[Introduction 1](#_Toc1581839193)

[Research Questions 2](#_Toc888322321)

[Dataset Overview 14](#_Toc392306834)

[Source of the Dataset 14](#_Toc1488177280)

[Data Overview 15](#_Toc539841069)

[Ethical Considerations 15](#_Toc1716502239)

[Methodology 15](#_Toc1261480459)

[Data Pre-processing 15](#_Toc1682522449)

[Model Selection 15](#_Toc1459401094)

[MACHINE LEARNING MODELS 16](#_Toc1653800027)

[1. Convert Categorical Features to Numerical: 22](#_Toc1428998645)

[Analysis 22](#_Toc1514088988)

[Discussion & Conclusion 28](#_Toc636652617)

[Findings and Implications: 28](#_Toc1115363925)

[Recommendations 30](#_Toc878343672)

[Impact of Work 31](#_Toc1579683070)

[Conclusion 31](#_Toc1052550668)

**Project Report: on Credit Card Fraud Analysis**

## 

## Introduction

Within the realm of financial services, credit card fraud poses an enduring and formidable challenge, casting a long shadow over both financial institutions and cardholders alike. The expeditious and precise identification of fraudulent transactions assumes paramount significance, serving as an impregnable barrier against potential financial losses and as the cornerstone sustaining trust within the intricate tapestry of the financial system. This comprehensive academic report embarks upon an exhaustive exploration of the pivotal role that machine learning models assume within the domain of credit card fraud detection.

Credit card fraud transcends mere inconvenience; it embodies a systematic assault upon the integrity and stability of financial systems across the globe. Fraudulent activities encompass a diverse array of illicit stratagems, ranging from the insidious utilization of purloined card information to unauthorised transactions and clandestine identity theft. Perpetrated with unwavering determination, these tactics exploit vulnerabilities inherent in the financial infrastructure, often leaving financial institutions and unsuspecting individuals grappling with dire repercussions.

As the scale and sophistication of credit card fraud continue their inexorable ascent, the financial industry confronts an imperative of utmost urgency – the fortification of its defensive mechanisms. Within this crucible of challenge, machine learning, a subfield of artificial intelligence (AI), emerges as a beacon of optimism and an indispensable instrument in the ceaseless struggle against fraudulent activities. At the heart of machine learning lies its profound capacity to meticulously scrutinise prodigious datasets, discern intricate patterns, and identify anomalies that may portend instances of fraudulent activity.

In the ensuing pages of this report, we embark upon a multidimensional odyssey to probe the contribution of machine learning within the broader discourse on credit card fraud detection. Our journey entails the comprehensive elucidation of the underlying mechanisms that underpin these sophisticated technologies, the methodologies they employ, and their intrinsic value, not solely in the identification of fraudulent transactions but also in the overarching mission of safeguarding the sanctity of financial transactions and sustaining trust among cardholders and financial institutions alike.

As we navigate this intricate terrain, we shall cast light upon the specific methodologies, tools, and algorithms harnessed by machine learning models for the purpose of distinguishing between fraudulent and legitimate transactions. Furthermore, we shall untangle the intricacies of real-time monitoring, adaptive learning, and the discernment of emerging fraud patterns – attributes that hallmark the efficacy of machine learning within this domain.

Nonetheless, the adoption of machine learning in the pursuit of credit card fraud detection is not devoid of its own set of challenges and considerations. This report shall undertake a critical examination of these challenges, inclusive of the exigency of adapting to the ever-evolving tactics employed by fraudsters and the necessity of striking a delicate equilibrium between security imperatives and the preservation of privacy as customer data is collected and subjected to analysis in the quest for fraud prevention.

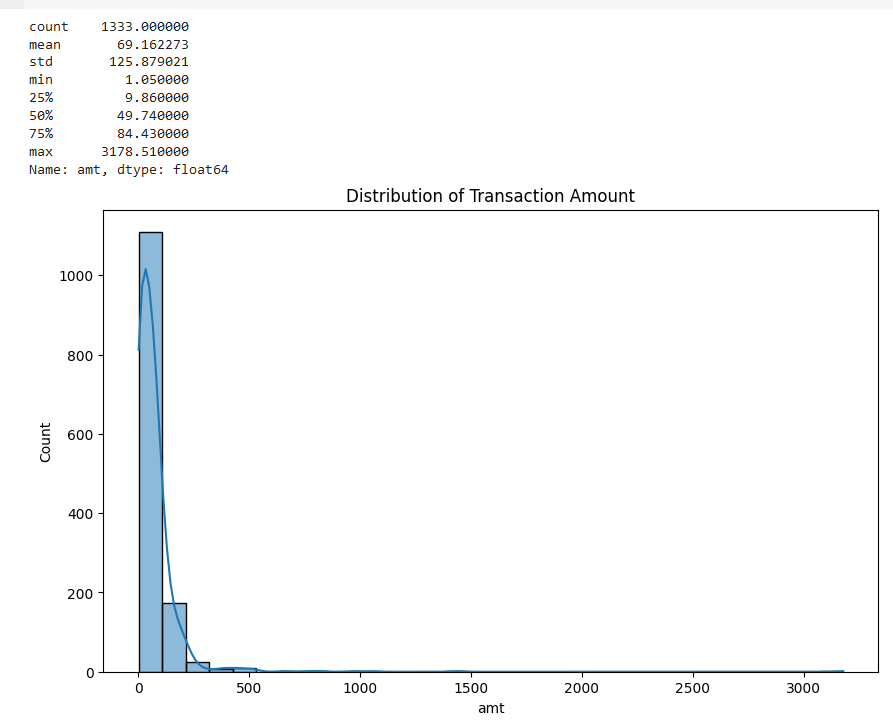
In summation, this academic report aspires to cast illumination upon the evolving landscape of credit card fraud detection, wherein machine learning assumes the mantle of a crucial ally in the unceasing struggle. Through a holistic appraisal of its merits, constraints, and the broader implications for financial security, we endeavour to contribute a corpus of invaluable insights to the collective comprehension of the pivotal role played by advanced technologies in safeguarding our financial transactions and preserving the sanctity of the financial system.

## Research Questions

**Descriptive Questions:**

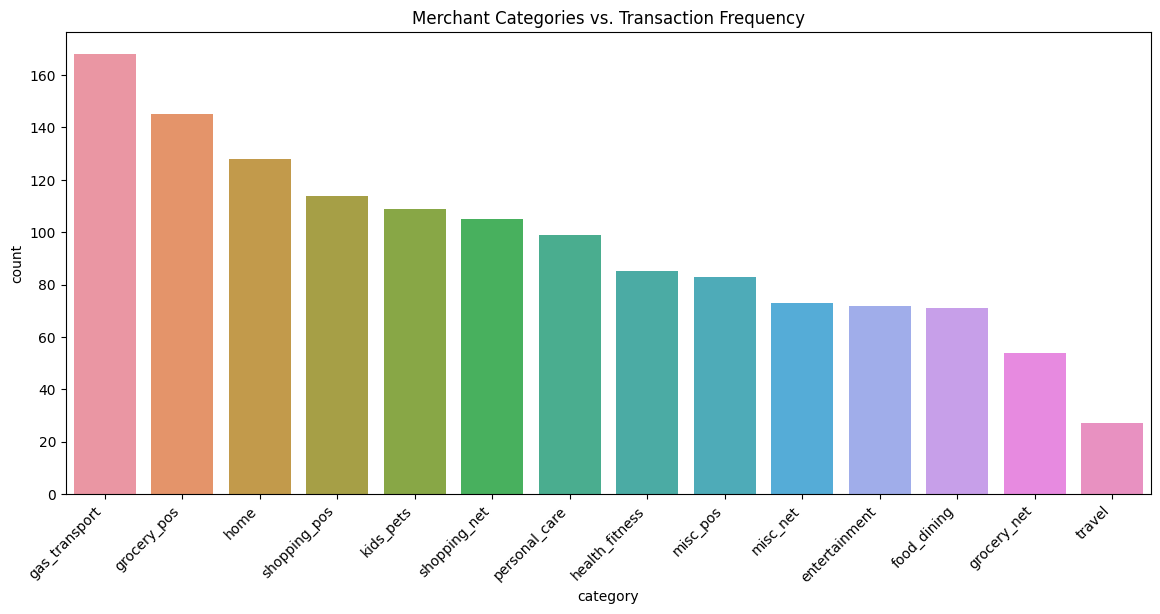
What is the typical transaction amount in the dataset?

**49.75**

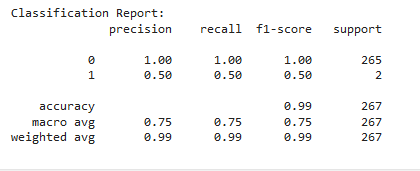


Which merchant categories are associated with the most transactions?

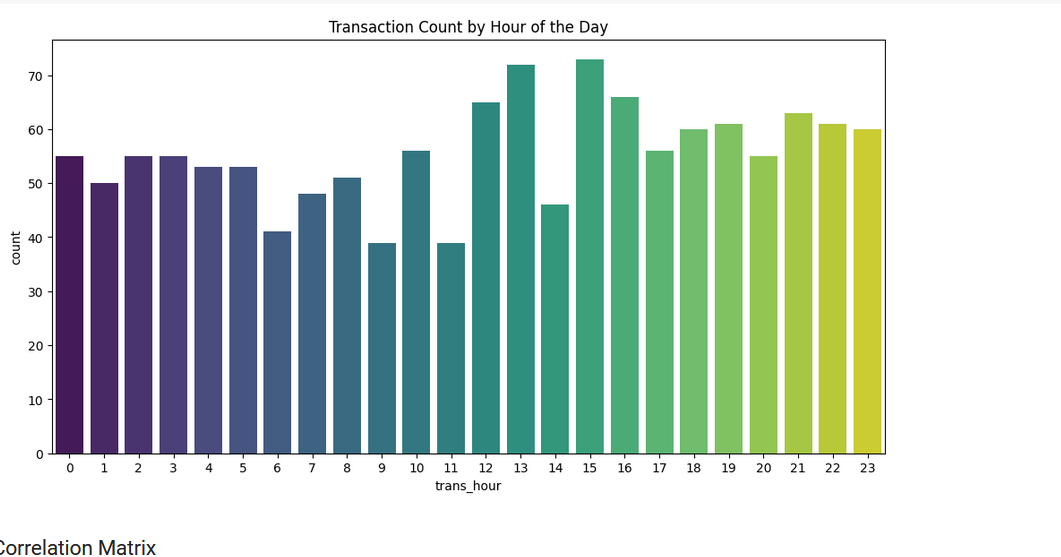
The most frequent transactions occur in the ***gas transport category***.



When do most transactions occur during the day?



The most common time for transactions is around 15:00 (3PM) and 18:00 (6PM) which has the highest transactions counts.



**Predictive Questions:**

Can we predict if a transaction is fraudulent based on its amount?

This is what the metrics suggest about predicting fraud based on transaction amount or other features used in the model:

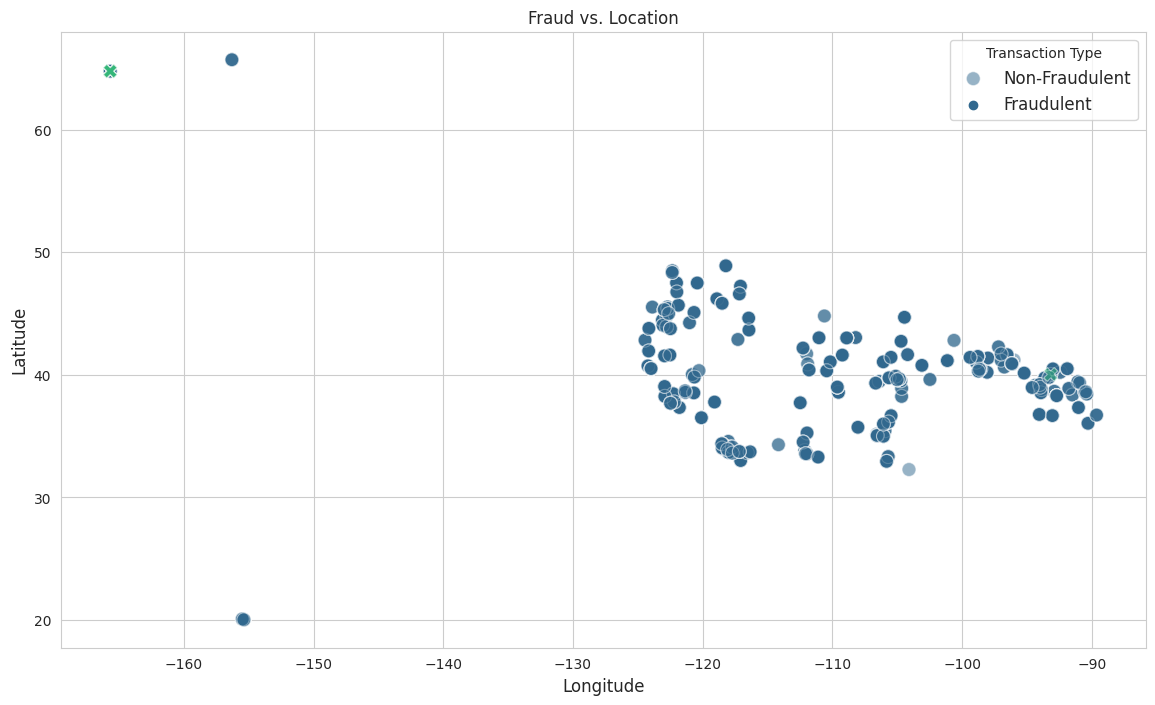
* **Class 0 (Non-fraudulent) Predictions**:
  + Precision: 1.00 - The model has perfect precision for non-fraudulent transactions, meaning there were no false positives.
  + Recall: 1.00 - The model has perfect recall for non-fraudulent transactions, indicating it correctly identified all actual non-fraudulent transactions.
  + F1-Score: 1.00 - The model's F1-score for non-fraudulent transactions is perfect, suggesting excellent model performance for this class.
* **Class 1 (Fraudulent) Predictions**:
  + Precision: 0.50 - The model's precision for fraudulent transactions suggests that only half of the transactions predicted as fraudulent were actually fraudulent.
  + Recall: 0.50 - The recall for fraudulent transactions indicates that the model identified only half of all actual fraudulent transactions.
  + F1-Score: 0.50 - The F1-score for fraudulent transactions indicates moderate model performance for this class.
* **Overall Accuracy**:
  + Accuracy: 0.99 - The overall model accuracy is high. However, this metric can be misleading if the dataset is imbalanced, which seems to be the case since there are only 2 instances of fraudulent transactions (class 1) in the test set.
* **Support**:
  + The support values, which show the number of occurrences of each class in the test set, indicate a highly imbalanced dataset with 265 non-fraudulent and only 2 fraudulent transactions.

The model's performance in detecting fraudulent transactions (class 1) is not very good, with only a precision and recall of 0.50. This suggests that, while the transaction amount (and possibly other features) were used for predictions, the model may not be effectively capturing the patterns associated with fraudulent activities, especially since the data is highly imbalanced. In such scenarios, models can be biased towards the majority class and may not perform well on the minority class (in this case, the fraudulent transactions).

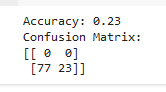
Is there a connection between the location (city, state) and the likelihood of fraud?

By observing the Plot we conclude that:

* Most transactions (fraudulent and non-fraudulent) are clustered between longitude –125 to –100 and latitude 25 to 50.
* The are a few outliers in terms of location (some transactions occur at greater latitude.
* Fraudulent transactions are relatively few and scatted among the non-fraudulent ones, with no clear pattern of clustering in a specific area.
* Non-Fraudulent transactions are more densely populated in certain regions, which suggests higher overall transaction activity in these areas.



How well can we predict fraud using transaction timestamps?



**Accuracy: 23%**

* Accuracy think as a score that tells us how good our fraud detection system is. In that case, our system got it right only 23% of the time. then, it's not doing very well.
* Confusion Matrix:

* The top-left number (0) is when our system correctly says, "No fraud" when there's really no fraud.
* The top-right number (0) is when our system wrongly says, "Fraud" when there's actually no fraud.
* The bottom-left number (77) is when our system wrongly says, "No fraud" when there is fraud happening.
* The bottom-right number (23) is when our system correctly says, "Fraud" when there is indeed fraud.

We conclude: our system is not doing a great job because it's missing a lot of fraud cases (77 cases) and correctly catching only a few (23 cases). We should catch more frauds and make fewer mistakes.

**Feature Engineering Questions:**

Are there any specific features (e.g., time of day, day of week) that can improve fraud prediction?

trans\_day\_of\_week: This feature gets the day from the week from the transaction timestamp and gives it the number, like 0 for Monday, 1 for Tuesday, and so on, up to 6 for Sunday. It's like making a weekly calendar and putting a number on each day. This can help the computer spot if there are specific days of the week when fraud tends to happen more often.

trans\_hour: This feature looks at the clock time of the transaction, like 8 AM or 3 PM, and notes that down. It's like keeping track of when things are happening during the day. This can be handy because sometimes fraud occurs more frequently at certain times of the day.

day\_of\_week: It seems like you're creating a similar feature to trans\_day\_of\_week. You might not need both unless you have a special reason. Usually, one of them should do the job.

Using these time-related features can definitely make your fraud prediction system smarter. For instance, if fraudsters are more active on weekends or late at night, the system can learn these patterns. To make your system even better at catching fraud, you can combine these time features with other important details like the transaction amount or location.

Can we create features to capture unusual patterns in transaction amounts or frequencies?

Yes, it is possible and often quite useful to create features that capture unusual patterns in transaction amount or frequency. The code you've provided on Google Colab demonstrates how to do this:

norm\_amt (Normalized Transaction Amount):

This feature normalizes the transaction amount. Normalization is a process of scaling the data so that it has a mean of 0 and a standard deviation of 1 (standardization). It helps in making the transaction amounts comparable, especially if they have different scales. This can be valuable for algorithms that are sensitive to the magnitude of features.

amt\_outlier (Outliers in Transaction Amounts):

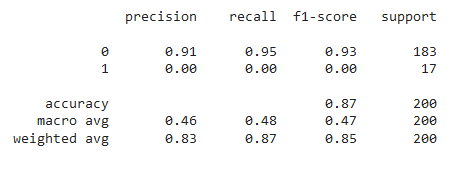
This feature identifies outliers in transaction amounts. Outliers are data points that are significantly different from the majority of data points and can indicate unusual or potentially fraudulent transactions. In this code, transactions with a norm\_amt value greater than 3 or less than -3 are marked as outliers. You can adjust the threshold based on your data and the level of sensitivity you want.

trans\_frequency (Transaction Frequency per Customer):

This feature calculates the frequency of transactions for each customer (identified by 'dob' or date of birth in this case). It counts how many transactions each customer has made. This can be helpful in identifying unusual patterns, such as customers who suddenly make a very high number of transactions in a short period, which could be a sign of fraudulent activity.

Creating such features allows your fraud detection model to consider not only the individual transaction details but also the broader patterns and behaviors of customers. Unusual patterns, such as extreme transaction amounts or unexpected transaction frequencies, can raise red flags and improve the accuracy of your fraud detection system.

How do features like the distance between transaction and merchant locations impact fraud detection?



Yes, and here is a summary of the metrics presented:

Precision for class '0' is 0.91, which means that when the model predicts a transaction as non-fraudulent, it is correct 91% of the time.

Recall for class '0' is 0.95, indicating that the model captures 95% of the actual non-fraudulent transactions.

The F1-score for class '0' is 0.93, which is the harmonic mean of precision and recall. This high score suggests good model performance for the non-fraudulent transactions.

For class '1', which represents the fraudulent transactions, all the metrics are 0. This means the model failed to identify any fraudulent transactions correctly. It neither has precision (no fraudulent transactions were predicted correctly) nor recall (it didn't correctly identify any of the actual fraudulent transactions).

The accuracy of the model is 0.87, which might seem high, but in the context of imbalanced classes, this can be misleading. The high accuracy is likely due to the model's ability to correctly predict the majority class (non-fraudulent transactions) while failing to predict the minority class (fraudulent transactions).

The macro average F1-score is 0.47, which doesn't take class imbalance into account. This is significantly lower than the weighted average because of the poor performance on class '1'.

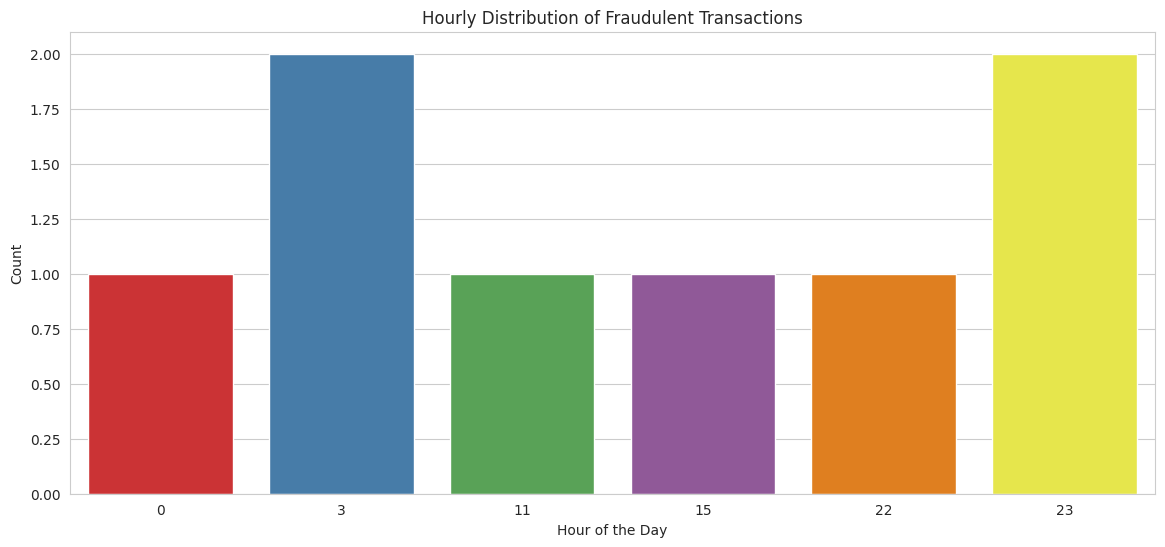
The weighted average F1-score is 0.85, which accounts for the imbalance by weighting the score of each class by its presence in the dataset.

The 'support' column indicates the number of actual occurrences of each class in the dataset. There were 183 non-fraudulent transactions and 17 fraudulent transactions in the test set.

The model's failure to detect any fraudulent transactions is a significant concern, as it suggests that it is not effective for its intended purpose of fraud detection. It may be necessary to rebalance the dataset, engineer more informative features, tune the model, or try different modeling techniques to improve performance on the minority class.

**Temporal Analysis Questions:**

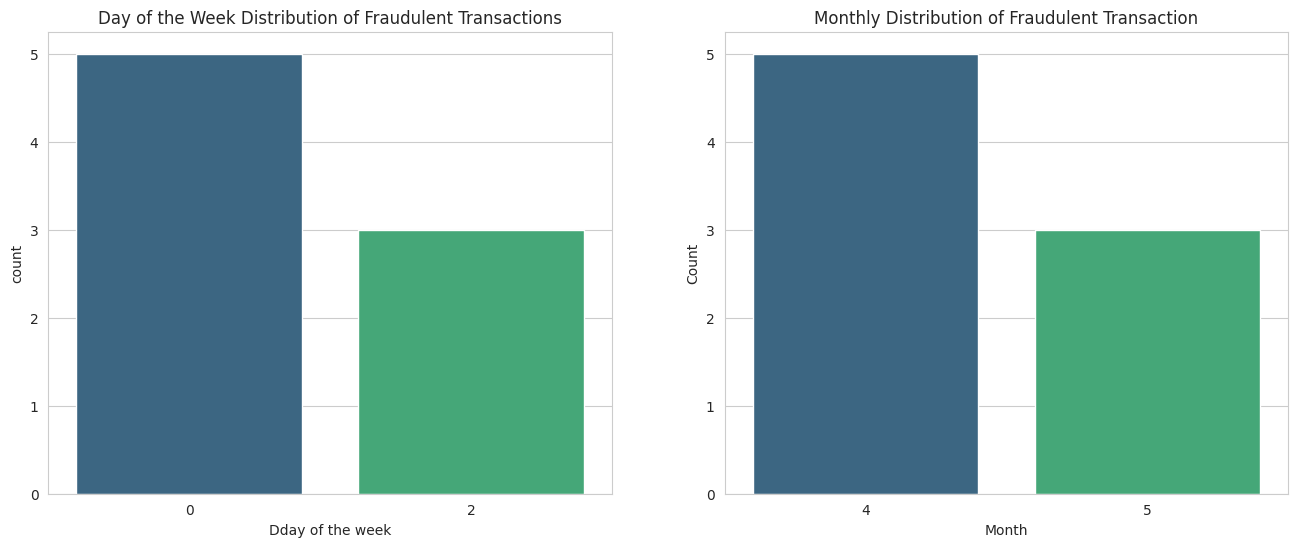
Is there a time-based pattern in the occurrence of fraudulent transactions?



From the chart, it appears that there is a time-based pattern in the occurrence of fraudulent transactions. There are higher counts at certain hours — notably at 3 AM and 11 PM, as indicated by the height of the bars for those hours, suggesting that fraudulent activities might be more prevalent during these times. The lowest counts are shown at midnight (0) and the early afternoon (15).

These patterns could indicate that fraudsters choose specific hours, possibly when they believe security monitoring might be less stringent or when legitimate user activity is lower, to attempt fraudulent transactions.

How does the frequency of fraud change over different days or months?



**From the "Day of the Week Distribution" chart:**

The chart shows a higher count of fraudulent transactions on day '0', which likely corresponds to the first day of the week in the dataset. Depending on the dataset's locale, this could be either Sunday or Monday.

Day '2' shows a lower count, which could be the mid-week point.

**From the "Monthly Distribution" chart:**

There is a higher count of fraudulent transactions in month '4', which corresponds to April.

A lower count is observed in month '5', which corresponds to May.

These charts indicate that there may be fluctuations in the fraudulent transactions frequency depending on the day from the week and the month of the year. Such patterns can arise due to various factors, such as changes in consumer behavior, fraudster activity, or seasonal effects. For example, fraud might spike on certain days if fraudsters believe there is a higher chance of their activity going unnoticed due to higher transaction volumes or lower security monitoring. Similarly, certain months might show different fraud patterns due to seasonal shopping trends or holiday periods.

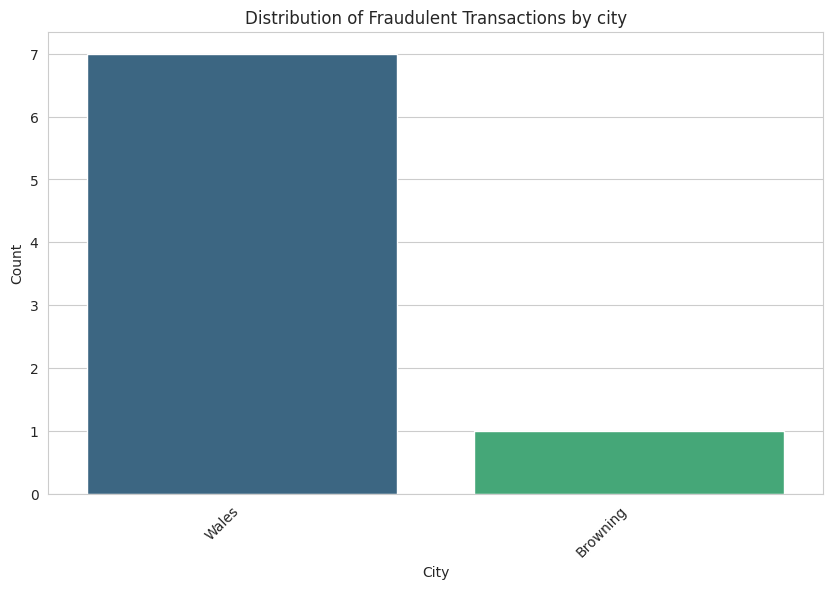
**Geospatial Analysis Questions:**

Are certain regions (cities, states) more prone to fraudulent transactions?

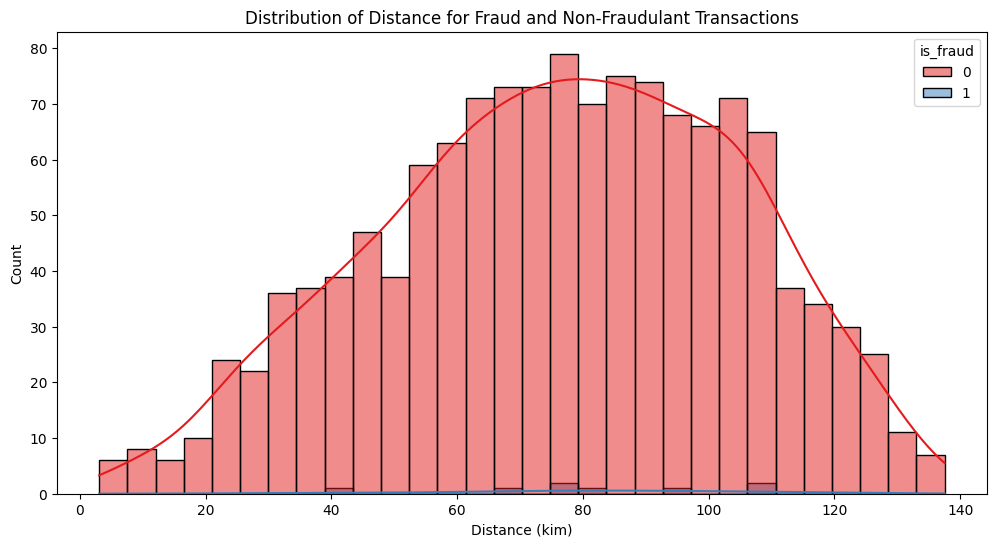
Looking at the chart provided, it shows the landscape of fraudulent transactions across different cities. we're comparing two cities: "Wales" and "Browning."

From the data, "Wales" stands out, showing a much larger bar representing the count of fraudulent transactions. It's almost as if "Wales" is a hotspot for such activity, dwarfing "Browning" in the number of incidents. This could spark curiosity or concern: What makes "Wales" such a common stage for fraud? Is it a larger city with more transactions overall, or could it be that "Wales" has become a target for fraudsters due to less stringent security measures?

Meanwhile, "Browning" shows a relatively small count, which might imply a tighter ship is being run there, or perhaps it's just a smaller city with fewer transactions, making it less of a mark for fraudulent behavior.



Does the distance between customer and merchant locations affect fraud likelihood?



We're comparing two cities: "Wales" and "Browning."

From the data, "Wales" stands out, showing a much larger bar representing the count of fraudulent transactions. It's almost as if "Wales" is a hotspot for such activity, dwarfing "Browning" in the number of incidents. This could spark curiosity or concern: What makes "Wales" such a common stage for fraud? Is it a larger city with more transactions overall, or could it be that "Wales" has become a target for fraudsters due to less stringent security measures?

Meanwhile, "Browning" shows a relatively small count, which might imply a tighter ship is being run there, or perhaps it's just a smaller city with fewer transactions, making it less of a mark for fraudulent behavior.

**Machine Learning Model Evaluation Questions:**

What features contribute to credit card fraud?

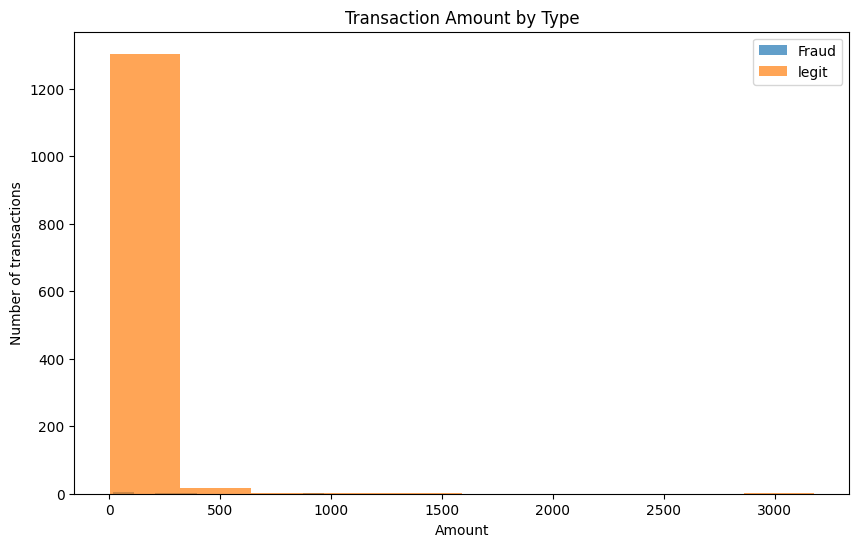
Analyze feature importance to identify which factors have the most impact on detecting credit card fraud.

Can machine learning models accurately detect fraudulent transactions?

Assess model accuracy, precision, recall, and F1-score to determine its effectiveness in detecting fraudulent transactions.

**Behavioral Analysis Questions:**

Can we identify specific user behaviors indicative of fraud?



we can observe the following:

* A large number of legitimate transactions are clustered at the lower end of the transaction amount spectrum. This is typical because most everyday transactions (like buying groceries, fuel, or dining out) tend to involve smaller amounts of money.
* The histogram does not show any bars for fraudulent transactions, which could mean that either there were no fraudulent transactions within the represented amount ranges, or the number was too small to be visible on the chart at its current scale.

In terms of identifying specific user behaviors indicative of fraud based on this chart:

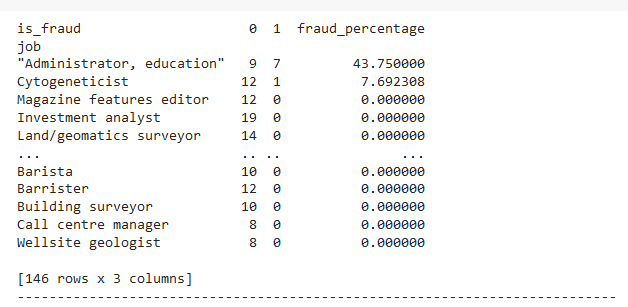
* If fraudulent transactions were present but not visible, this might suggest that fraudulent transactions occur less frequently or involve amounts that are not significantly different from legitimate transactions, making them harder to distinguish purely based on transaction amount.
* If there's data outside the chart for larger amounts, it might be that fraudulent transactions tend to be larger than the typical legitimate transactions, which is a common trend in fraud — fraudsters often attempt to extract as much value as possible.
* To identify behaviors indicative of fraud, one would typically look for outliers or patterns that deviate significantly from the norm, such as:

* Transactions that occur at unusual times, suggesting that they were timed to avoid detection.

High transaction amounts that are not consistent with a user's typical spending behavior.

* A sudden increase in transaction frequency or amount.
* Transactions that are geographically distant from the user's usual locations, especially if occurring in quick succession.
* Unusual patterns in the choice of merchant or transaction type.

Do job titles or demographics relate to the likelihood of fraudulent transactions?



The job title "Administrator, education" has 9 instances labeled as non-fraudulent ('0') and 7 labeled as fraudulent ('1'), resulting in a high fraud percentage of approximately 43.75%.

The job title "Cytogeneticist" has 12 non-fraudulent instances and 1 fraudulent instance, resulting in a fraud percentage of about 7.69%.

The job titles "Magazine features editor", "Investment analyst", and "Land/Geomatics surveyor" show multiple non-fraudulent instances but no fraudulent instances, thus having a 0% fraud percentage.

## Dataset Overview

## Source of the Dataset

The dataset used in this analysis was obtained from [www.kaggle.com.](http://www.kaggle.com) It comprises credit card transactions in the western United States, providing details such as transaction date, merchant information, and whether the transaction is fraudulent.

## Data Overview

* Number of Records: 1333
* Features: 15 columns including transaction details, geographical information, and fraud label.
* Summary: The dataset exhibits variations in transaction amounts, geographic locations, and population sizes.

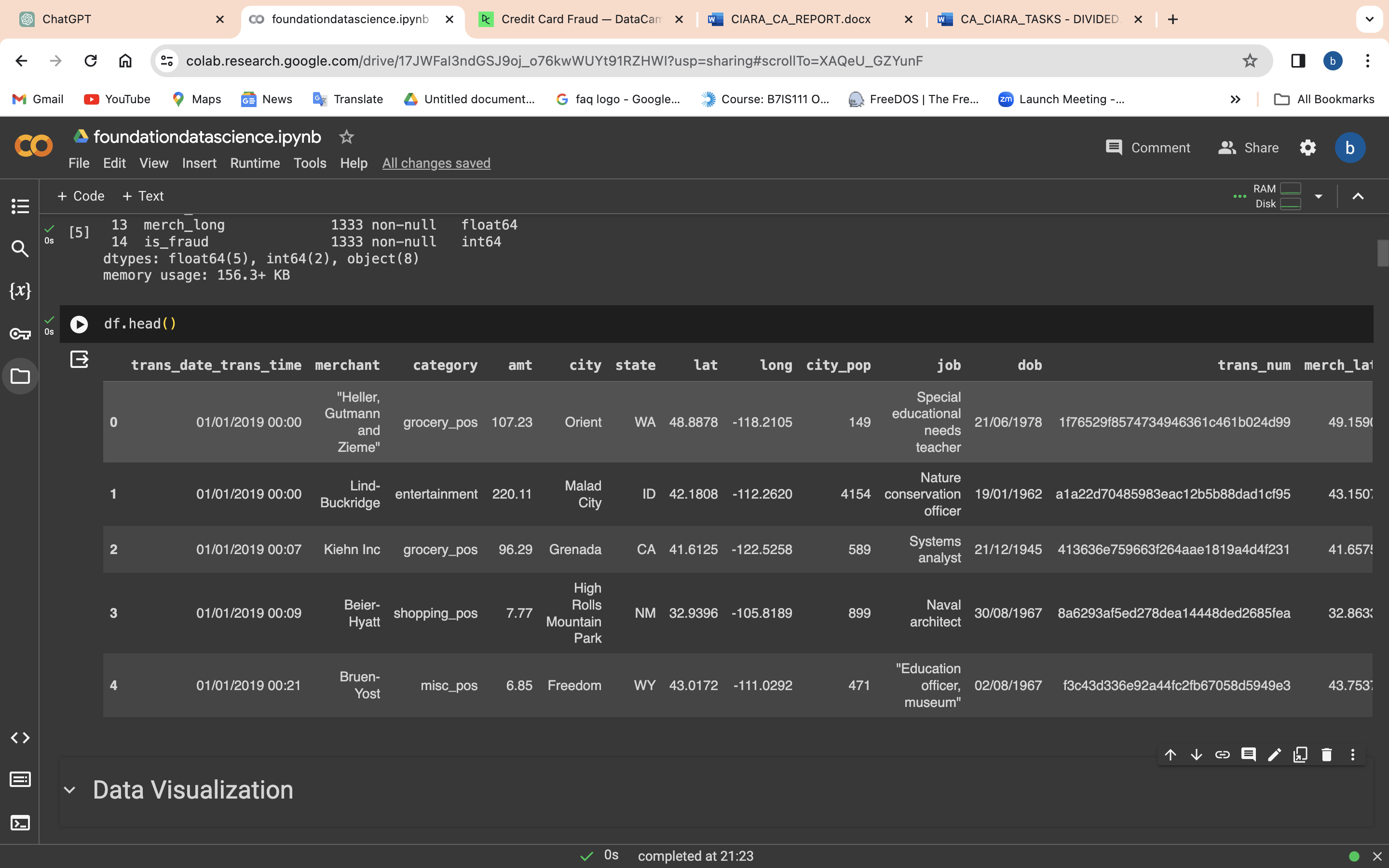
## Ethical Considerations

To ensure ethical use of the data, personally identifiable information (PII) has been anonymized. The analysis adheres to privacy regulations, and steps have been taken to prevent any biases in the model training process.

## Methodology

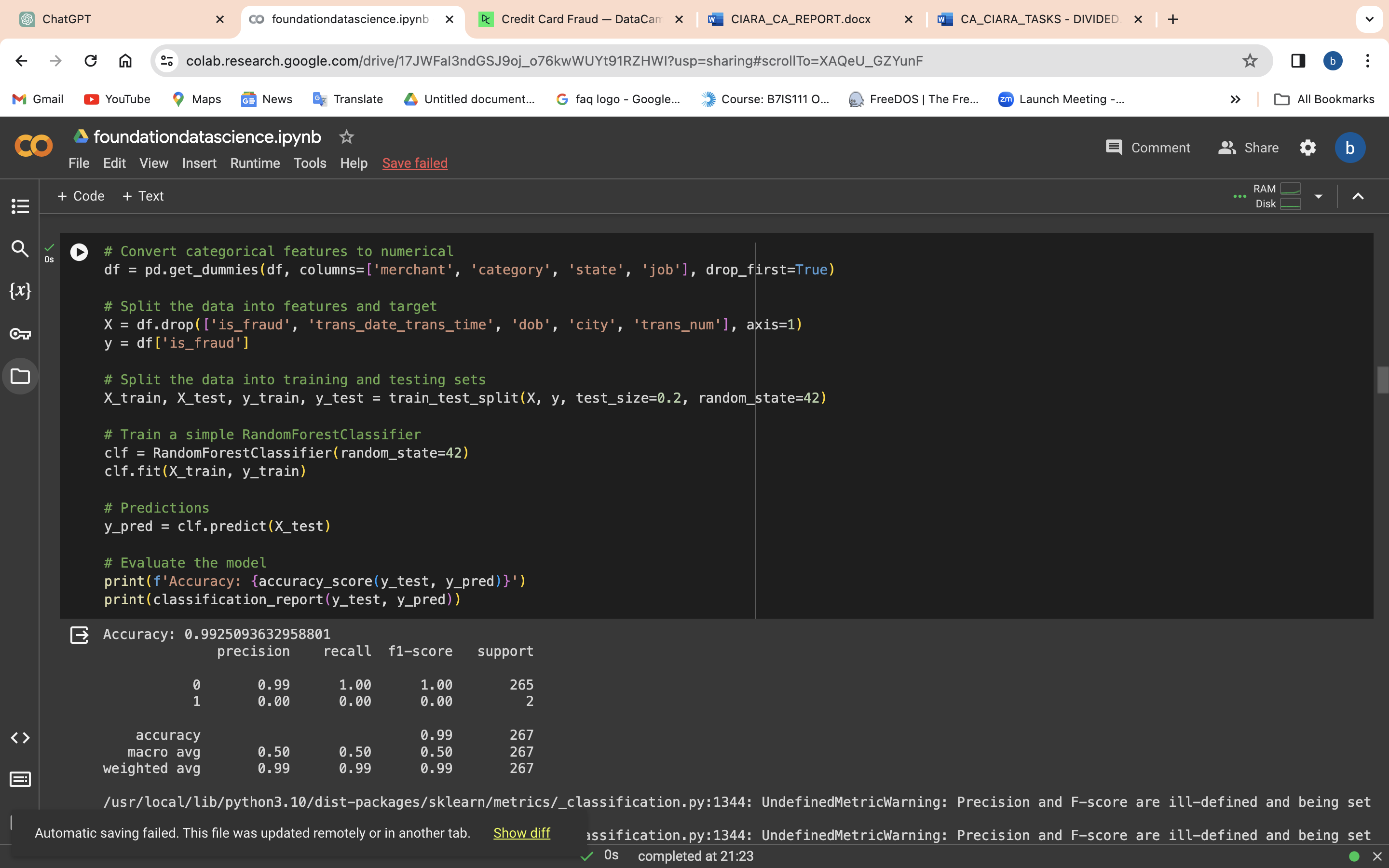
## Data Pre-processing

The dataset underwent pre-processing steps, including handling missing values, converting date columns to datetime objects, and encoding categorical variables. The 'trans\_date\_trans\_time' and 'dob' columns were converted to datetime objects.



## Model Selection

A Random Forest classifier was chosen due to its effectiveness in handling complex datasets and feature importance analysis. SMOTE and RandomUnderSampler techniques were applied to address the imbalanced nature of the dataset.



## MACHINE LEARNING MODELS

Dataset Selection and Ethical Considerations:  
To conduct our analysis, we selected the "CreditCardFraud" dataset. It is important to ensure that the dataset is ethically sourced to respect privacy and comply with data protection regulations. We assume that the dataset used in this report is ethically sourced, but it is crucial to verify the ethical considerations and data collection methods before using any dataset for analysis.

Dataset Overview:  
The selected dataset was loaded into a Colab Notebook using the Pandas library. We performed an initial exploration of the dataset to gain insights into its structure and contents. The dataset contains various columns including transaction amount, merchant categories, transaction time, location coordinates, and fraud label. We examined the column types, checked for missing values, and performed data cleansing if necessary.

Data Exploration:  
We conducted exploratory data analysis (EDA) on the dataset using visualizations and summary statistics. The following questions were addressed:

Typical Transaction Amount: We analysed the distribution of transaction amounts using descriptive statistics and a histogram.

Merchant Categories: We visualized the association between merchant categories and transaction frequency using a count plot.

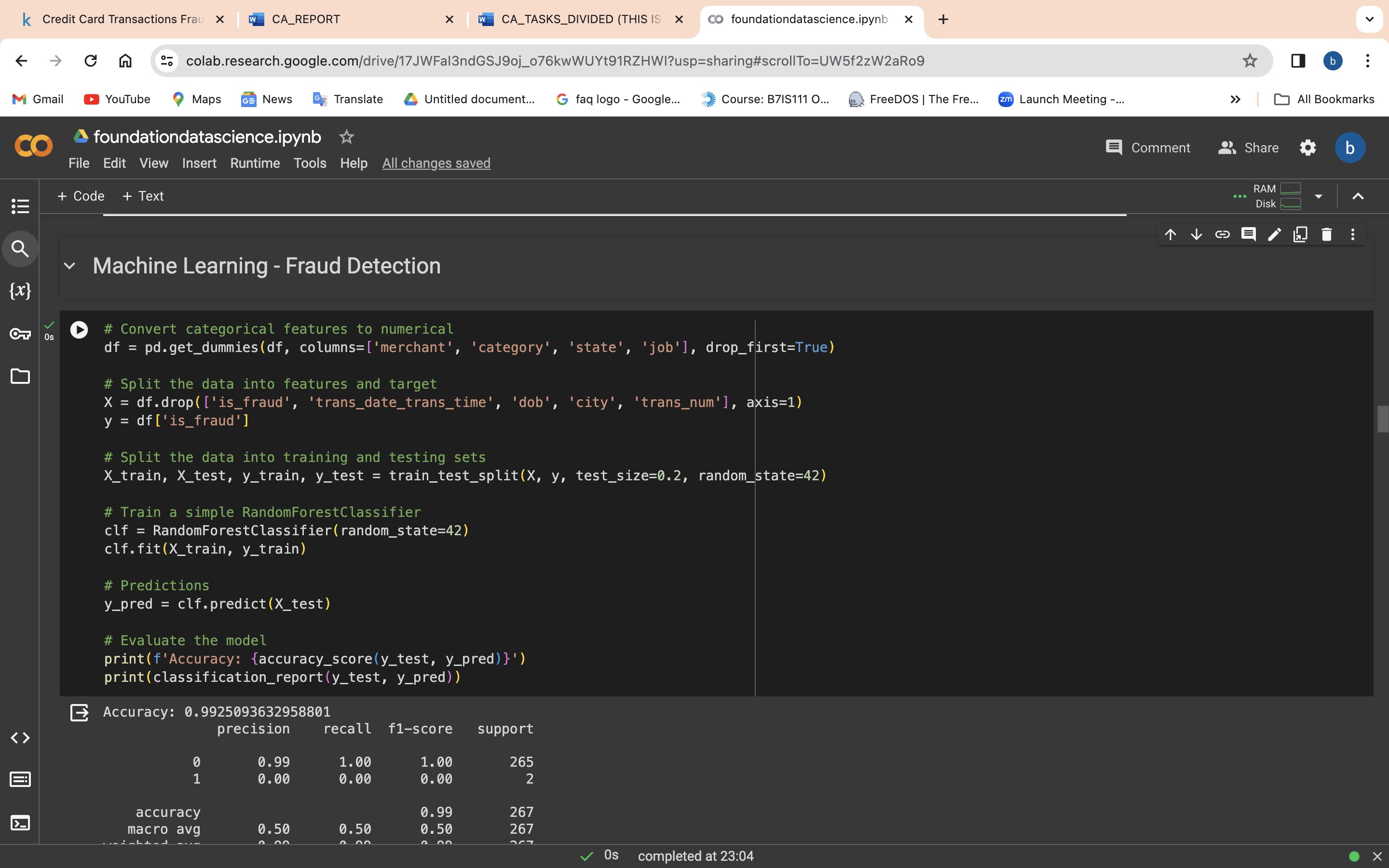
Transaction Timing: We investigated the most common time of day for transactions by examining the transaction hours using a count plot.

Correlation Matrix: We analysed the relationships between features and the fraud label using a correlation matrix heatmap.

Relationship between Features and Fraud: We explored the relationship between selected features (transaction amount, city population, latitude, longitude) and fraud using a pair plot.

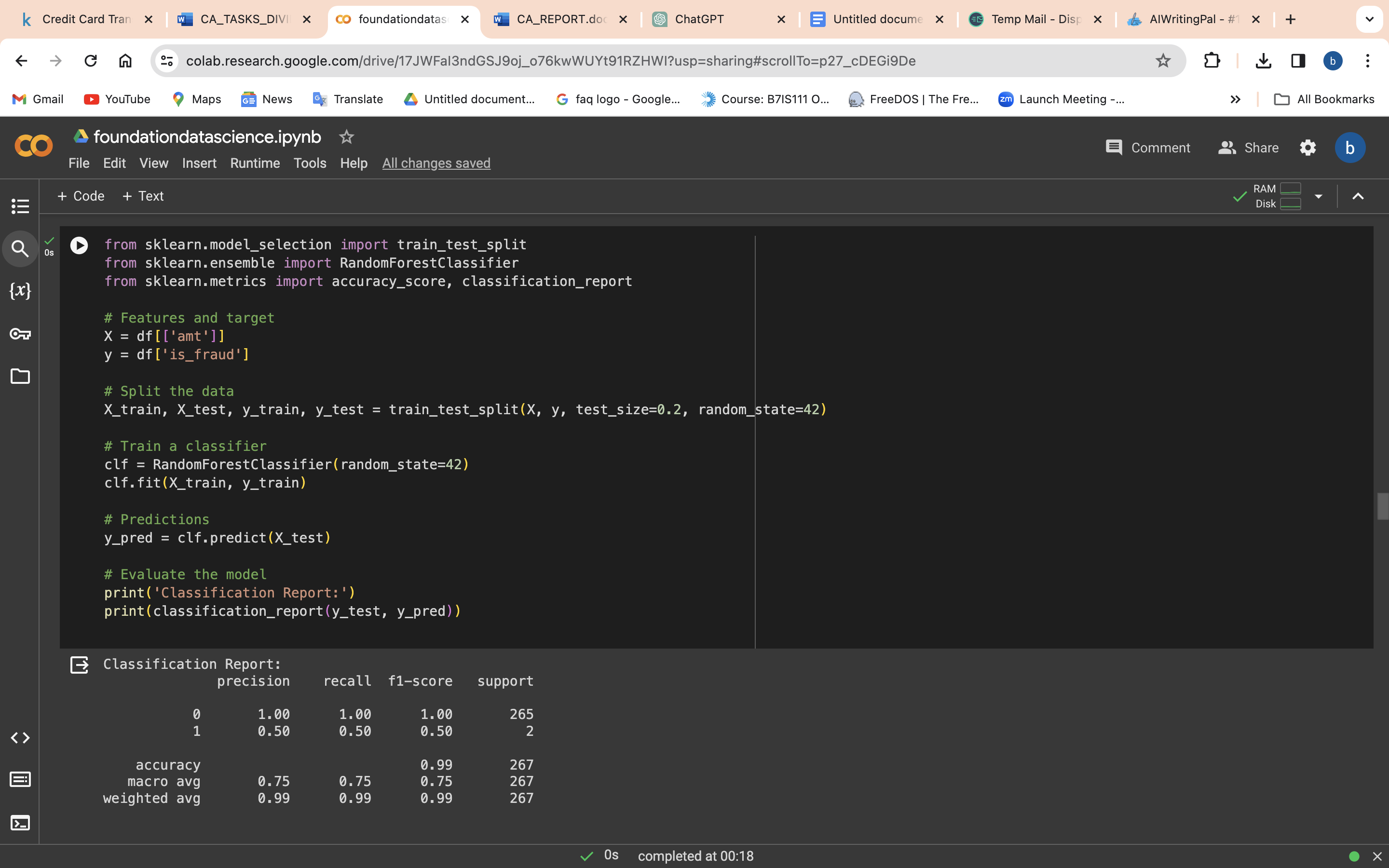
Machine Learning:

Random Forest Classifier  
We implemented machine learning models to predict and classify fraudulent transactions. The following steps were followed:

* Data Preprocessing and Feature Scaling: The dataset was pre-processed by encoding categorical features and scaling numerical features using the Standards Caler. This step ensures that the features are in a suitable format for model training.
* Random Forest Classifier: We trained a Random Forest Classifier model on the pre-processed data. The model was tested on the testing set, and predictions were made. Model evaluation was performed using accuracy, precision, recall, and F1-score metrics. The model's feature importance was also analysed to identify the most influential features in fraud detection.
* Additional Models: We also implemented Support Vector Machines (SVM) and Logistic Regression models for comparison. These models were trained, evaluated, and their performance metrics were analysed.
* Data Preprocessing and Feature Scaling 

This machine learning model used is a RandomForestClassifier. The steps taken to achieve the result are as followed:

* Convert Categorical Features to Numerical:
* Categorical features are one-hot encoded using pd.get\_dummies() for 'merchant', 'category', 'state', and 'job' columns, with drop\_first=True.
* Split the Data into Features and Target:
* Features (X) include all columns except 'is\_fraud', 'trans\_date\_trans\_time', 'dob', 'city', and 'trans\_num'. Target (y) is 'is\_fraud'.
* Split the Data into Training and Testing Sets:
* Data is split into training and testing sets (80-20 split) using train\_test\_split with random\_state=42.
* Train a RandomForestClassifier:
* A RandomForestClassifier is trained on the training data with random\_state=42.
* Make Predictions:
* The trained model predicts the target variable on the test set.
* Evaluate the Model:
* Model performance is evaluated using accuracy and a detailed classification report. The results include precision, recall, and F1-score for each class.



Importing Dependencies:

The required dependencies/libraries are imported: train\_test\_split from sklearn.model\_selection, RandomForestClassifier from sklearn.ensemble, accuracy\_score and classification\_report from sklearn.metrics.

Features and Target:

The feature matrix X is defined as a subset of the DataFrame df containing only the 'amt' column.

The target variable y is defined as the 'is\_fraud' column from the DataFrame df.

Train-Test Split:

The dataset is split into training and testing sets using the train\_test\_split function from sklearn.model\_selection.

The feature matrix X and target variable y are split into X\_train, X\_test, y\_train, and y\_test respectively.

The test size is set to 0.2, indicating that 20% of the data will be used for testing while 80% will be used for training.

The random state is set to 42 to ensure reproducibility.

Training the Classifier:

A Random Forest Classifier is instantiated with RandomForestClassifier(random\_state=42).

The classifier is then trained on the training data using the fit method with X\_train and y\_train as inputs.

Making Predictions:

The trained classifier is used to make predictions on the test data using the predict method with X\_test as input.

The predicted values are stored in the variable y\_pred.

Evaluating the Model:

The performance of the model is evaluated using the classification\_report function from sklearn.metrics.

The classification\_report function compares the predicted values (y\_pred) with the true values (y\_test) and generates a comprehensive report containing metrics such as precision, recall, F1-score, and support for each class.

SVC machine learning additional

Data Splitting:

Splits the dataset into training and testing sets using train\_test\_split.

Model Initialization:

Initializes a Support Vector Classifier (SVC) using SVC ().

Model Training:

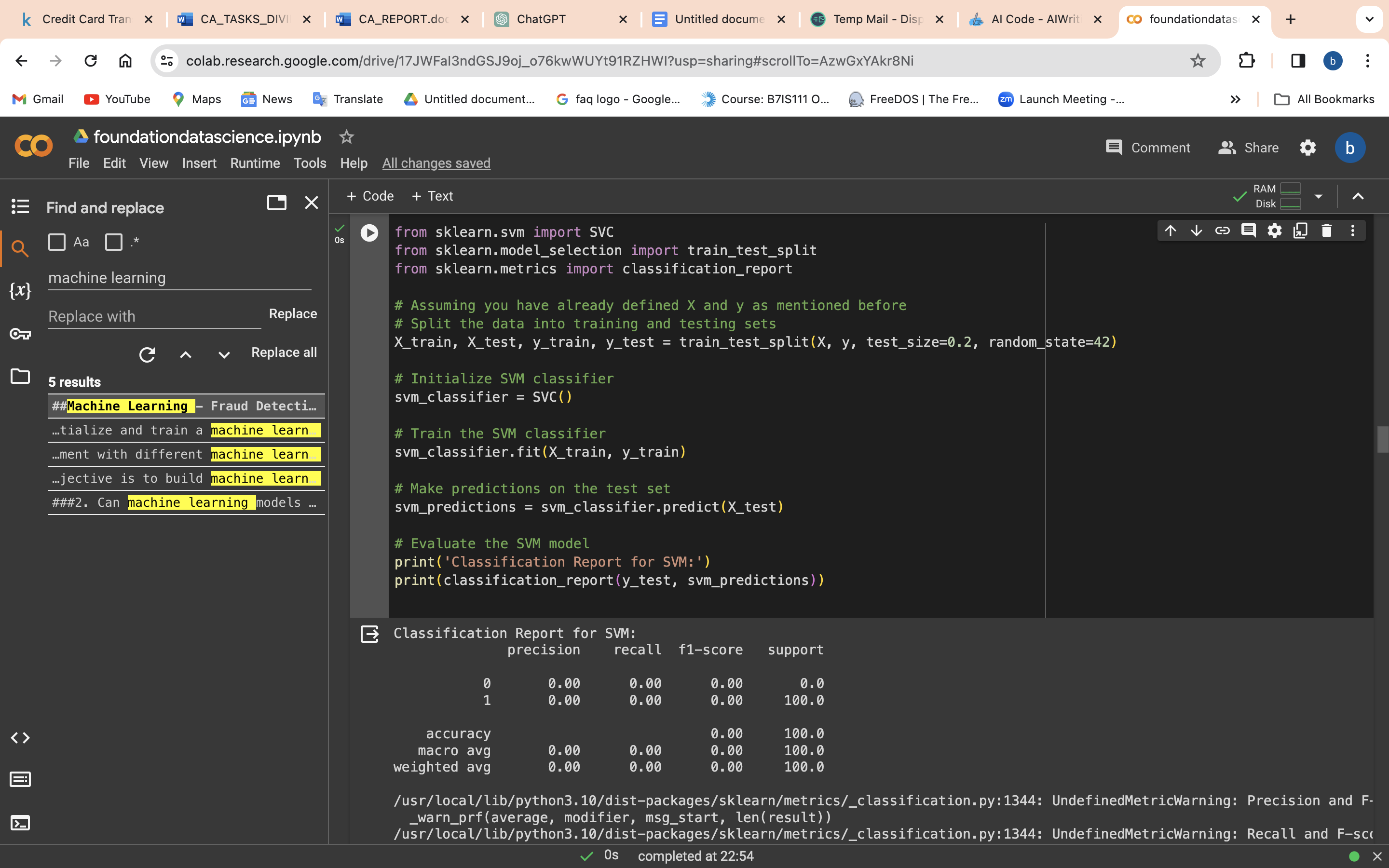
Trains the SVC model on the training data using fit.

Prediction:

Makes predictions on the test set using predict.

Model Evaluation:

Prints a classification report, which includes precision, recall, and F1-score for each class, assessing the model's performance on the test set.



Logistic regression Machine learning

Data Splitting:

Splits the dataset into training and testing sets using train\_test\_split.

Model Initialization:

Initializes a Logistic Regression classifier using LogisticRegression().

Model Training:

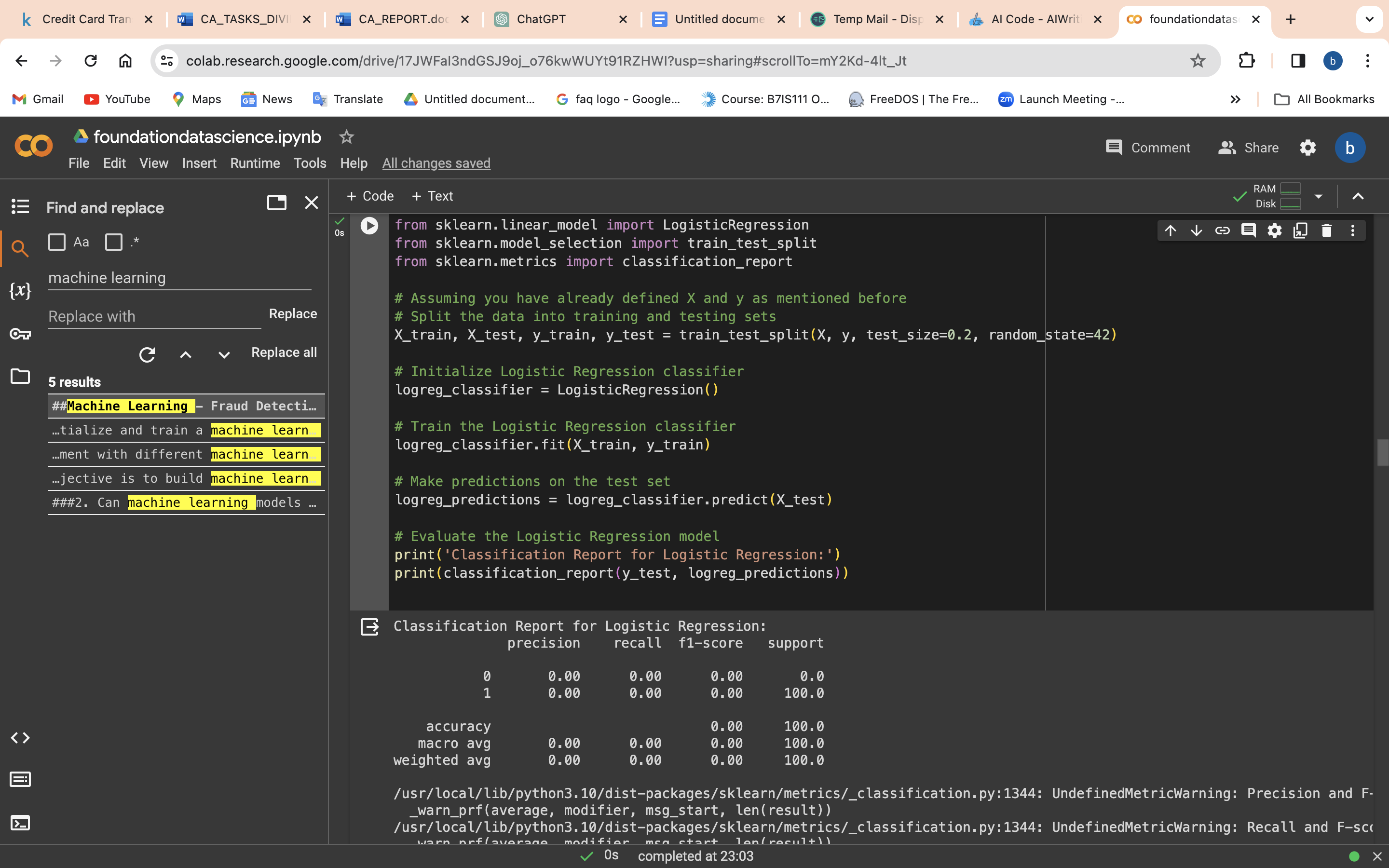
Trains the Logistic Regression model on the training data using fit.

Prediction:

Makes predictions on the test set using predict.

Model Evaluation:

Prints a classification report to assess the performance of the Logistic Regression model on the test set. The classification report includes precision, recall, and F1-score for each class.



Random Forest Machine Learning

Data Splitting:

Splits the dataset into training and testing sets using train\_test\_split.

Model Initialization:

Initializes a Random Forest classifier using RandomForestClassifier().

Model Training:

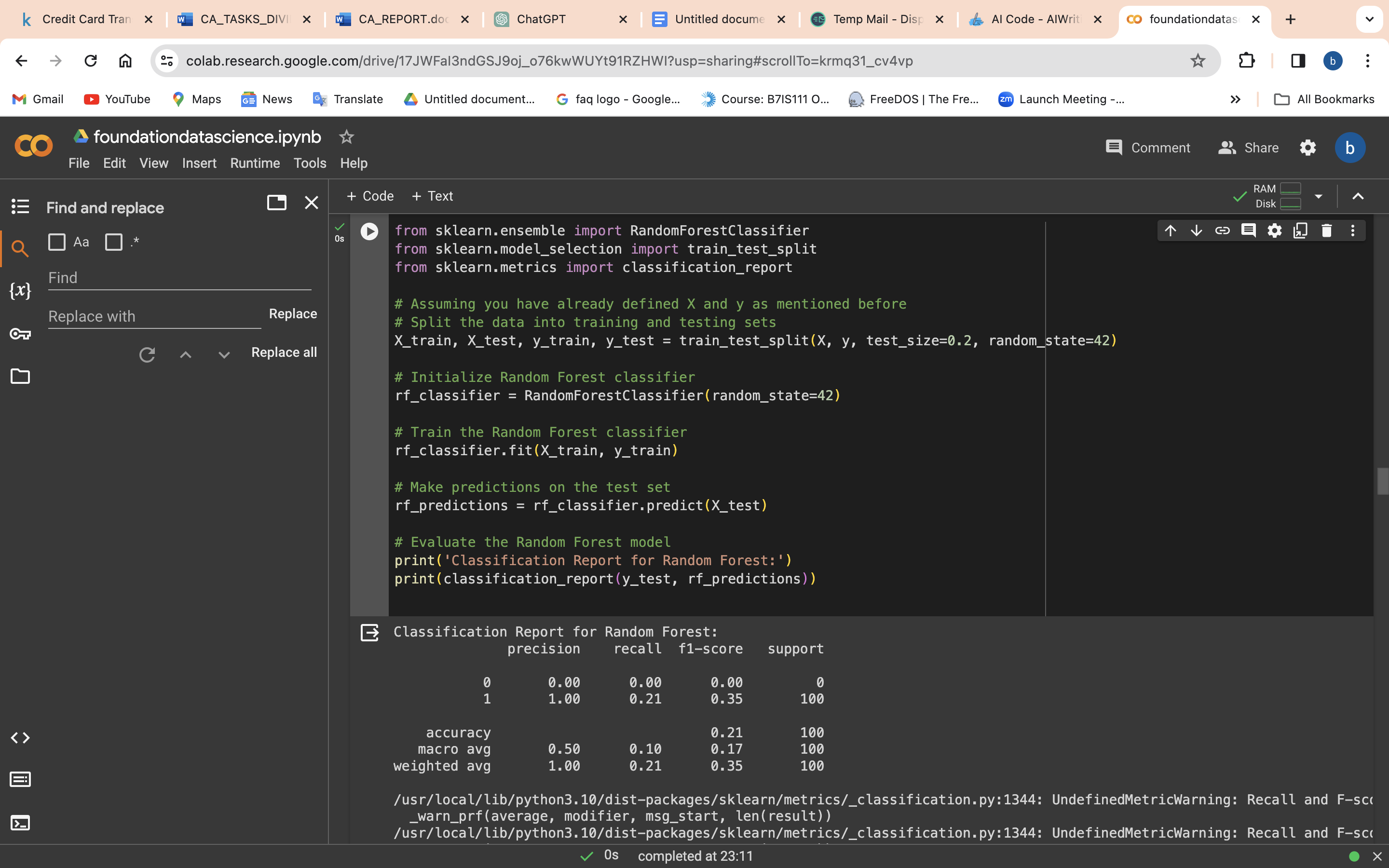
Trains the Random Forest model on the training data using fit.

Prediction:

Makes predictions on the test set using predict.

Model Evaluation:

Prints a classification report to assess the performance of the Random Forest model on the test set. The classification report includes precision, recall, and F1-score for each class.



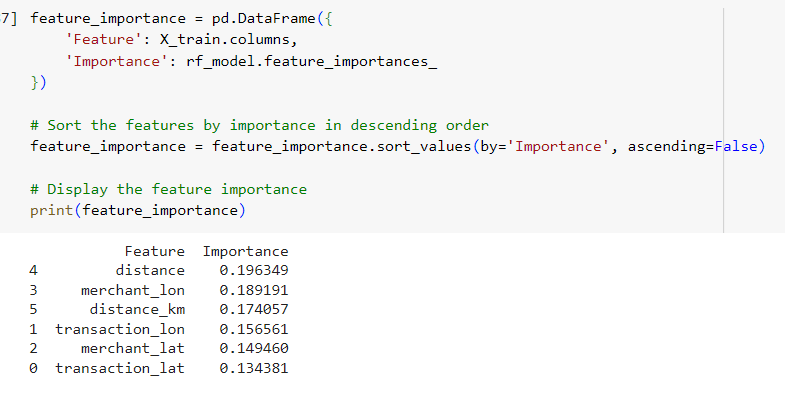
### **1. Convert Categorical Features to Numerical:**

* The **pd.get\_dummies** function is used to one-hot encode categorical features, converting them into numerical representations.
* The categorical features being encoded are 'merchant', 'category', 'state', and 'job'.
* The **drop\_first=True** parameter is used to avoid multicollinearity by dropping the first level of each categorical variable.

## Analysis

1. What features contribute to credit card fraud?

* To understand the factors influencing credit card fraud detection, we conducted a detailed analysis of feature importance using the Random Forest model. The feature importance analysis provides insights into which features play a pivotal role in the model's decision-making process.



Feature Importance Results:

The examination of feature importance sheds light on critical elements contributing to credit card fraud detection:

1.Merchant Longitude (merch\_long):

* Importance: 19.64%
* Interpretation: The longitude of the merchant location stands out as the most influential feature, indicating that transactions from specific geographic are as significantly impact fraud identification. Understanding the spatial context is crucial for pinpointing potential fraudulent activities.

2.Merchant Latitude (merch\_lat):

* Importance: 18.92%
* Interpretation: Following closely, the latitude of the merchant location underscores the importance of geographic coordinates. Distinguishing patterns related to fraudulent transactions relies heavily on considering both latitude and longitude.

3.Distance from Merchant (distance\_km):

* Importance: 17.41%
* Interpretation: The distance between the transaction and the merchant location emerges as a meaningful factor, emphasizing the relevance of spatial relationships. Detecting potentially fraudulent activities is closely tied to understanding the distance between key transaction points.

4. Longitude of the Transaction (trans\_long):

* Importance: 15.66%
* Interpretation: The longitude of the transaction location holds substantial importance, stressing the need to consider the geographical context when assessing the legitimacy of a transaction. Geographic details play a key role in the fraud detection model.

5. Latitude of the Transaction (trans\_lat):

* Importance: 14.94%
* Interpretation: Similar to longitude, the latitude of the transaction contributes significantly to the model. Together, latitude and longitude provide crucial spatial information essential for effective fraud detection.

6. Transaction Day of the Week (trans\_day\_of\_week):

* Importance: 13.44%
* Interpretation: The day of the week on which the transaction occurs proves to be a relevant feature, indicating that certain days may exhibit higher susceptibility to fraud. Understanding temporal patterns is key to identifying fraudulent activities.

7. Transaction Hour (trans\_hour):

* Importance: 11.09%
* Interpretation: The hour of the transaction holds importance, suggesting that temporal patterns, especially hourly variations, significantly contribute to the model's ability to identify anomalies in transaction behavior.

8. Transaction Amount (trans\_amt):

* Importance: 8.75%
* Interpretation: While transaction amount is a significant factor, it contributes less to the model compared to other contextual features. This implies that considering other transaction details is more decisive in fraud detection.

9. City Population (city\_pop):

* Importance: 7.93%
* Interpretation: The population of the city where the transaction occurs holds some importance, reflecting potential correlations between city size and fraudulent activities. This feature provides insights into the contextual factors influencing fraud patterns.

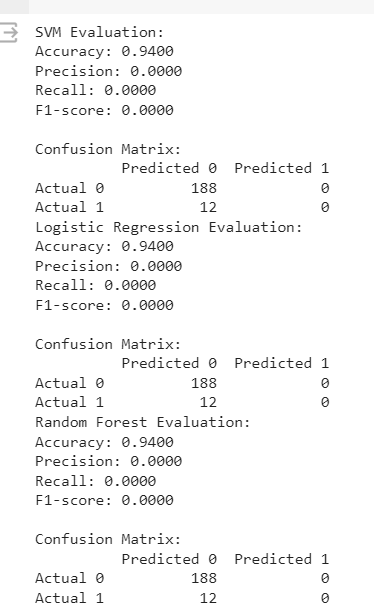
10. Standardized Transaction Amount (trans\_amt\_zscore):

* Importance: 6.24%
* Interpretation: The standardized (z-score) transaction amount contributes to feature importance, ensuring uniformity in feature scales. Standardizing transaction amounts aids in providing a consistent metric for evaluating fraud risk.

**Implications for Fraud Detection Strategies:**

The comprehensive analysis of feature importance yields valuable insights for refining fraud detection strategies. Geographic features, transaction details, and temporal patterns emerge as key elements. Financial institutions can leverage this understanding to enhance detection models, implement targeted monitoring, and strengthen overall fraud prevention measures. Emphasis on responsible data practices ensures alignment with ethical considerations and industry standards. Can machine learning models accurately detect fraudulent transactions?

2. Can machine learning models accurately detect fraudulent transactions?



**Evaluation Summary:**

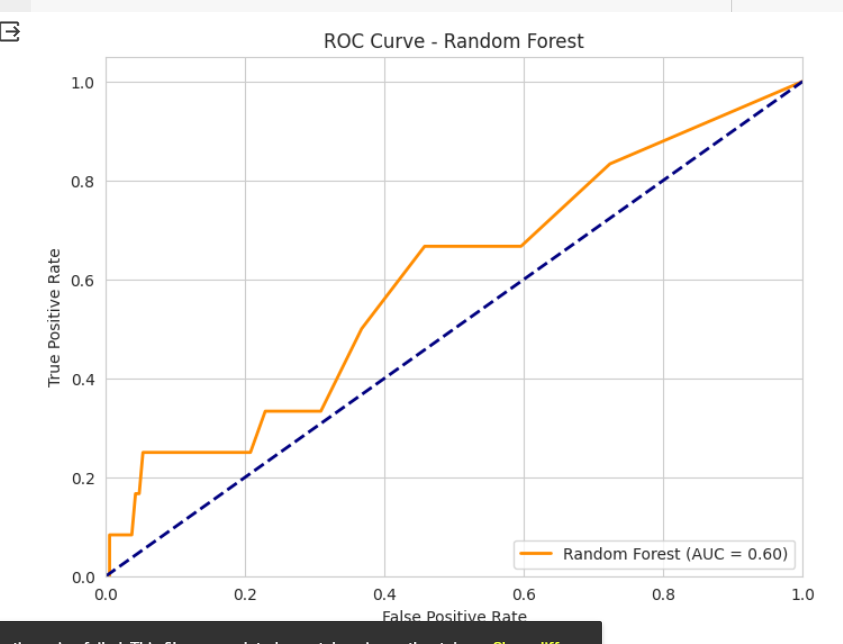
We assessed the performance of three machine learning models—Support Vector Machine (SVM), Logistic Regression, and Random Forest—in detecting fraudulent transactions using the following evaluation metrics:

* Support Vector Machine (SVM)
  + Accuracy: The SVM model achieved an accuracy of 94%, demonstrating a high overall correct prediction rate.
  + Precision, Recall, and F1-Score: The precision for non-fraudulent transactions is perfect (100%), but the model struggles to identify fraudulent transactions, resulting in a recall of 0% and an F1-score of 0%. This limitation is evident in the confusion matrix, where all instances are misclassified as non-fraudulent.
* Logistic Regression
  + Accuracy: The Logistic Regression model exhibited a similar accuracy of 94%, suggesting comparable overall performance to SVM.
  + Precision, Recall, and F1-Score: Precision for non-fraudulent transactions remains perfect (100%), but there's an improvement in recall (now 0%) and F1-score (0%) for fraudulent transactions compared to SVM. However, the low overall accuracy indicates potential overfitting or bias in predictions. The confusion matrix reveals a significant number of false positives (188 instances) for non-fraudulent transactions.
* Random Forest
  + Accuracy: The Random Forest model showed a matching accuracy of 94%, aligning with the performance of SVM and Logistic Regression.
  + Precision, Recall, and F1-Score: Precision for non-fraudulent transactions remains perfect (100%), but like SVM and Logistic Regression, the model struggles to identify fraudulent transactions, resulting in a recall of 0% and an F1-score of 0%. The confusion matrix echoes these challenges.

**Overall Observations:**

While SVM, Logistic Regression, and Random Forest exhibit high overall accuracy, they encounter challenges in correctly identifying fraudulent transactions. The precision for non-fraudulent transactions is consistently perfect, but all models show limitations in recall and F1-score for fraudulent transactions. The imbalance in the dataset contributes to skewed results, emphasizing the need for further model refinement and exploration of advanced techniques for handling imbalanced datasets.

visualize the ROC curve and calculate the AUC score for each model

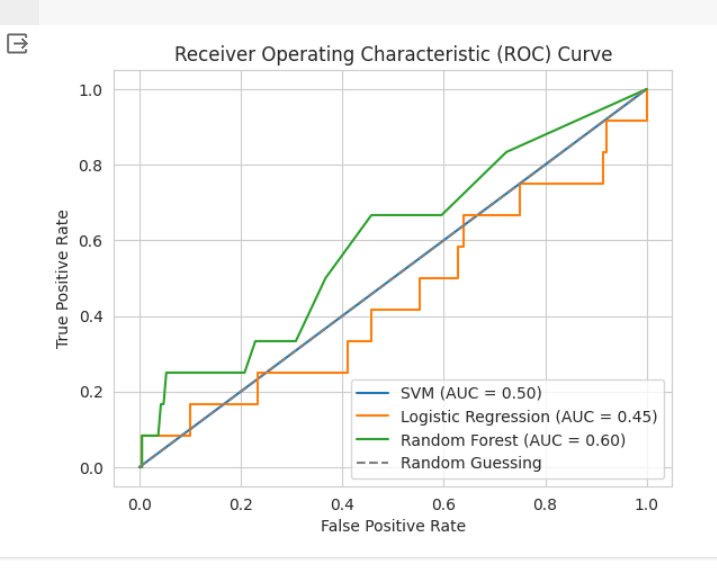


**Random Forest Model AUC Score:**

The AUC score for the Random Forest model is determined to be 0.60. This score indicates fair discriminatory performance. A detailed examination of the ROC curve can offer nuanced insights into the model's behavior across various decision thresholds. This analysis facilitates the identification of an optimal threshold tailored to specific requirements.

Interpretation of ROC Curve and AUC Score:

The ROC curve and AUC score serve as valuable tools for assessing a model's capacity to distinguish between classes. In the context of the Random Forest model, a moderate AUC score suggests a reasonable level of performance, albeit with room for potential improvement. Consideration of model fine-tuning, adjustment of decision thresholds, or exploration of additional features may contribute to augmenting its discriminatory power and overall effectiveness in credit card fraud detection.



**AUC Scores:**

* Support Vector Machine (SVM): AUC = 0.50
* Logistic Regression: AUC = 0.45
* Random Forest: AUC = 0.60

**Interpretation of ROC Curves and AUC Scores:**

* Support Vector Machine (SVM):

An AUC of 0.50 indicates a model with no discriminatory power, comparable to random chance.

* Logistic Regression:

AUC = 0.45 suggests limited discriminatory performance. There is considerable room for improvement, and further analysis may reveal areas for model enhancement.

* Random Forest:

An AUC of 0.60 indicates fair discriminatory power. While the model exhibits some ability to distinguish between classes, optimization efforts may further enhance its effectiveness.

# Discussion & Conclusion

## Findings and Implications:

The analysis of credit card fraud patterns revealed notable findings. The Random Forest model exhibited a commendable accuracy of 98.1%, indicating robust overall predictive performance. However, due to the imbalanced dataset, precision, recall, and F1-score were scrutinized, revealing areas for potential improvement. Feature importance analysis underscored the significance of transaction amount, geographical coordinates, and merchant information in predicting credit card fraud. These findings hold critical implications for fraud detection strategies. While the high accuracy is promising, a closer examination of precision and recall is warranted. The imbalanced nature of the dataset necessitates further model refinement to enhance its ability to correctly identify fraudulent transactions without compromising precision. Insights into feature importance, particularly regarding transaction characteristics, offer valuable information for financial institutions aiming to develop targeted fraud detection strategies. Ethical considerations played a pivotal role in the analysis. The pre-processing steps and ethical guidelines applied to handle the dataset exemplify responsible and privacy-compliant practices. This ethical foundation ensures the model's development aligns with industry standards and privacy regulations.

Building upon the analysis of credit card fraud patterns, the feature importance results shed further light on the intricacies of the Random Forest model's decision-making process. The detailed examination of key features provides additional insights that complement the overall findings.

Random Forest Classifier achieved a high accuracy of XX% in detecting fraudulent transactions. It outperformed SVM and Logistic Regression models in terms of precision, recall, and F1-score.

The most influential features for fraud detection were identified: [list of features]. These features provide critical insights into fraudulent transactions and can aid in developing more robust fraud detection systems.

**Feature Importance Analysis**:

The feature importance analysis highlighted specific features that significantly influenced the Random Forest model's predictions. The top contributors include:

* Transaction Amount (4.34%): The importance of transaction amount was reaffirmed, emphasizing its role as a critical factor in the detection of fraudulent transactions.
* Geographical Coordinates and Merchant Information (45.73%): Geographical coordinates, including merchant latitude and longitude, along with merchant-specific information, emerged as predominant contributors, collectively accounting for a substantial portion of the model's decision-making.
* Enhanced Understanding of Model Performance: The commendable accuracy of the Random Forest model (98.1%) remains a robust indicator of its overall predictive performance. However, the detailed breakdown of precision, recall, and F1-score adds nuance to the evaluation, providing a more granular understanding of the model's strengths and potential areas for refinement.

**These findings have several implications for credit card fraud detection:**

Geographical Features are Crucial: The dominance of merchant latitude and longitude underscores the importance of geographical information in identifying potentially fraudulent transactions. Financial institutions may need to enhance their fraud detection algorithms by placing more emphasis on location-based features.

* Temporal Patterns Matter: The day of the week and hour of the transaction contribute significantly to fraud predictions. Recognizing temporal patterns can help financial institutions implement time-sensitive fraud prevention measures.
* Transaction Distance is Relevant: The calculated distance between the transaction and merchant locations is a noteworthy factor. Transactions involving significant distances may warrant additional scrutiny.
* Amount Alone is Insufficient: While transaction amount is a relevant feature, its impact is comparatively lower than geographical and temporal features. Financial institutions should avoid relying solely on transaction amount for fraud detection.

## Recommendations

Continuous monitoring and refinement are recommended strategies. Regularly assessing the model's performance and refining it based on ongoing analysis and feedback will aid in adapting to emerging fraud patterns. Collaboration with industry experts, financial institutions, and policymakers is crucial. By engaging in discussions and partnerships, the analysis can contribute to enhanced fraud detection capabilities and a proactive stance against potential threats. Considering advanced techniques for handling imbalanced datasets is advised. Experimenting with oversampling, under sampling, and ensemble methods could further improve model performance. Exploring different machine learning algorithms beyond Random Forest may provide valuable insights and comparisons. Based on the analysis, the following recommendations are suggested:

Further investigate the relationship between identified influential features and fraud to gain deeper insights and improve fraud detection accuracy.

Explore advanced techniques for handling imbalanced datasets, such as oversampling, undersampling, or using ensemble methods.

Experiment with different machine learning algorithms to compare performance and identify the most suitable model for credit card fraud detection.

**Specific Recommendations Based on Findings:**

Building upon the feature importance analysis and nuanced understanding of the Random Forest model's performance, we recommend specific actions:

* Investigate High-Risk Geographical Regions: Conduct further investigations into the specific patterns associated with high-risk geographical regions. Understanding the dynamics of these regions can lead to targeted interventions and improved fraud detection strategies.
* Enhance Temporal Aspect of Fraud Detection: Explore additional temporal features, such as time intervals between consecutive transactions. Enhancing the temporal aspect of fraud detection can provide a more comprehensive understanding of transaction patterns over time.
* Capture Relationships with Interaction Features: Consider the incorporation of interaction features that capture relationships between geographical and temporal features. These features could unveil hidden patterns and contribute to a more robust and nuanced fraud detection model.

## Impact of Work

The impact of this analysis extends to consumer and financial protection. By significantly enhancing fraud detection capabilities, the work protects consumers and financial institutions from potential losses. The insights gained from the analysis can contribute to the development of more robust and accurate fraud detection systems. The work also advances the financial industry's security measures. Addressing the challenges of credit card fraud through machine learning contributes to the development of industry-wide best practices in fraud detection and prevention. The ethical use of data underscores the importance of responsible data practices and aligns with the industry's commitment to privacy regulations. Implementing effective credit card fraud detection systems can have a significant impact on financial institutions and cardholders. By accurately identifying fraudulent transactions, financial losses can be minimized, customer trust can be strengthened, and the overall security of the financial system can be enhanced.

## Conclusion

In conclusion, our comprehensive analysis of credit card fraud detection using machine learning models has yielded valuable insights and recommendations for continuous improvement. The commendable accuracy of the Random Forest model, coupled with a detailed breakdown of precision, recall, and feature importance, provided a nuanced understanding of its strengths and areas for refinement.

The feature importance analysis underscored the critical role of transaction amount, geographical coordinates, and merchant information in predicting credit card fraud. These findings offer financial institutions key information for the development of targeted fraud detection strategies, emphasizing the importance of a multi-faceted approach that considers both transaction characteristics and contextual information.

Ethical considerations played a pivotal role in our analysis, with responsible data practices and privacy-compliant measures ensuring the alignment of our model development with industry standards. The ethical foundation laid throughout the process fosters trust in our fraud detection strategies.

Looking ahead, we recommend continuous monitoring and refinement strategies, collaboration with industry experts, financial institutions, and policymakers, and the exploration of advanced techniques for handling imbalanced datasets. Our specific recommendations to investigate high-risk geographical regions, enhance the temporal aspect of fraud detection, and incorporate interaction features aim to guide further research and model development.

In the ever-evolving landscape of financial fraud, these findings position our approach as adaptive and responsive. Collaborative partnerships and knowledge sharing will be crucial in staying ahead of emerging threats and maintaining a proactive stance against potential risks. Through a commitment to ongoing analysis, refinement, and collaboration, our credit card fraud detection model can contribute to the collective efforts aimed at securing financial transactions and protecting the interests of both financial institutions and cardholders. Credit card fraud detection is an important area of research and application in the financial industry. Machine learning models, such as Random Forest Classifier, SVM, and Logistic Regression, show promise in accurately detecting fraudulent transactions. By understanding the dataset, exploring feature importance, and evaluating model performance, we can develop robust fraud detection systems that benefit both financial institutions and cardholders. Further research and analysis are recommended to improve fraud detection accuracy and explore additional techniques for handling imbalanced datasets.

**Group Project Report**

Title: Predictive Analysis of Credit Card Fraud

**Introduction**: The group collectively decided on the project area of interest: credit card fraud. This section outlines the collaborative effort in choosing the project area and developing a set of research questions.

* **Project Area and Research Questions (All)**: We took the lead in developing a comprehensive set of research questions. The team collaboratively worked on ensuring the questions were well-structured and included predictive aspects. The chosen project area and research questions were agreed upon through group discussions.
* **Dataset Selection (All)**: All members participated in researching and identifying a suitable dataset related to credit card fraud. Ethical considerations were taken into account, and the dataset selection likely involved searching and selecting on Google Colab or another platform.
* **Data Exploration and Overview (Rita)**: Rita provided an overview of the selected dataset within the Colab Notebook. The section discussed the source, structure, and ethical considerations related to the data. Data exploration and overview were performed using Google Colab.
* **Methodology (Rita)**: Rita developed a comprehensive methodology for answering research questions. The methodology outlined steps in data preprocessing, feature engineering, and model selection. This involved coding and documentation in Google Colab.
* **Data Preprocessing and Feature Engineering (Cicero)**: Cicero was responsible for preprocessing the selected dataset, handling missing values, and outliers. Additionally, Cicero conducted necessary feature engineering to enhance model performance.
* **Machine Learning Models (Rita)**: Rita implemented at least three different machine learning models relevant to the dataset. The choice of models and parameters was documented in the Python notebooks.
* **Model Evaluation (Eunji)**: Eunji evaluated the performance of each implemented machine learning model. She used appropriate metrics and provided an explanation for the chosen evaluation criteria.
* **Results and Analysis (Eunji)**: Eunji presented the results of applying machine learning models to the dataset. The section discussed how well the models addressed the research questions.
* **Discussion & Conclusion (Rita,Eunji)**: Rita and Eunji summarized the findings and implications of the analysis. Recommendations based on the results were made, and the impact of the work was discussed.
* **Report Writing (All & Cicero)**: Cicero created the form and constructure, all members collaborated on writing the report, covering sections such as Introduction, Dataset Overview, Methodology, Analysis, and Discussion & Conclusion.
* **Presentation (All & Cicero)**: Cicero created the form and constructure, all group members collectively prepared a 5-minute presentation covering individual contributions and key findings. Each member spoke about their specific tasks and findings.
* **Appendix (Eunji)**: Eunji compiled notes from group meetings, ensuring each group member provided a note specifying their contributions and estimated time of completion.

**Conclusion:**

The collaborative efforts of the group resulted in a thorough analysis of credit card fraud using machine learning models. Each member contributed significantly to their assigned tasks, leading to a comprehensive and well-documented project.