

Report on

Project= “**Predicting Customer Churn in a Telecommunications Company**”

(as a part of assignment)

For the role of

**“SDE(Data Science)”**

**Submitted** **to**: **Submitted By:**

Subjected Company= **SpeakX**  Name=Amandeep Singh

Registration No.=12101332

**Objective:**

The primary objective of this project is to develop a predictive model that can identify customers at risk of churning, enabling the company to take proactive measures to retain them. Churn refers to the phenomenon where customers discontinue using a service or product. In this context, churn specifically refers to customers who are likely to discontinue their subscription or service with the company.

**Key Components:**

1. **Data Collection and Preparation:** Gather relevant data on customer demographics, usage patterns, and past behaviors. Clean and preprocess the data to ensure its quality and suitability for modeling.
2. **Feature Selection and Engineering:** Identify relevant features (independent variables) that are predictive of churn. This may include demographic information, service usage history, contract details, etc. Create new features if necessary to improve predictive performance.
3. **Model Development:** Train and evaluate predictive models using historical data. Explore various machine learning algorithms such as logistic regression, decision trees, random forests, or gradient boosting techniques. Select the best-performing model based on evaluation metrics like accuracy, precision, recall, and F1 score.

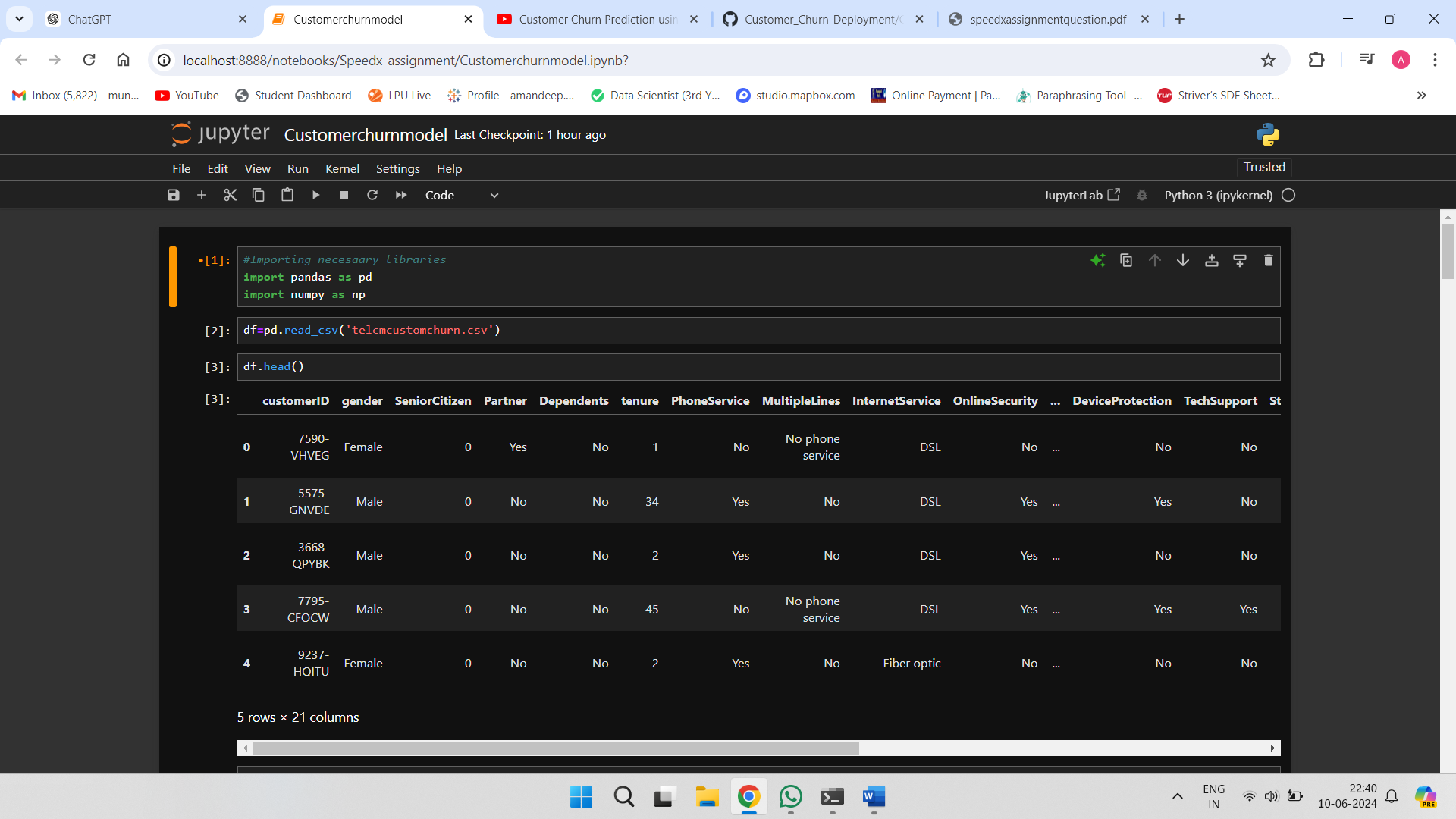
**Challenges faced during the project:**

* Data Quality: Poor data quality, such as missing values, inconsistencies, or inaccuracies, can hinder model performance. Cleaning and preprocessing the data effectively is crucial but can be time-consuming, especially if the dataset is large and diverse.
* Imbalanced Data: Churn prediction datasets often suffer from class imbalance, where the number of churn instances is much smaller than non-churn instances. Imbalanced data can lead to biased models that prioritize accuracy on the majority class. Techniques such as oversampling, undersampling, or using appropriate evaluation metrics are necessary to address this challenge.
* Feature Selection: Identifying relevant features that contribute to churn prediction is essential for model effectiveness. However, selecting the right features from a large pool of potential variables can be challenging. It requires domain expertise and experimentation to determine which features are most informative.
* Model Interpretability: While complex models like Random Forest can provide accurate predictions, they may lack interpretability, making it difficult to understand the reasoning behind the predictions. Interpretable models such as decision trees or simpler ensemble methods may be preferred in some cases, even if they sacrifice a bit of predictive accuracy.
* Model Evaluation: Evaluating the performance of churn prediction models is challenging, especially when dealing with imbalanced data. Traditional metrics like accuracy may not provide a complete picture of model performance. Metrics such as precision, recall, F1-score, and ROC AUC are more suitable for assessing performance in imbalanced datasets.

**Data Collection and Preprocessing:**

Dataset used for the assignment: <https://www.kaggle.com/datasets/blastchar/telcocustomer-churn>

Know your dataset:

Employing very functions such as describe(),head(),shape(),column(),dype(),isnull().sum() we acknowledged various properties of dataset to find missing values if any . This will make the dataset ready for analysis.**A screenshot of a computer

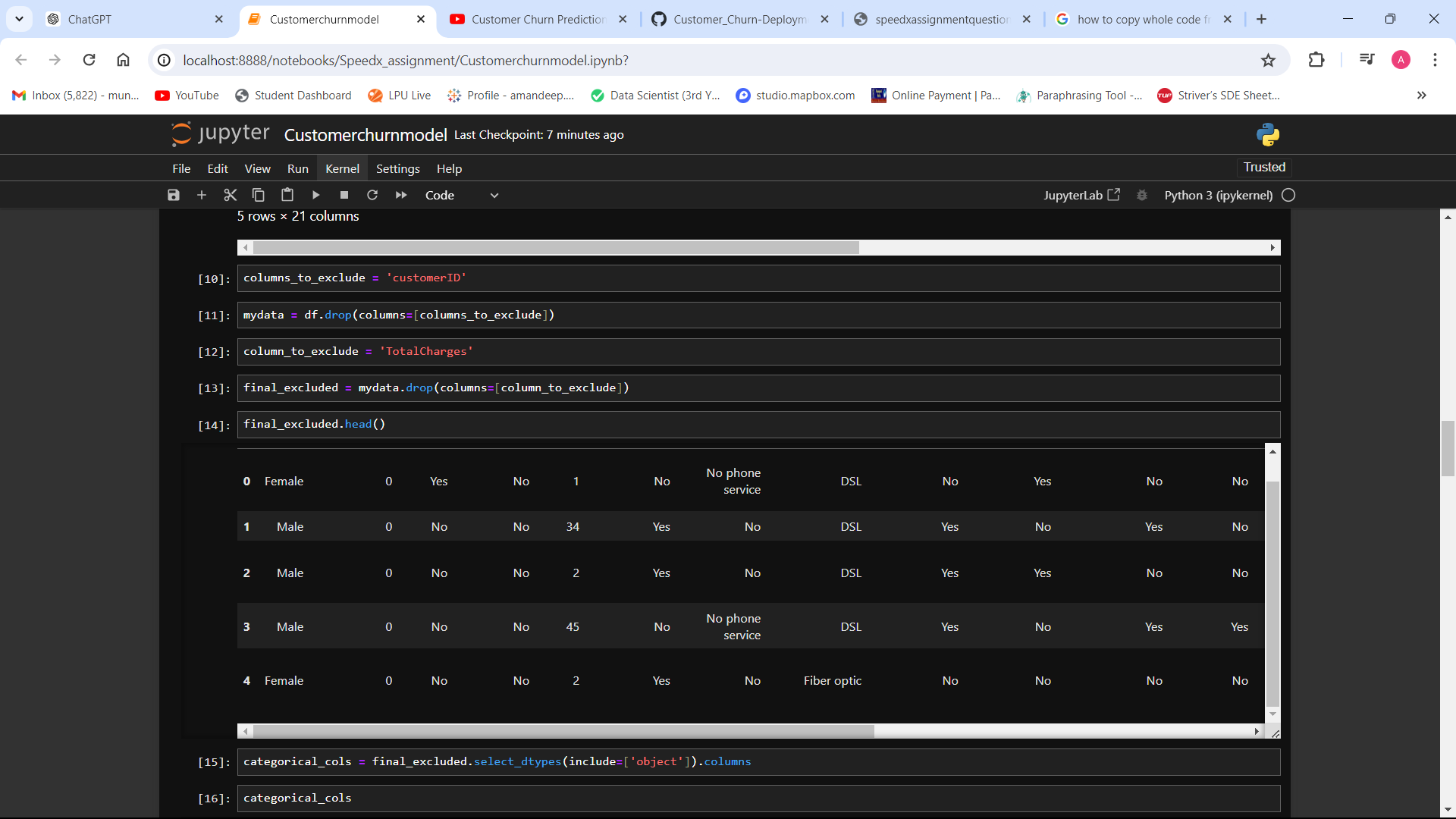
Description automatically generatedA screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

Excluding unnecessary columns:



Categorical columns:

Using the Lable Encoder , we successfully translated the string column values(object) to numeric to make the dataset ready for statistical analysis and model development.

A screenshot of a computer

Description automatically generated

**Exploratory Data Analysis (EDA):**

Performing EDA on the dataset to understand customer behaviour and factors influencing churn. EDA plays a crucial role in feature selection and engineering by identifying the most relevant variables and creating new features that may enhance predictive power. It also helps in formulating hypotheses and guiding subsequent analyses, leading to more informed decision-making and actionable insights.

Overall, EDA is not just a preliminary step but an ongoing and iterative process throughout the data analysis lifecycle, helping to uncover hidden patterns, validate assumptions, and derive meaningful conclusions from the data.

By applying necessary libraries like matplotlib and pandas we created a heatmap plot to visualize and understand the data properly.

A screenshot of a computer

Description automatically generated

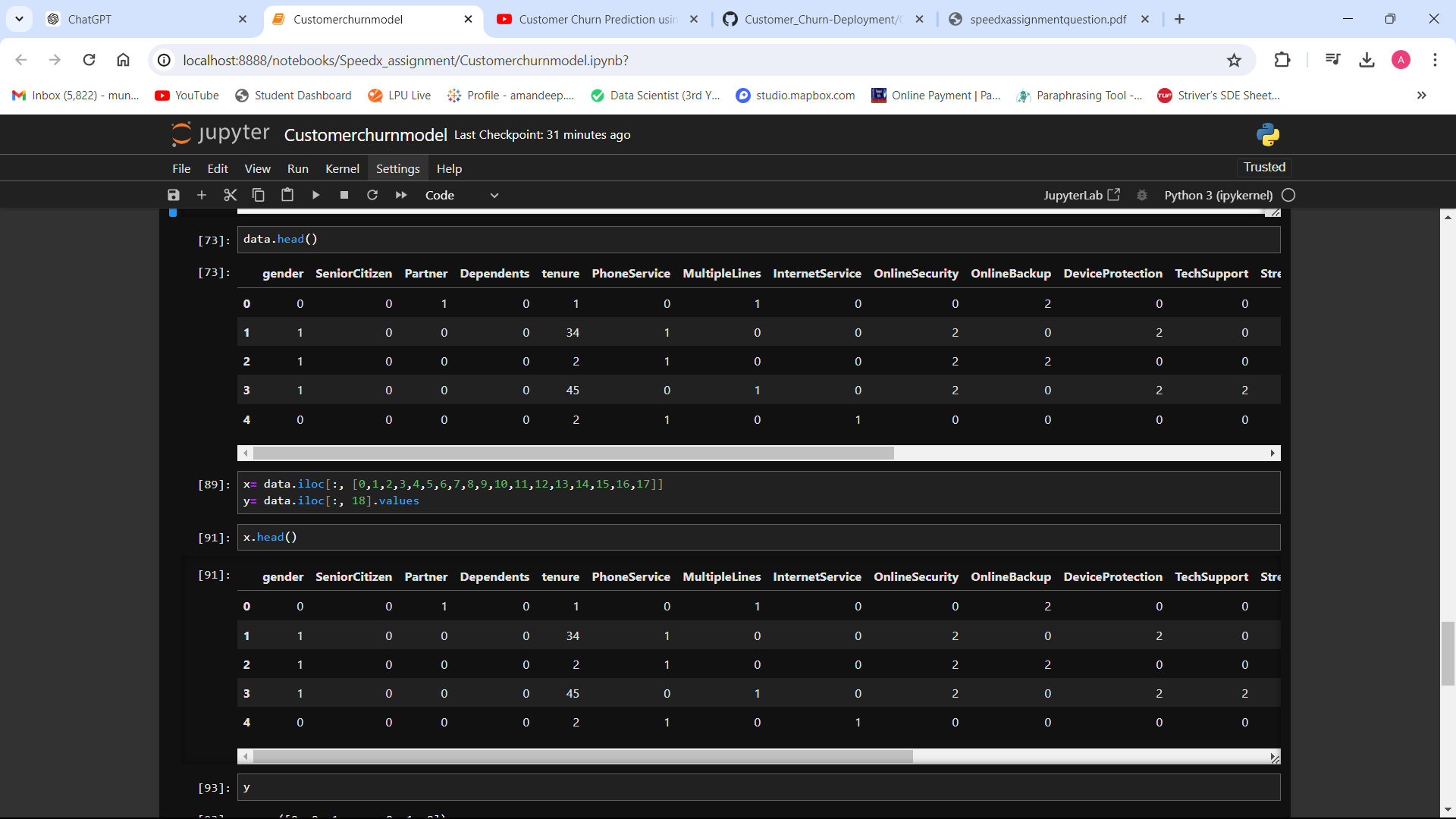
A screen shot of a graph

Description automatically generated

**Building the Churn Prediction Model:**

A churn prediction model is a tool used by businesses to anticipate when customers are likely to stop using their services or products. By analyzing various factors like customer behaviour, interactions, and demographics, the model forecasts the probability of churn.

Separating the dependent and independent variables by defining them in separate x and y lists.



Building the Prediction Model : **Random Forest Classifier**

**Significance of the model used :**

Random Forest Classifier is particularly well-suited for churn prediction tasks due to its specific advantages:

1. **Handling Imbalanced Data**: Churn prediction datasets often suffer from class imbalance, where the number of churn instances is much smaller than the number of non-churn instances. Random Forest inherently handles class imbalance well by constructing decision trees based on bootstrapped samples of the data and combining their predictions through voting.
2. **Non-linear Relationships**: Churn prediction involves understanding complex interactions between various factors that contribute to customer attrition. Random Forest can effectively capture non-linear relationships between features and the target variable, allowing it to model the intricate nature of churn behavior.
3. **Feature Importance**: Identifying which factors contribute most to churn is critical for developing effective retention strategies. Random Forest provides a measure of feature importance, allowing businesses to prioritize resources on addressing the most influential factors contributing to churn.
4. **Robustness to Noise**: Churn prediction datasets may contain noisy or irrelevant features. Random Forest is robust to noisy data and can handle irrelevant features by selecting informative ones during the tree-building process, thus improving the model's performance.
5. **Scalability**: Random Forest is parallelizable, making it suitable for processing large volumes of data commonly encountered in churn prediction tasks. It can efficiently handle datasets with numerous features and instances, scaling well with computational resources.
6. **Interpretability**: While Random Forest is an ensemble of decision trees, it still offers a degree of interpretability by providing feature importance rankings and decision paths. Understanding which features contribute most to churn allows businesses to gain insights into customer behavior and tailor retention strategies accordingly.
7. **High Accuracy**: Random Forest generally yields high accuracy in classification tasks, including churn prediction. Its ability to reduce overfitting and generalize well to unseen data contributes to its effectiveness in accurately identifying potential churners.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Evaluation Results:**

Accuracy: It can be seen over here that the Random Forest Classifier model achieved the accuracy of about 78.6 percent which is enough to predict the customer churn based on different columns data values.

Similarly, the **precision score** achieved is: 0.6363636363636364.

Precision focuses on the correctness of positive predictions. It measures the proportion of true positive predictions (correctly identified instances of a particular class) among all positive predictions (instances predicted as belonging to that class).

**Recall value** is: 0.450402144772118

Recall, also known as sensitivity, measures the proportion of true positives predicted correctly out of all actual positive instances. It's crucial when the cost of false negatives is high.

**F1 score** achieved: 0.5274725274725275

The F1 score is the harmonic mean of precision and recall. It provides a balance between these two metrics, making it useful in situations where there's an uneven class distribution. F1 score reaches its best value at 1.