Assignment 1  
Regression Models

short line

Balakumaran Sivanesan

24883302  
28 – April - 2024

36106 - Machine Learning Algorithms and Applications

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[**1. Business Understanding 2**](#_if9iwijz2pg3)

[a. Business Use Cases 2](#_bnjfuh5ewogw)

[**2. Data Understanding 3**](#_yg03y3w5hig9)

[**3. Data Preparation 4**](#_tj77rrmixzhi)

[**4. Modeling 5**](#_6vwh14nc44ga)

[**5. Evaluation 6**](#_mngiyx5tet6w)

[a. Results and Analysis 6](#_25v57zsj5m7g)

[b. Business Impact and Benefits 6](#_hkob76wu4d6q)

[c. Data Privacy and Ethical Concerns 6](#_uf1z6gbsejg6)

[**6. Conclusion 8**](#_9hr11g79asdx)

[**7. References 9**](#_ma3y8ytjvp1y)

# Business Understanding

**Scenario Overview:**

The primary application of this project is at a car reseller company that aims to enhance its marketing strategies by predicting which existing customers are likely to repurchase a vehicle. By accurately identifying these potential buyers, the company can target them specifically with customized marketing campaigns, promotions, and offers.

**Challenges:**

* Data Complexity: Analyzing massive datasets containing characteristics like past purchases, service history, demographic data, and more is necessary to understand trends in consumer behavior. Because of this complexity, developing precise prediction models may be difficult.
* Accuracy and Reliability: Developing a model that reliably predicts customer behavior is crucial. Inaccuracies can lead to misdirected marketing efforts, resulting in wasted resources and potential loss of customer trust.
* Integration with Existing Processes: Technical difficulties may arise when attempting to integrate new predictive models with CRM systems and marketing workflows; major process changes may also be necessary.

**Motivation:**

* Resource Optimization: There is a limited amount of money and time available for marketing. The business may more effectively deploy these resources by using predictive analytics to identify potential customers, ensuring that marketing efforts are focused on those who are most likely to make a purchase.
* Boost Sales Conversions: The probability of turning leads into sales rises dramatically when you target clients who are expected to be prepared for a new buy. This focused strategy increases each campaign's efficacy while also increasing sales.
* Customer Retention: Personalized engagement tactics are essential for keeping consumers in the cutthroat vehicle resale industry. By using predictive modeling, the business may increase customer happiness and loyalty by providing timely incentives to consumers who may be thinking about making another purchase.

Machine learning algorithms are particularly essential when it comes to anticipating client repurchase behavior for an automobile reseller organization because of their capacity to examine intricate and diverse data and reveal latent patterns that conventional statistical techniques can ignore. Here are some reasons:

* + - * Handling Large and Complex Datasets
      * Pattern Recognition
      * Predictive Accuracy
      * Scalability
      * Customization and Personalization

**Key Objectives:**

This project's main goal is to use machine learning to identify the clients who are most likely to buy cars from the auto reseller business again. The initiative is to allow focused and effective marketing methods that are more likely to result in sales by precisely identifying these clients. The ultimate goal of this predictive capacity is to raise overall profitability and improve client retention rates. Identify the stakeholders and their requirements.

**Stakeholders and Their Requirements:**

The firm has to improve its marketing strategies and customer retention programs in order to expand marketing department: Accurate projections of consumer repurchase behavior are necessary for marketing department stakeholders to optimize campaign targeting and budget allocation.

**Stakeholder Requirements with Machine Learning:**

Machine learning models can detect patterns and trends suggestive of repurchase behavior by examining past consumer data, including demographics, purchase history, and interactions. These models allow for precise forecasting of which clients are most likely to purchase a new vehicle. As a result, the business may direct marketing efforts towards these people and put customised retention plans in place, ultimately achieving stakeholder goals of increasing sales income, strengthening customer loyalty, and raising customer satisfaction.

# Data Understanding

**Overview of the Dataset:**

The dataset employed for predicting customer repurchase probabilities consists of 131,337 entries, each representing individual customer interactions with a car reseller company. Sourced from the company’s internal sales and service records, this dataset includes comprehensive data on customer demographics, vehicle details, and service usage.

**Variables and Features :**

The dataset comprises the following key variables:

* ID: A unique identifier for each customer.
* Target: A binary indicator where 1 signifies that the customer has purchased more than one vehicle, and 0 indicates only one purchase.
* Age band: Categorical bands representing customer age groups, significant for analyzing purchase trends across different age demographics.
* Gender: Includes categories Male, Female, or Missing, useful for gender-based marketing strategies.
* Car model and Car segment: Provide insights into customer preferences for specific models and vehicle types, crucial for inventory and marketing alignment.
* Age\_of\_vehicle\_years, Annualised\_mileage: Reflect vehicle usage patterns that can indicate when a customer might be in the market for a new car.

**Missing Values and Limitations:**

Significant amounts of "age\_band" and "gender" data are missing from the dataset, which may affect how accurate demographic-based analyses are. Developing trustworthy prediction models will depend on how these missing variables are handled. The 69,308 missing data in "gender" may have an impact for gender-specific marketing tactics, while the 112,375 missing entries in "age\_band" present a problem for age-related insights.

**Exploratory Data Analysis (EDA)**:

* Descriptive Statistics: Computation of means, medians, and modes to understand central tendencies of numerical features.
* Correlation Analysis: Conducted to identify relationships between continuous variables and the target, using Pearson correlation coefficients visualized through heatmaps.
* Distribution Analysis: Frequency distributions for categorical data like car models and service usage to identify popular categories and service patterns.
* Missing Value Analysis: Identification of columns with missing data and their impact on the dataset’s comprehensiveness.

short dash

# Data Preparation

* Drop irrelevant columns: The 'ID' and 'age band' columns were eliminated from the dataset as they were deemed redundant. These columns were removed since it was found they were not relevant to predicting the likelihood that a customer would make another purchase.
* Encode category variables: The 'gender' column was encoded using Label Encoder, which converts categorical data into numeric values. This was necessary since machine learning algorithms typically work with numerical input, but categorical variables can only be included in the model by encoding them.
* One-hot encoding of categorical variables: The 'car\_model' and 'car\_segment' columns were one-hot encoded using pd.get\_dummies. This was necessary since the variables were category-based and had many values, making it impossible for Label Encoder to encode them. Using one-hot encoding, categorical data are converted into several binary variables, each of which indicates a possible category value.
* Separate the data into sets for testing and training: The data was split between training and testing sets using the train\_test\_split function from sklearn.model\_selection. This was carried out in order to train the model on a subset of the data and evaluate its performance on another subset.

short dash

# Modeling

Machine Learning Algorithms Used for Modeling

In this project, various machine learning algorithms were utilized, each chosen for its strengths in handling binary classification tasks. These include:

* Logistic Regression: A statistical model that estimates probabilities using a logistic function—ideal for binary outcomes.
* Decision Tree Classifier: A tree-like model of decisions and their possible consequences, including chances of event outcomes.
* Random Forest Classifier: An ensemble of decision trees, typically used for improving classification accuracy.
* Support Vector Machine (SVM): A powerful classifier that works well on smaller cleaner datasets, using a technique called the kernel trick to transform data and then based on these transformations, it finds an optimal boundary between the possible outputs.
* K-Nearest Neighbors (KNN): A non-parametric method that classifies data points based on the 'votes' of neighbors, with the data point being assigned to the class most common among its k nearest neighbors.

**Rationale Behind Selecting These Algorithms**

* logistic regression was selected, Because of its straightforward and effective in binary classification problems and offers a reliable baseline for performance comparison.
* Decision trees are helpful for comprehending decision-making processes, which is essential for obtaining corporate insights. They are also simple to grasp and apply.
* Random Forest improves on the decision tree's tendency to overfit, offering better generalization on unseen data.
* SVM works well in high-dimensional spaces which is perfect given the variety of characteristics in our dataset—it was chosen.
* KNN doesn't require retraining and can readily adjust to real-time data, it can produce predictions that accurately represent the most recent data inputs.

**Parameter Tuning and Model Selection Process:**

Parameter Tuning: Each model underwent a tuning process using techniques like GridSearchCV or RandomizedSearchCV, which search through a specified subset of hyperparameters to find the most effective combination for our dataset.

For KNN, parameters such as n\_neighbors, weights, and metric were optimized. n\_neighbors determines the number of neighbors to consider, weights affects how much influence neighbors have, and metric defines the distance metric used.

Model Selection: Models were evaluated based on several metrics, including accuracy, precision, recall, F1-score, and the area under the ROC curve. Cross-validation was used to ensure that our evaluation was robust against overfitting.

short dash

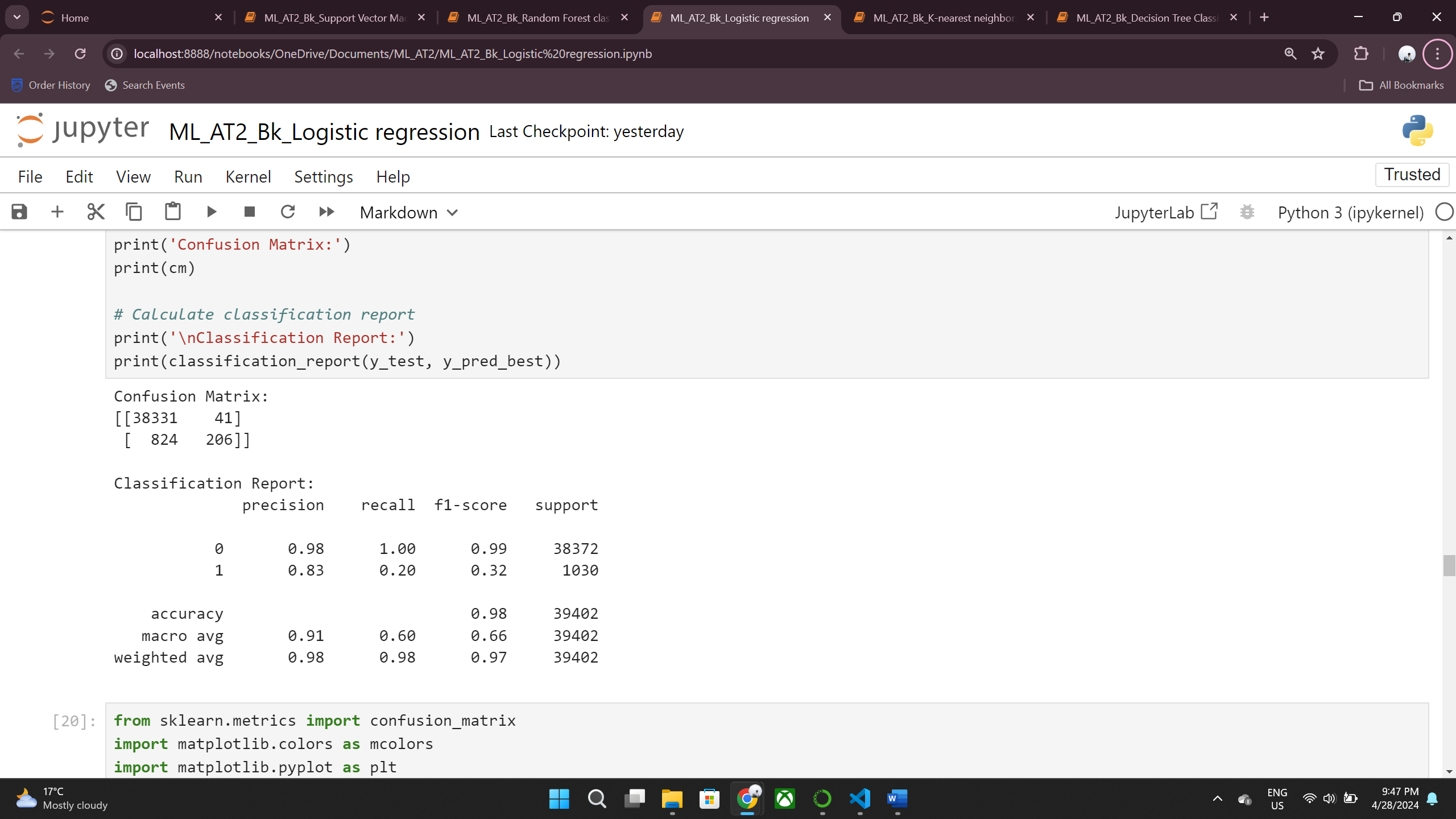
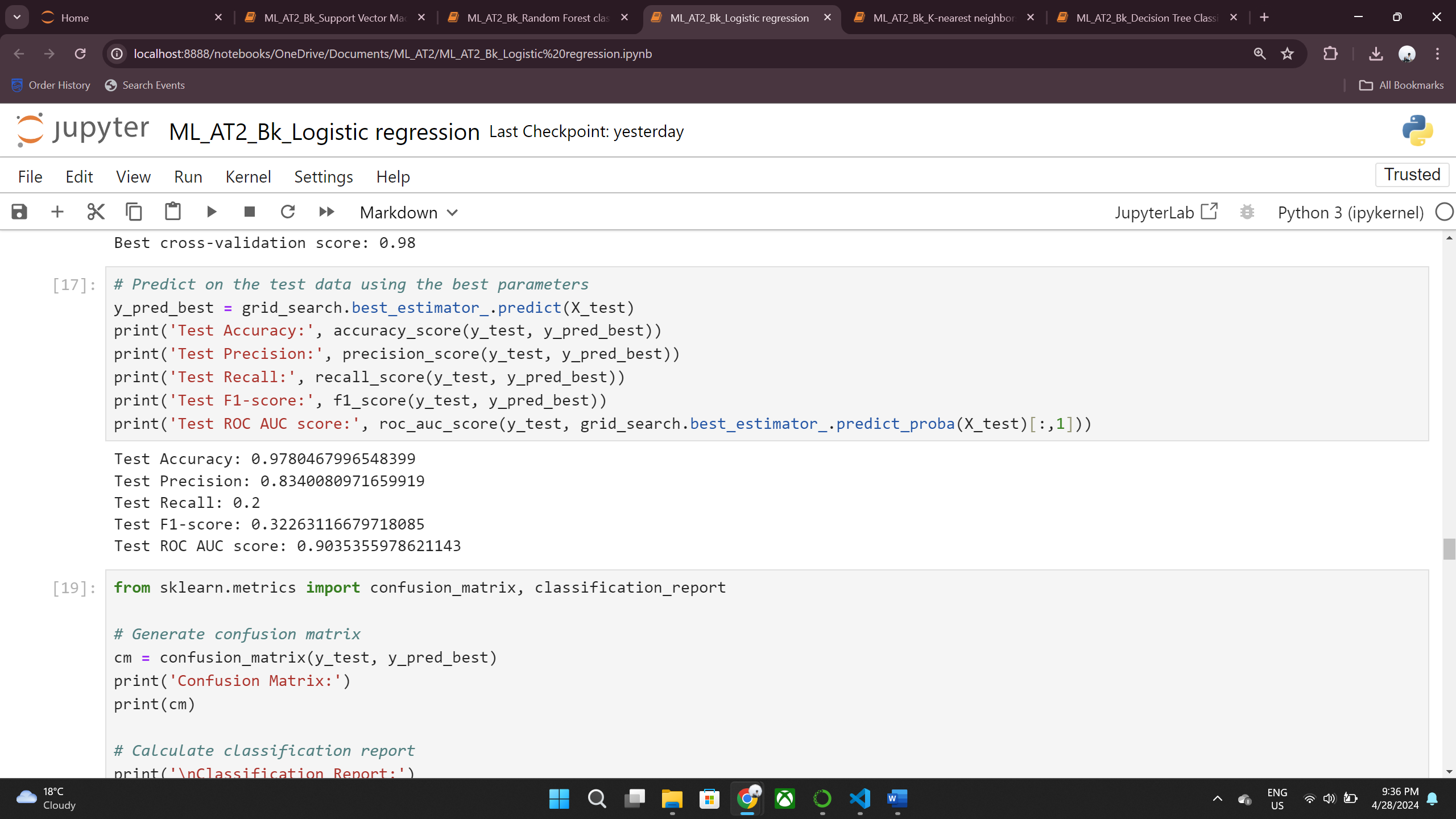
# Evaluation

## Results and Analysis

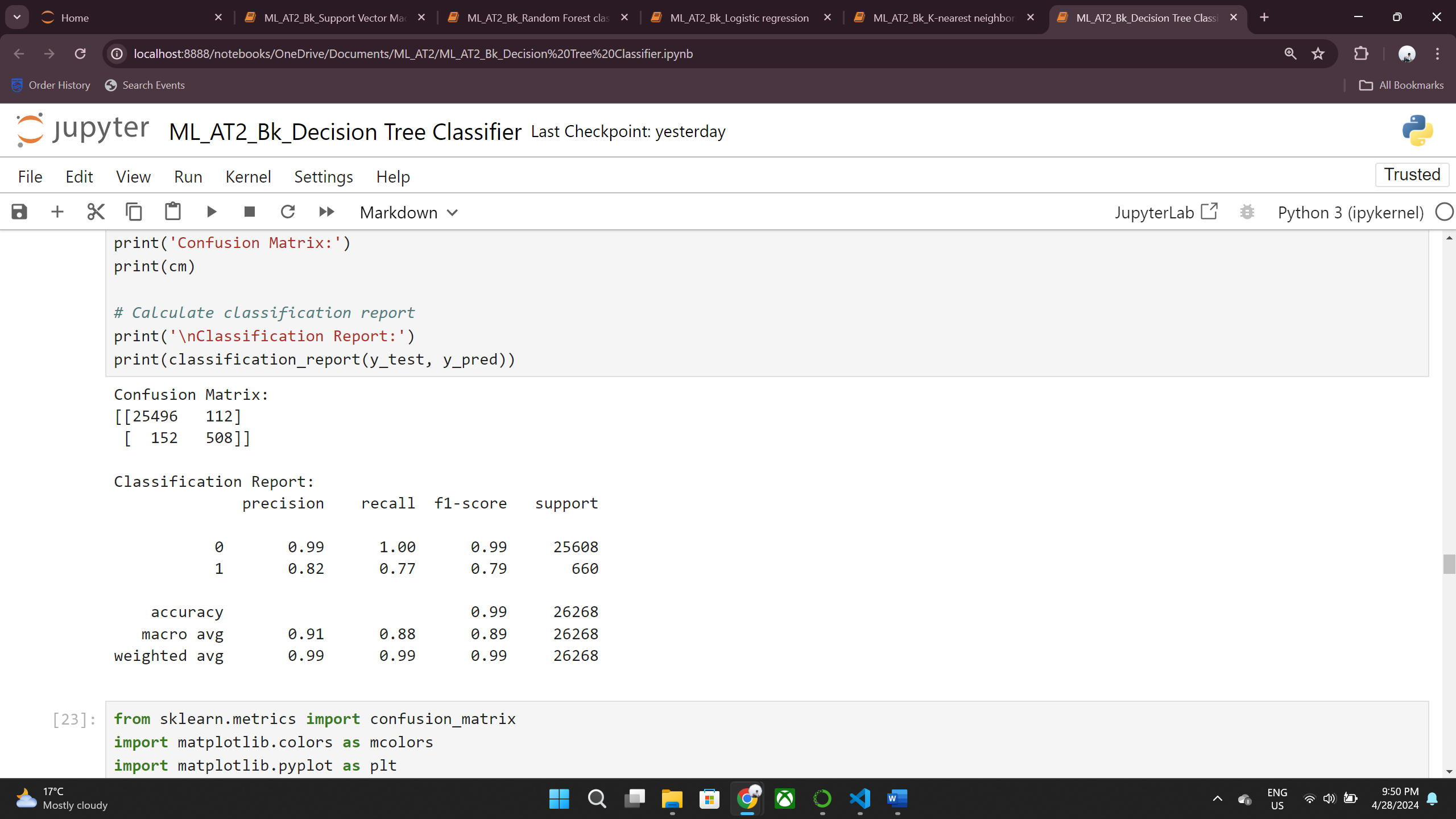
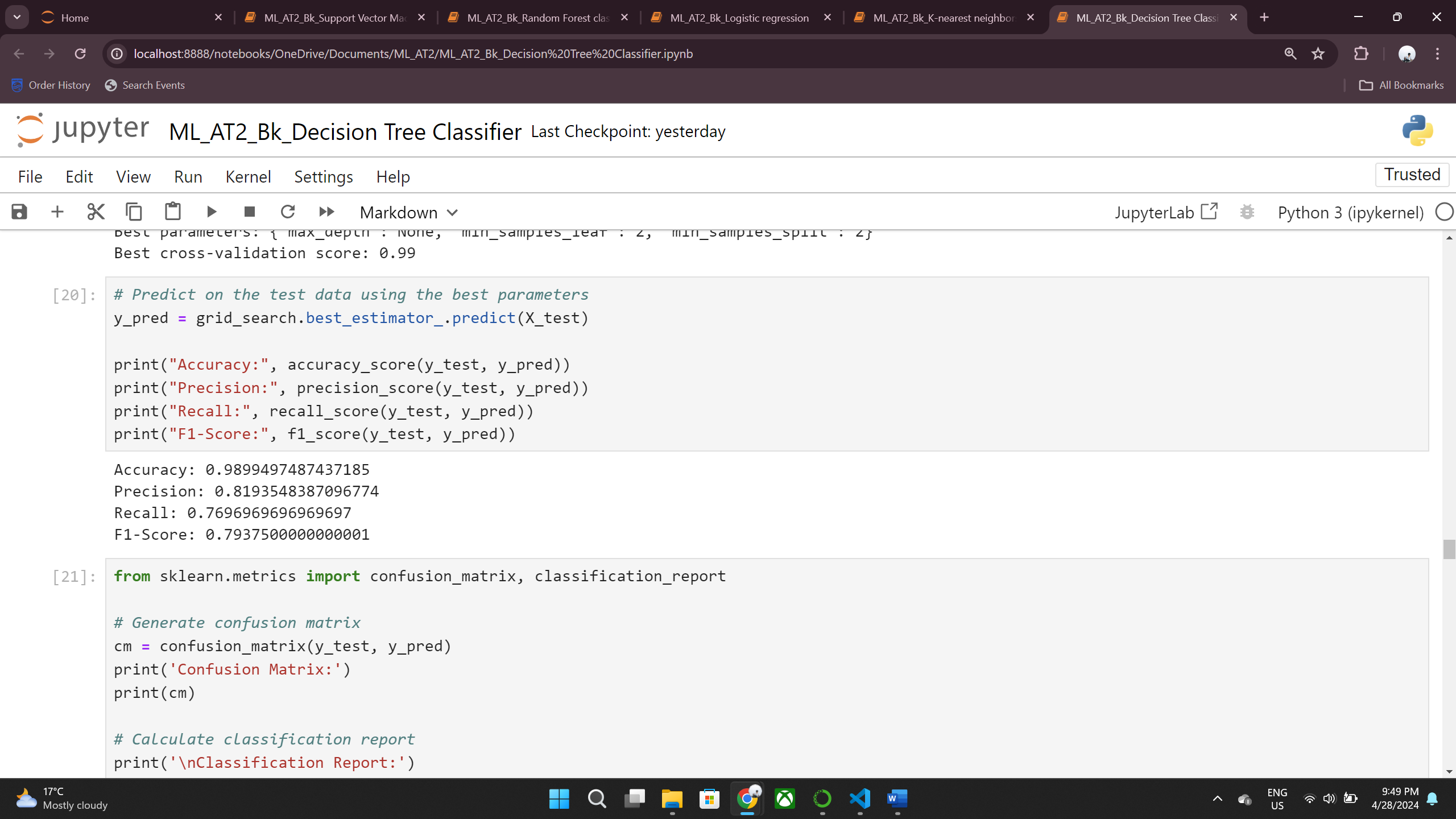
During the evaluation phase, the models were rigorously tested and measured against several performance metrics:

Model Performance Metrics:

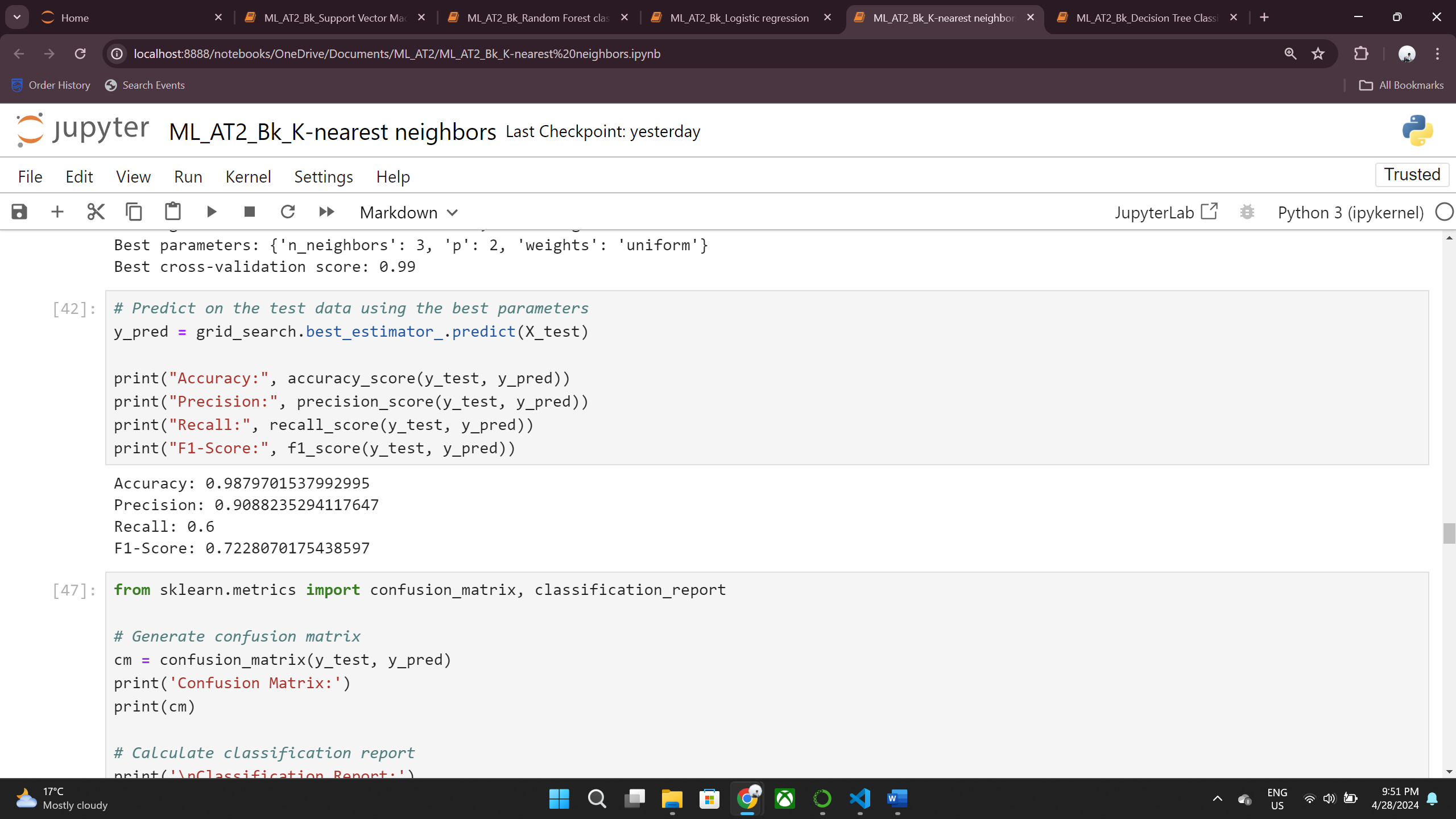
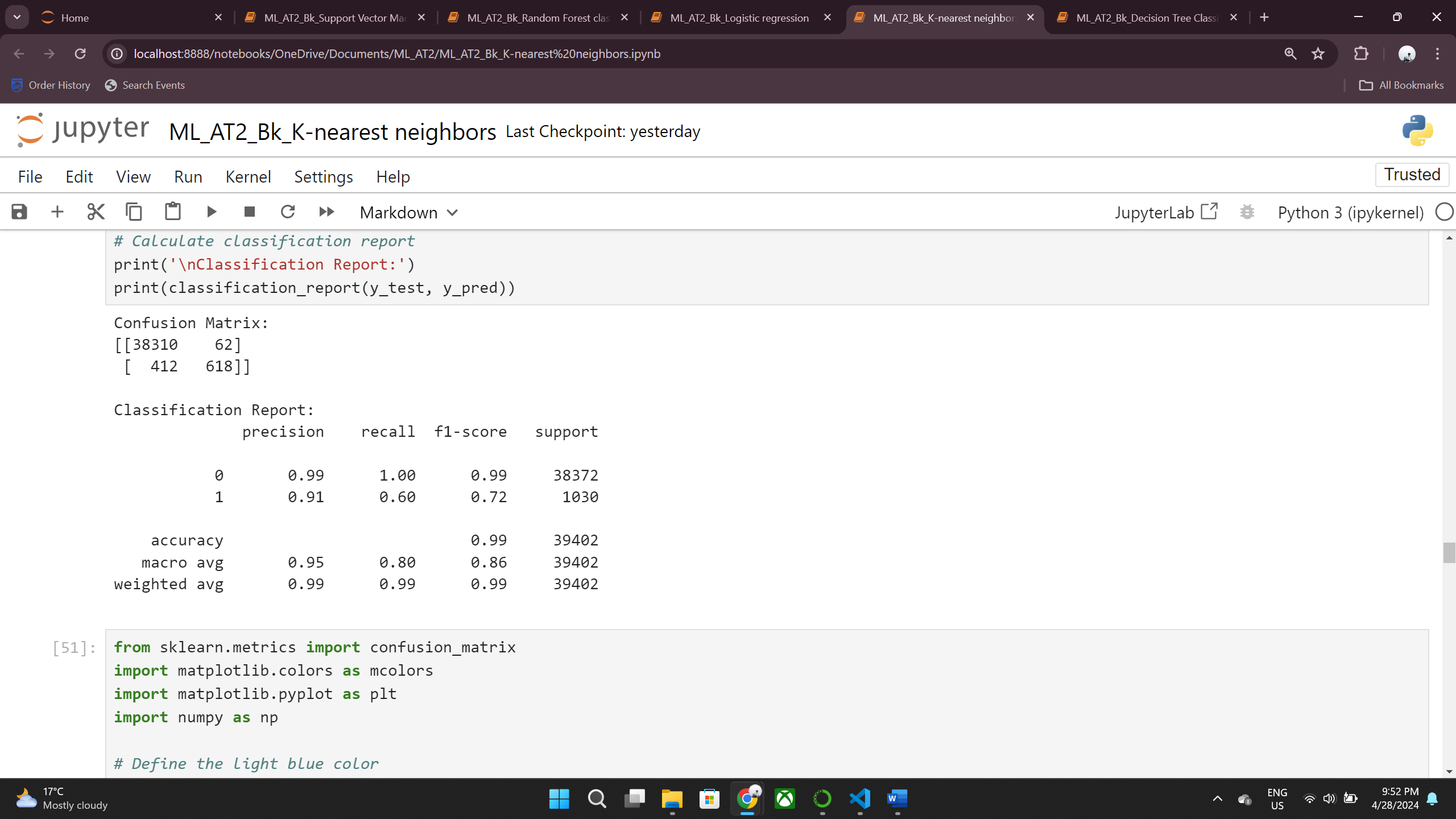
The Logistic Regression showed an accuracy of 97%, precision at 83%, recall at 20%, and an F1-score of 32%.



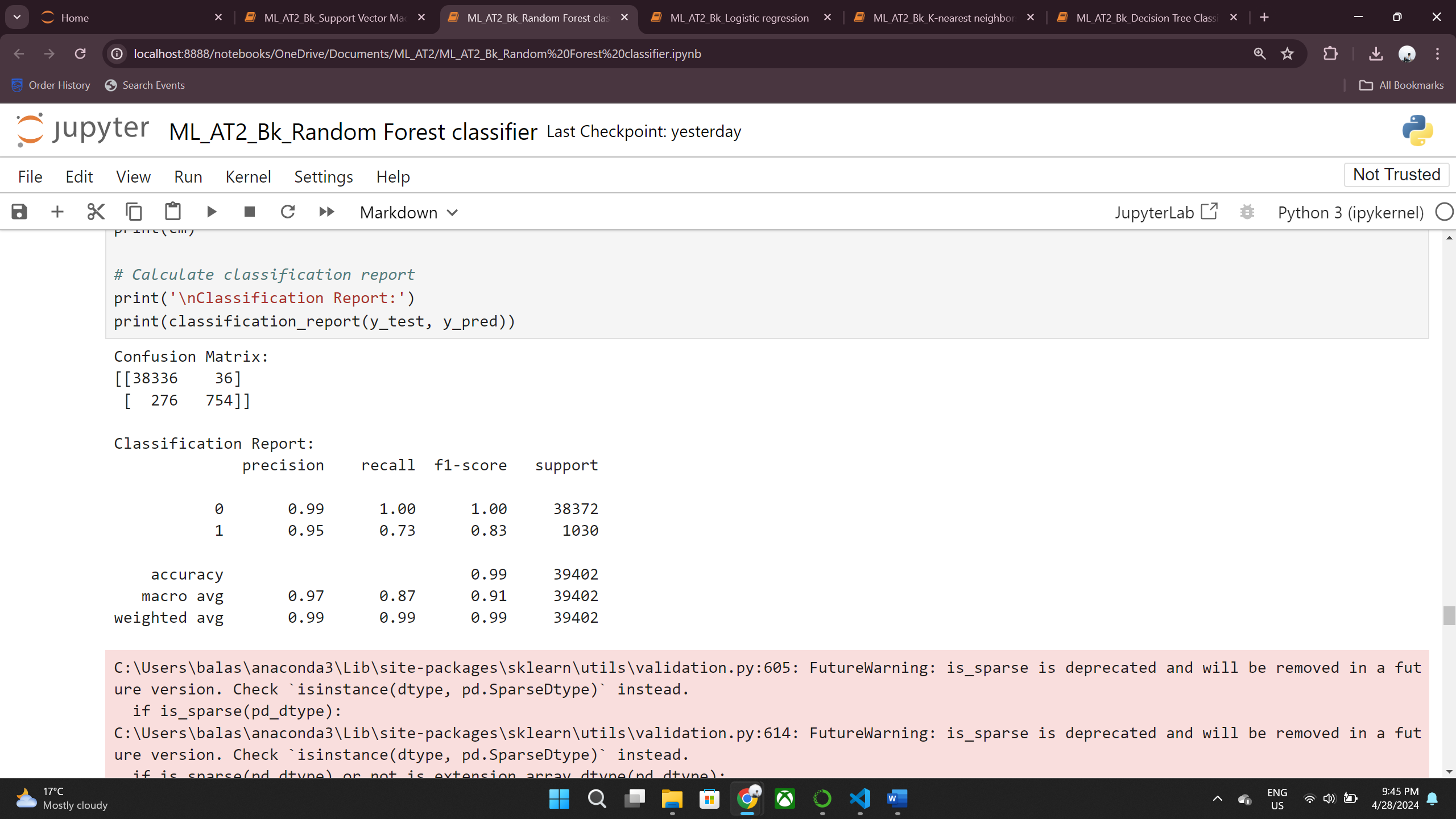
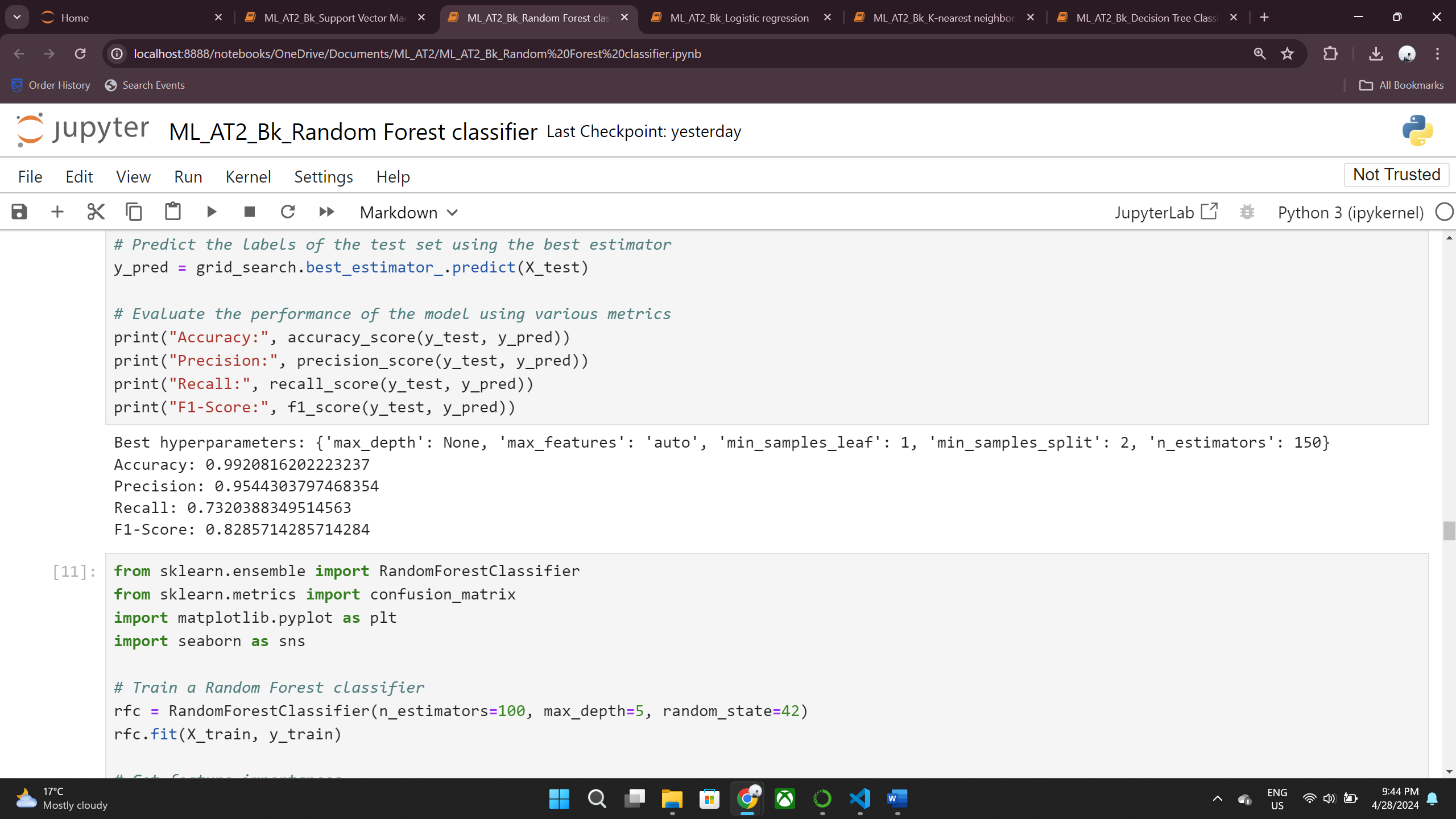
The Decision Tree Classifier (DTC) showed an accuracy of 99%, precision at 82%, recall at 77%, and an F1-score of 79%.



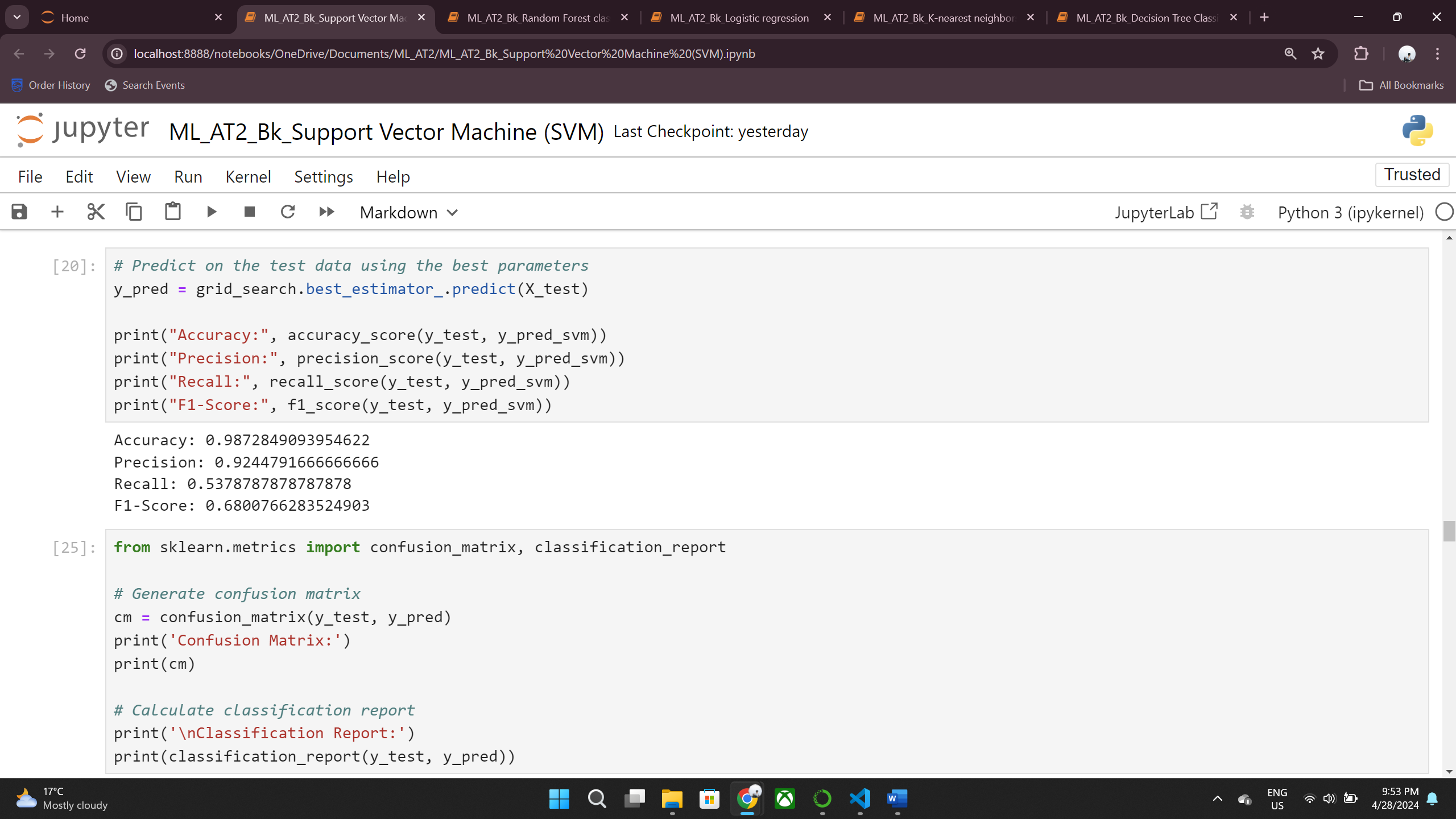
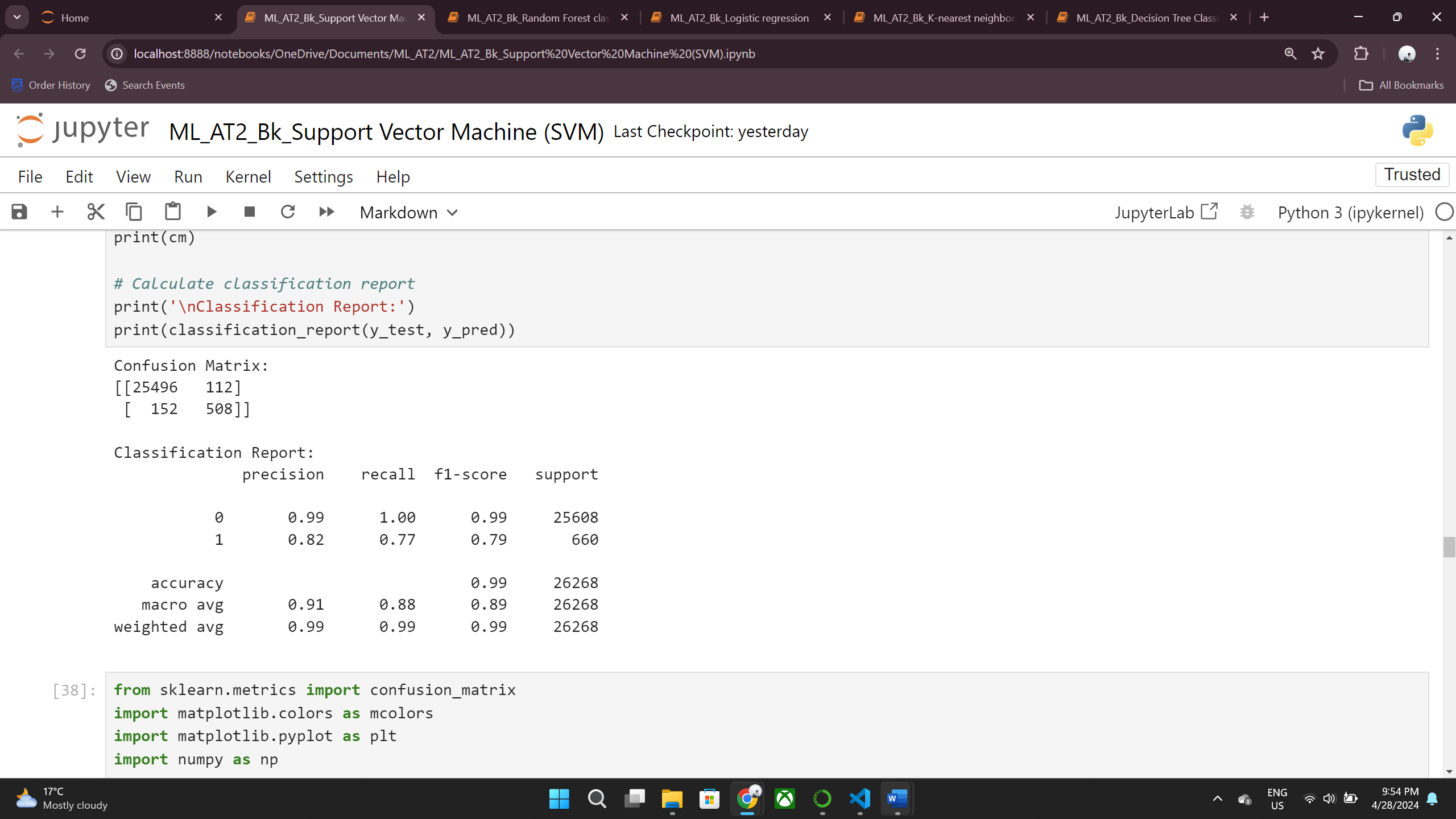
The K-Nearest Neighbors (KNN) yielded an accuracy of 98.79%, precision of 90.88%, recall of 60%, and an F1-score of 72.28%.

The Random Forest Classifier (RFC) outperformed with an accuracy of 99.2%, precision at 95.4%, recall at 73.2%, and an F1-score of 82.9%.



The Support Vector Machine(SVM) outperformed with an accuracy of 99%, precision at 93%, recall at 54%, and an F1-score of 68%.

## **Comparative Analysis:**

## The Random Forest Classifier leads with the highest accuracy and F1-score, reflecting a strong balance between precision and recall. The Decision Tree follows closely, with a slightly lower F1-score but high recall. The KNN model, while having a lower recall, excels in precision, suggesting it may be useful in contexts where false positives have a higher cost. Logistic Regression, despite its high precision, lagged significantly in recall, reducing its F1-score. SVM showed strong precision with moderate recall, resulting in a fair F1-score.

**Key Insights:**

* The ensemble approach of RFC minimizes overfitting and maximizes predictive accuracy.
* The significant discrepancy between LR’s precision and recall suggests the need for improved feature selection or model complexity.
* High precision in KNN and SVM indicates they are cautious models, likely to minimize false positives.
* The consistent accuracy across models suggests that the data is well-structured and relevant to the target variable.

## Business Impact and Benefits

Effect on Business Use Cases: Random Forest Classifier (RFC) With its industry-leading accuracy and well-balanced metrics, the has the potential to have a big influence on how well businesses can forecast consumer behavior. By putting this concept into practice, marketing resources may be optimized, potential customers can be targeted more effectively, and the expense of unqualified leads can be decreased.

Solving Identified Challenges: The approach specifically tackles the difficulty of determining which clients are most likely to become customers, allowing the company to concentrate its efforts in those areas where they have the greatest chance of producing conversions. For instance, the company may save money by decreasing false positives, which identifies clients with a low likelihood of making a transaction.

Precise forecasting of repurchase behavior is essential for businesses to maximize marketing initiatives and retention tactics, which in turn boosts client loyalty and profits. But it's crucial to understand that a predictive model is only one part of a larger strategic framework; other elements include things like customer service and product quality, both of which have a big impact on keeping customers.

## Data Privacy and Ethical Concerns

The project involves sensitive customer data, which raises significant privacy concerns, particularly regarding the storage, access, and processing of personally identifiable information (PII). Proper measures must be implemented to ensure data is handled securely to protect customer privacy, and some of the steps to ensure data privacy are Data protection, Bias monitoring, Transparency and Accountability.

short dash

# Conclusion

This project strategically used many machine learning models to anticipate consumer repurchase behavior in order to improve prediction tactics for a vehicle reseller firm. Out of all of them, the Random Forest model was distinguished by its exceptional equilibrium in performance parameters such as accuracy, precision, recall, and F1-score; this made it indispensable for fine-tuning marketing tactics and maximizing the use of available resources. The initiative made clear how crucial it is to handle data carefully, think ethically, and comply with privacy laws in order to build confidence and follow legal requirements. Going future, the project may be enhanced by utilizing ensemble approaches for robustness, expanding feature engineering, and consistently adjusting to market fluctuations. Future initiatives will emphasize the proper application of predictive analytics while also focusing on improving data privacy and minimizing biases. In summary, this research achieved its initial objectives and established a solid basis for further advancements in predictive modeling in the car resale industry.

short dashshort dash