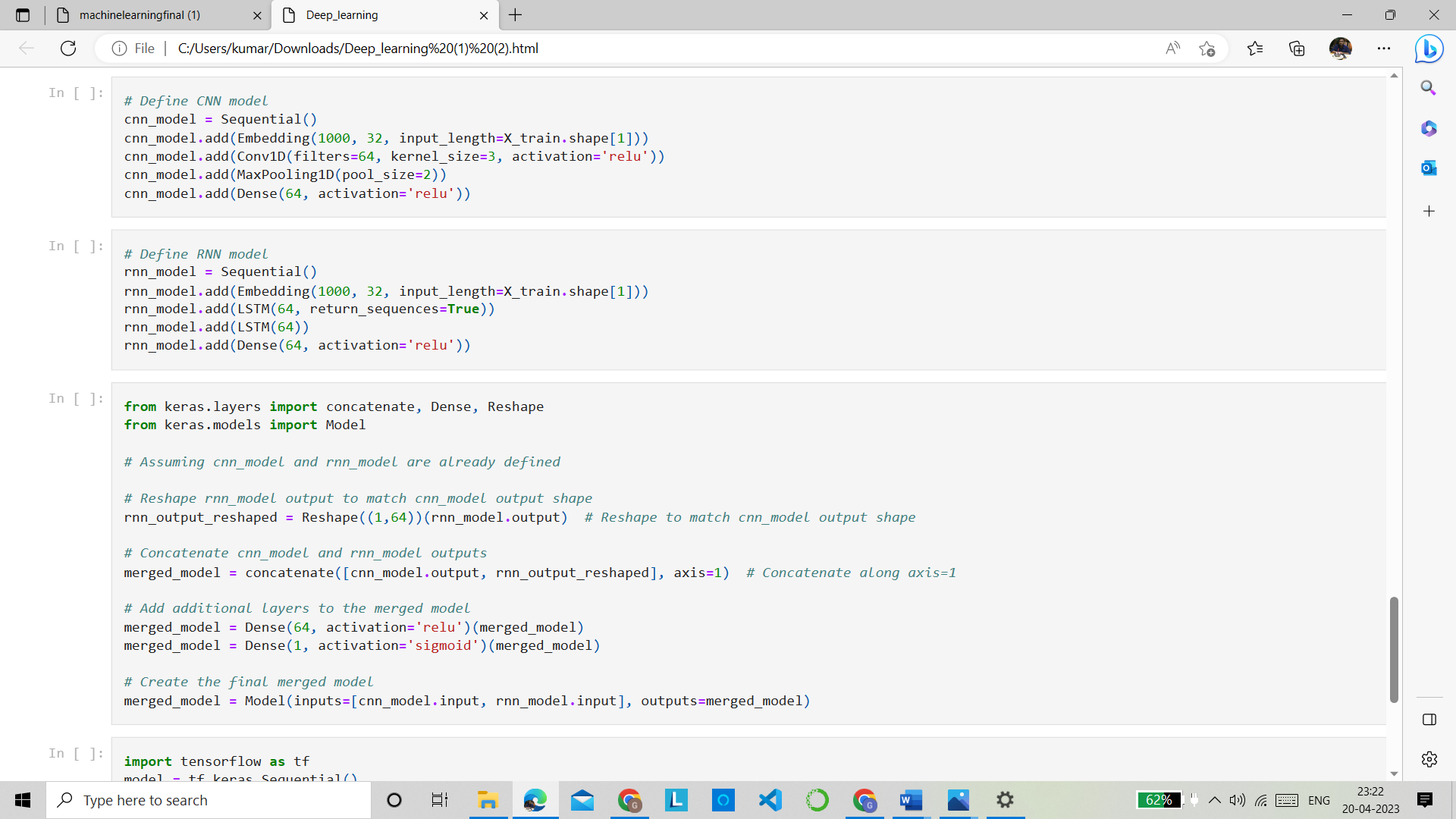
XGBoost



CNN



Hybrid Model(CNN+RNN)



1. Model Evaluation: The effectiveness of the predictive model is evaluated using various metrics such as accuracy, precision, recall, and F1 score. The model may need to be fine-tuned and optimized to improve its accuracy.
2. Deployment: After building and evaluating the predictive model, it is deployed in the production environment, where it continuously monitors the network for potential threats and attacks. Any alerts generated by the model are analyzed by human experts and appropriate action is taken to mitigate the threat.

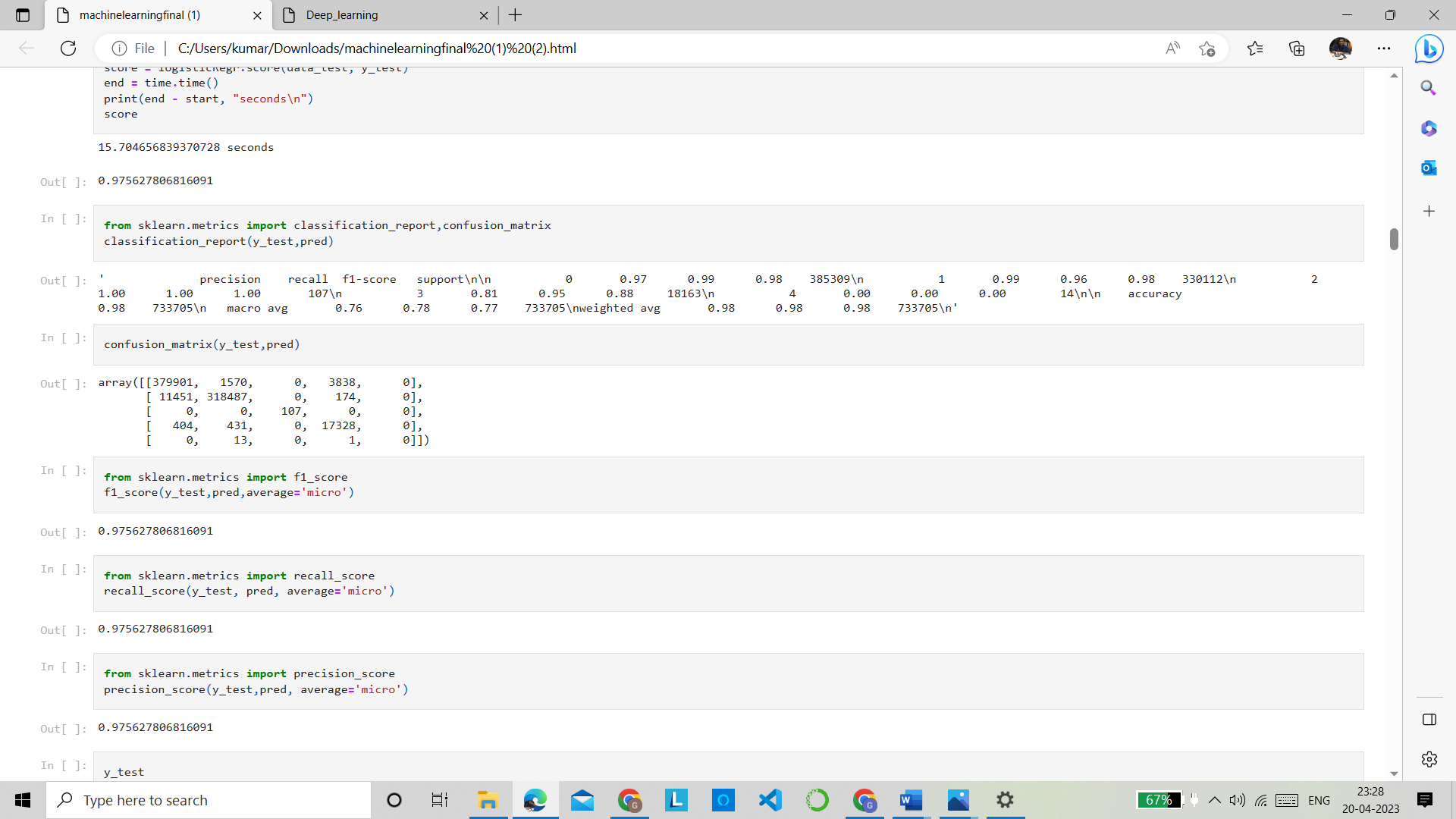
.

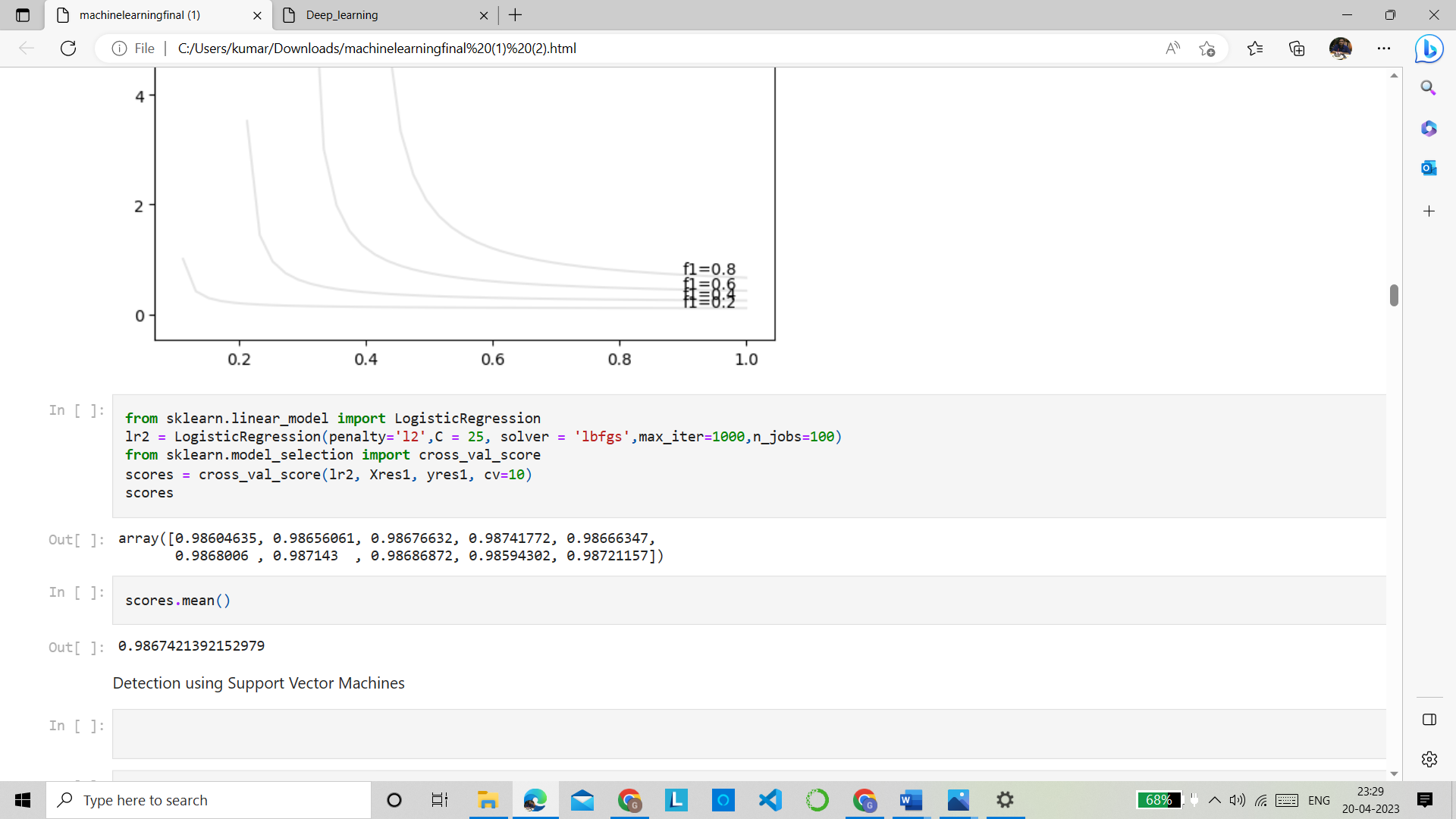
.

**Result Analysis:**

Results and inferences

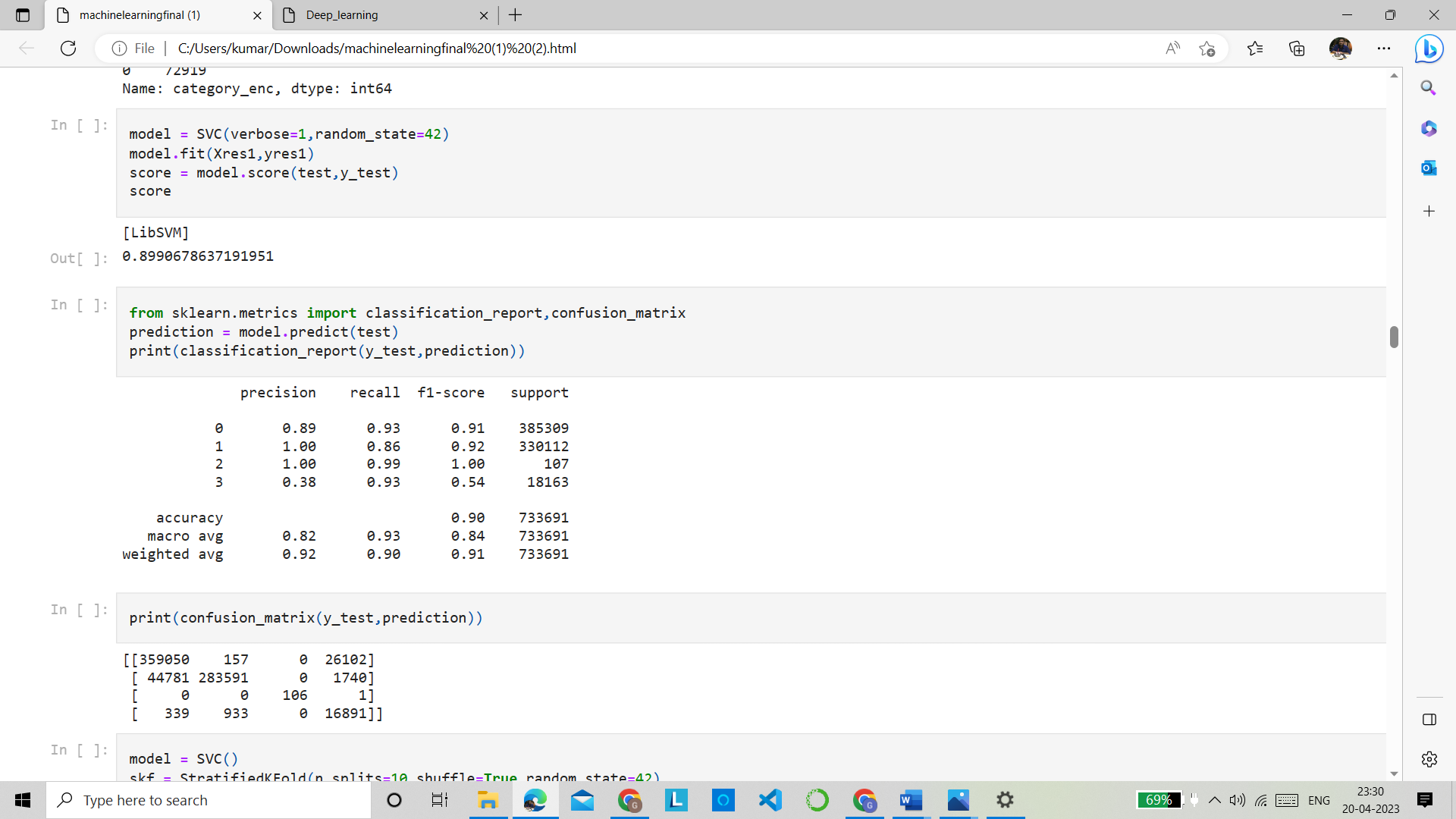
1. Logistic Regression

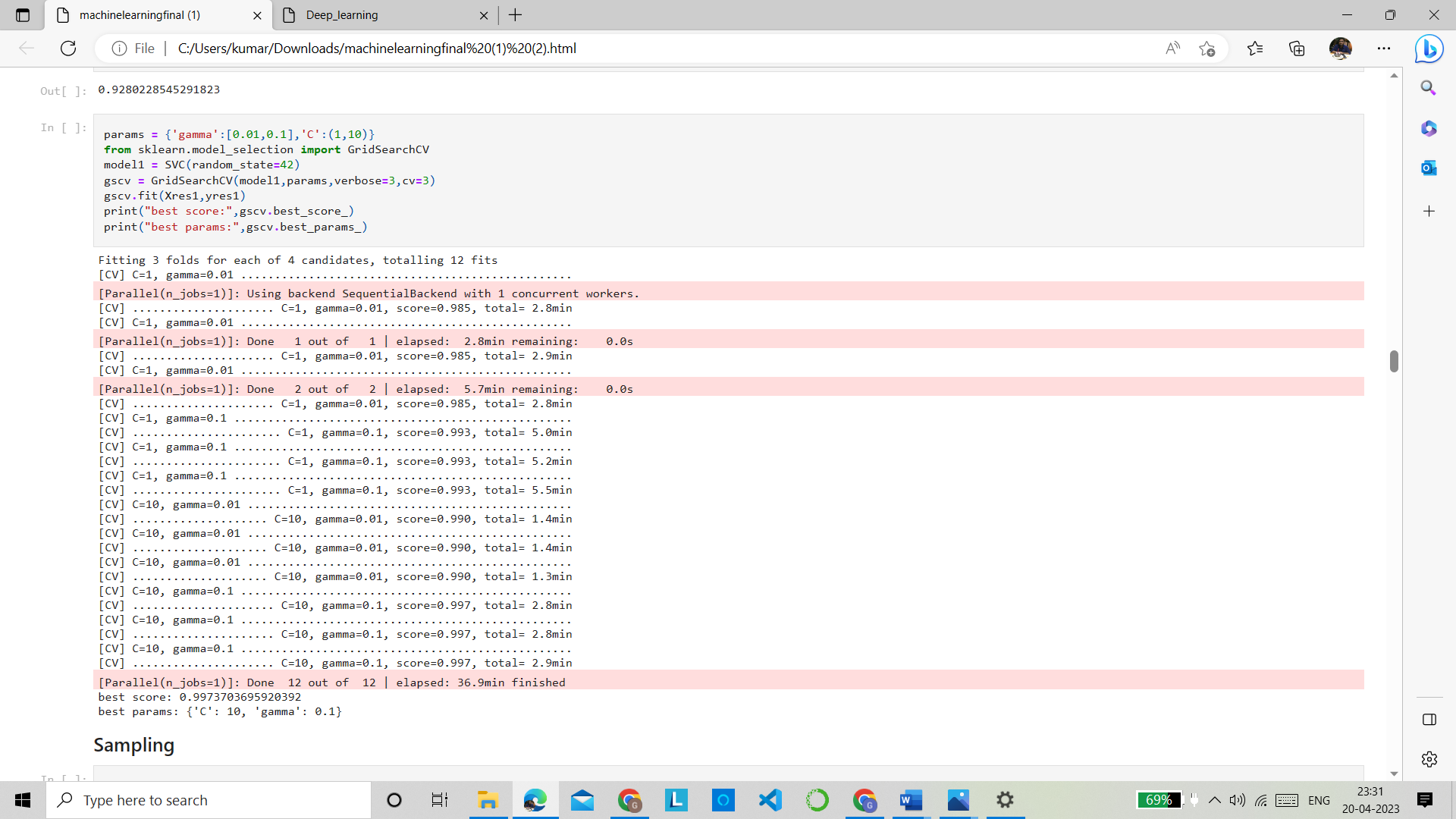




After performing Simple Logistic Regression we got an accuracy of 97.5627806816 For improving this accuracy we tried to perform K fold Cross validation(K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data) After performing K fold cross validation we got an accuracy of 98.6742139215 . So we can see that the accuracy increased after K fold cross validation.

1. SVM

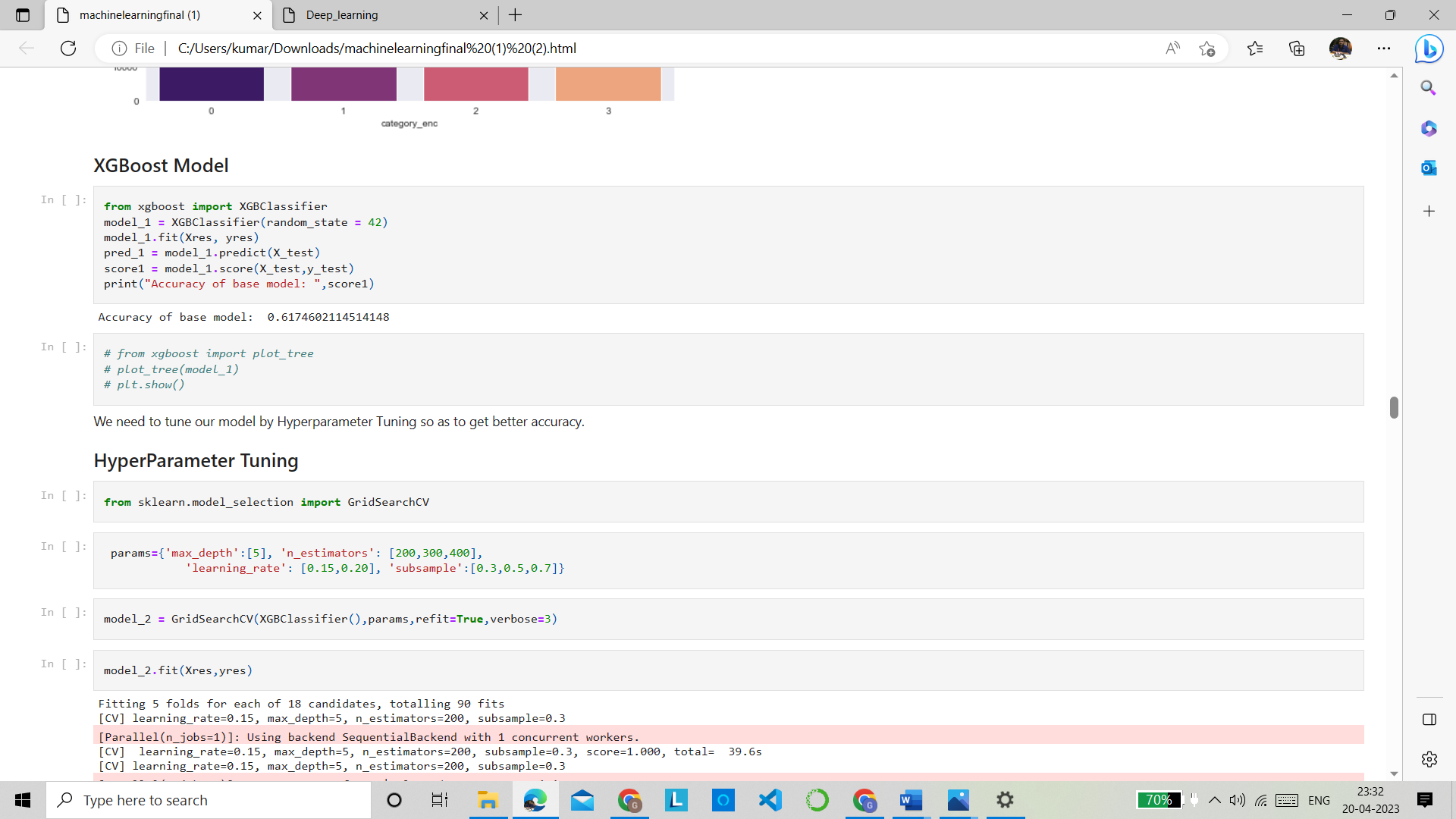


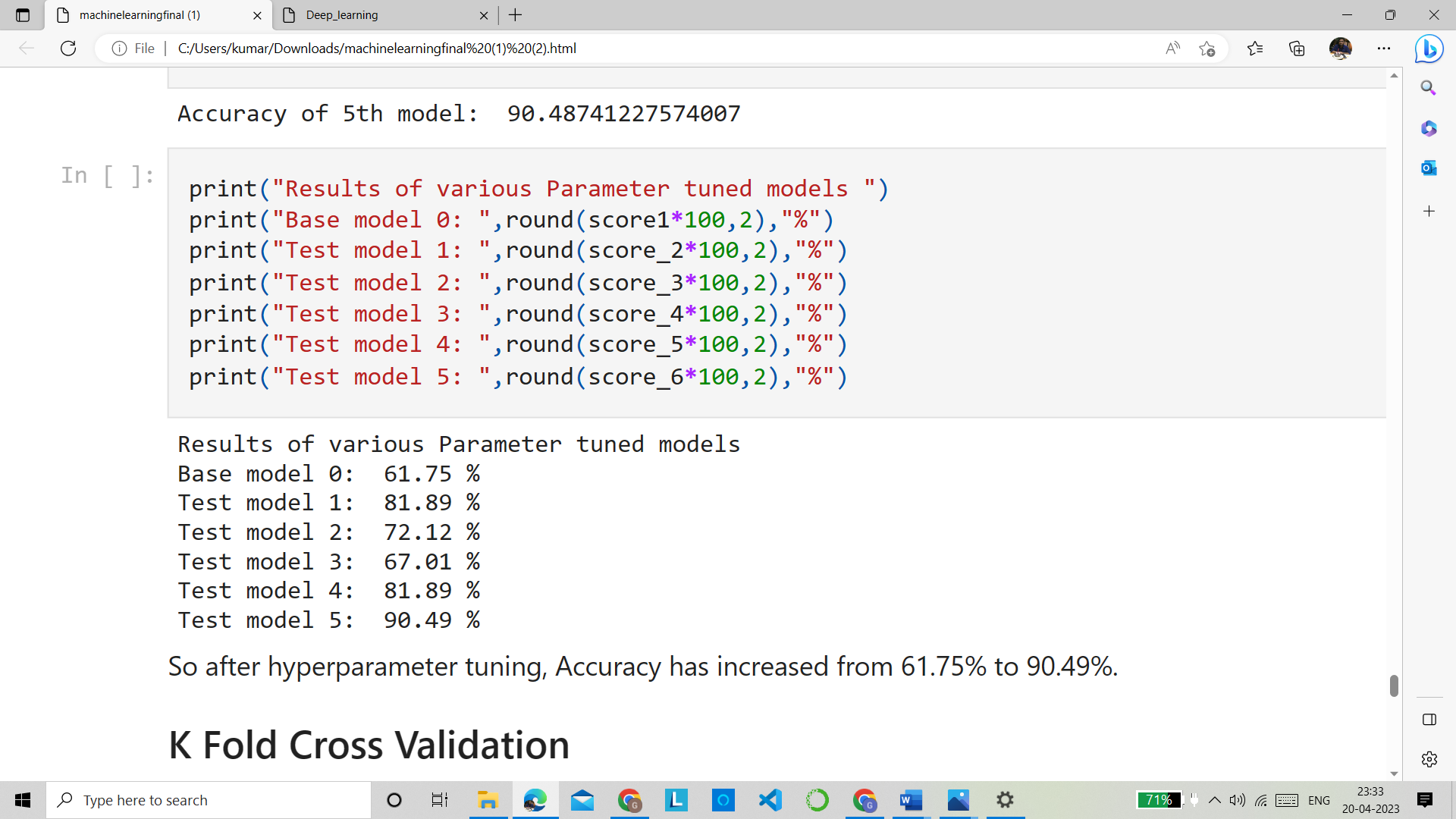


After performing the simple SVM we got an accuracy of 89.9067863719 , then we performed K fold cross validation then we achieved the accuracy of almost 92%

For increasing this accuracy we did Hyper-parameter Tuning(The grid search)Grid search is the simplest algorithm for hyperparameter tuning. Basically, we divide the domain of the hyperparameters into a discrete grid. Then, we try every combination of values of this grid, calculating some performance metrics using cross-validation. After performing the grid search we got the best accuracy among all models and it was 99.7370369592 . These were the best parameters best params: {'C': 10, 'gamma': 0.1}

1. XGboost





After performing Only XGboost we got an accuracy of around 62% . For increasing this accuracy we performed hyper -parameter tuning , after performing almost 5 test models ww got an accuracy of 90% which was the huge increase from 62%.

Results of various Parameter tuned models

Base model 0: 61.75 %

Test model 1: 81.89 %

Test model 2: 72.12 %

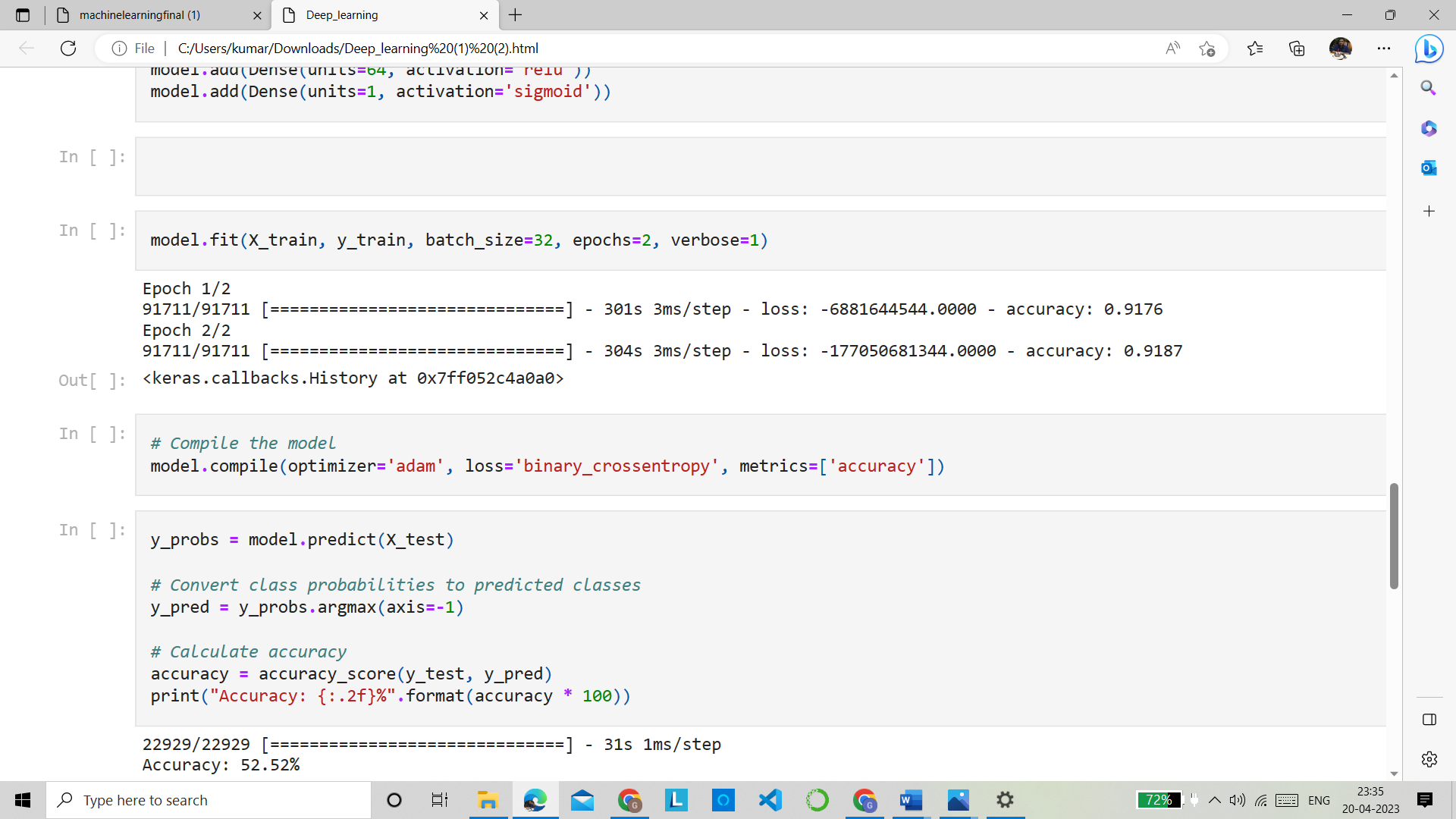
Test model 3: 67.01 %

Test model 4: 81.89 %

Test model 5: 90.49 %

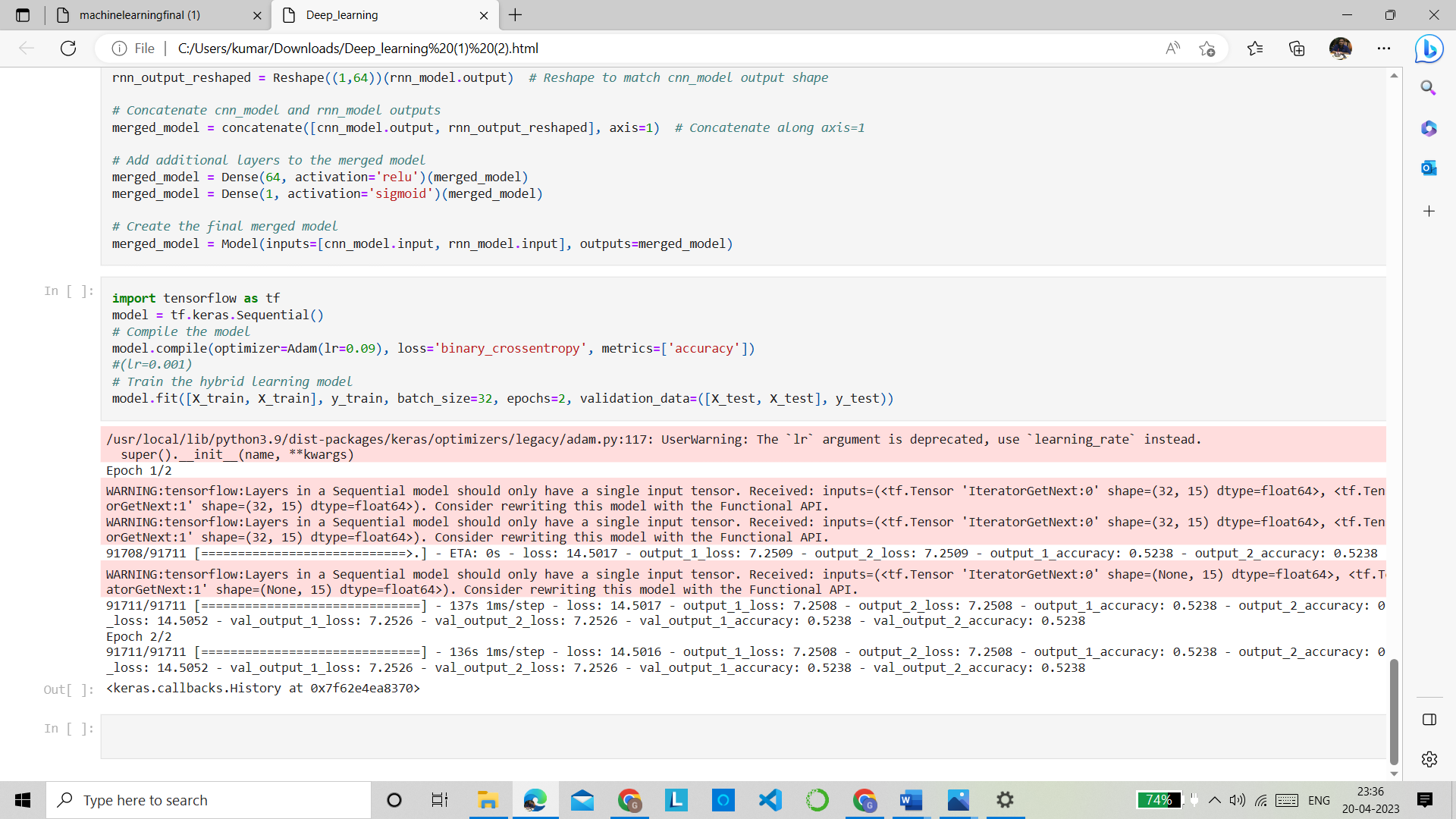
So after hyperparameter tuning, Accuracy has increased from 61.75% to 90.49%

4.CNN



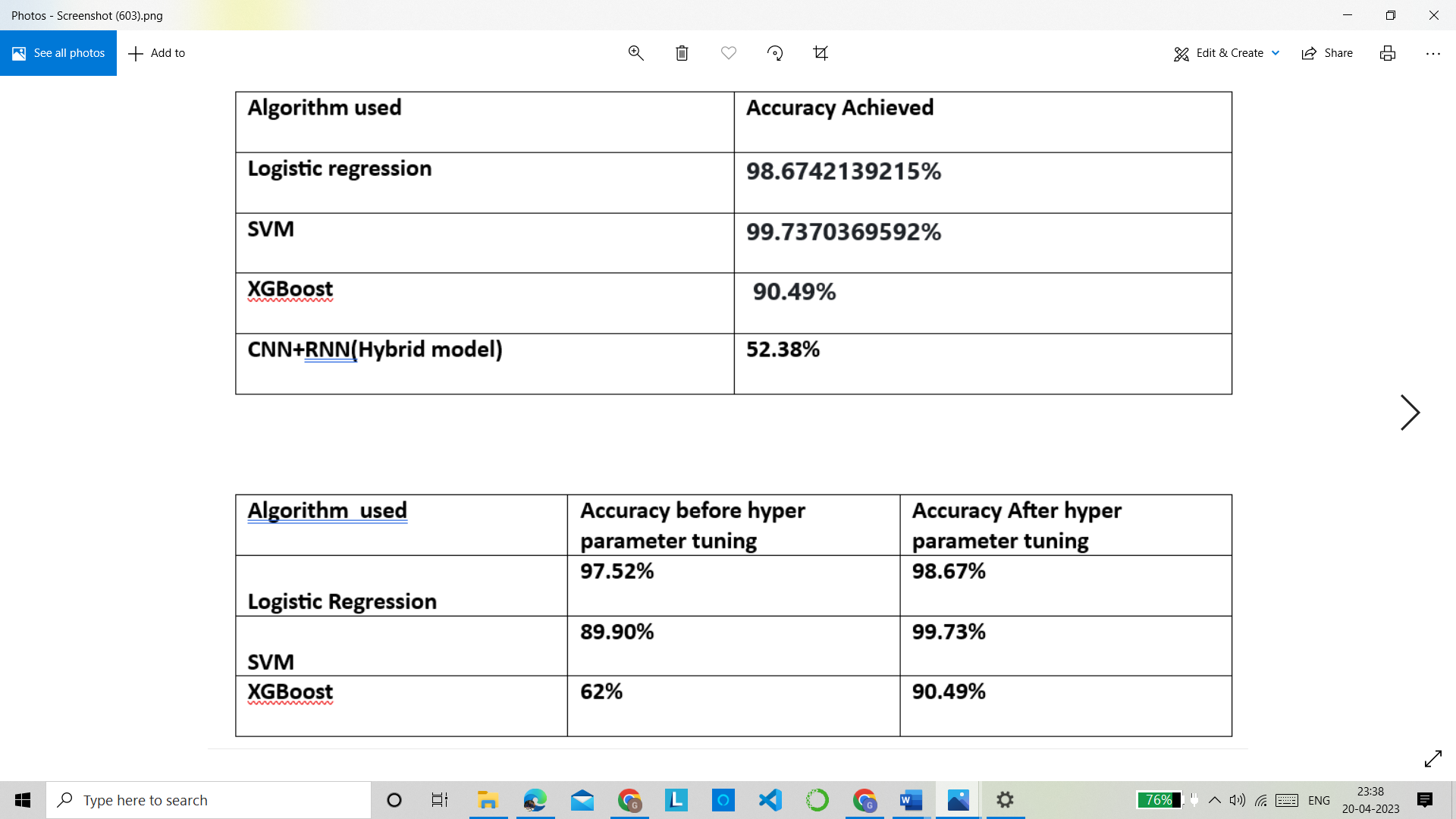
After training the Simple CNN model we got an accuracy of 52.52%

5.Hybrid Model(CNN+RNN)



After concatenating the CNN and RNN model we got an accuracy of around 52%.

Final Inferences:



**Logistic Regression :98.6742139215%**

**Support Vector Machines :99.7370369592%**

**XGBoost Classifier: 90.49%.**

**CNN:52.52%**

**Hybrid model :52.38%**

**So hereby we can conclude that SVM(Support Vector machines) performed best among all other algorithms with an accuracy of 99.737%**

**Conclusion-**

In conclusion, intrusion detection using machine learning and ensemble learning techniques such as logistic regression, SVM, XGBoost, and a hybrid model using CNN+RNN has shown promising results in detecting and mitigating cybersecurity threats. These approaches have been proven effective in classifying network traffic data and identifying potential intrusions with high accuracy and efficiency.

Logistic regression is a simple and interpretable model that can provide good results in certain scenarios, while SVM is known for its ability to handle complex data and achieve high classification accuracy. XGBoost, as an ensemble learning method, combines multiple models to improve accuracy and robustness, making it highly effective for intrusion detection tasks.

The hybrid model using CNN+RNN leverages the strengths of both convolutional neural networks (CNN) and recurrent neural networks (RNN) to capture spatial and temporal patterns in network traffic data, resulting in improved accuracy and detection rates. The CNN component is capable of extracting meaningful features from raw data, while the RNN component can capture sequential dependencies in the data, making it suitable for analyzing time-series data.

Overall, the use of machine learning and ensemble learning techniques, including logistic regression, SVM, XGBoost, and the hybrid model with CNN+RNN, has demonstrated promising results in enhancing intrusion detection capabilities. However, it is important to continuously update and refine these models with new data and adapt them to evolving cybersecurity threats to ensure effective protection against potential intrusions. Further research and development in this area are necessary to continue improving the accuracy and robustness of intrusion detection systems in the face of ever-changing cybersecurity landscape.

**Future Scope: -**

Future scope ideas for : Intrusion Detection using ML and Ensemble Learning

1. Improved Detection Accuracy: With advancements in machine learning algorithms and ensemble techniques, the future scope of intrusion detection is likely to yield higher detection accuracy compared to current models. This can lead to better identification and classification of various types of network intrusions, including known and unknown attacks.
2. Real-time Detection: Real-time detection of network intrusions is crucial to prevent potential damage. Future research can focus on developing more efficient and faster machine learning models that can provide real-time detection of intrusions, allowing organizations to respond and mitigate attacks in a timely manner.
3. Enhanced Feature Selection: Feature selection is a critical step in machine learning-based intrusion detection. Future research can focus on developing automated techniques for feature selection that can identify the most relevant and informative features from large-scale datasets, leading to more accurate and efficient intrusion detection models.
4. Explainable AI: Explainable AI is gaining importance in the field of intrusion detection to provide interpretable and understandable insights into the decision-making process of machine learning models. Future research can focus on developing intrusion detection models that are not only accurate but also explainable, allowing security analysts to understand and trust the decisions made by the models.

**References: -**

[1] ."A Machine Learning-based Approach for IoT Intrusion Detection using Bot-IoT Dataset" by R. G. Campos et al.

[2] "A Comparative Study of Machine Learning Techniques for Bot-IoT Traffic Classification" by R. Y. Mohtar et al.

[3] "Machine Learning Techniques for IoT Intrusion Detection: A Comprehensive Review and Comparative Study using Bot-IoT Dataset" by M. T. Siddiqui et al.

.[4] ."IoT Intrusion Detection using Machine Learning Techniques: A Case Study on Bot-IoT Dataset" by M. S. Hossain et al.

[5] <https://www.proquest.com/openview/f7ecfdb591084ffe6072aa316b547a2e/1?pq-origsite=gscholar&cbl=18750&diss=y>

[6] <https://www.ijraset.com/research-paper/using-machine-learning-in-detecting-iot-cyber-attacks>

[7] <https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=11851&context=theses>

[8] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9658941/>