Telco Customer Churn Predictive Modeling

Introduction

With costs increasing at the fastest pace in decades, inflation has unsurprisingly risen to the top of the agenda for telecommunications executives. The effect on each Telco will vary, of course. But inflation will put pressure on most of the budget, particularly personnel costs, energy, external spending on services, leases, and capital expenditures.

Therefore, it’s a no-brainer that customer retention is one of the most critical goals for businesses. Increasing customer retention rates by a mere 5% could increase profits by 25% to 95%. Many Telco will continue to focus on delivering a great customer experience which will make it easier for customers to accept price increases and not switch to competitors.



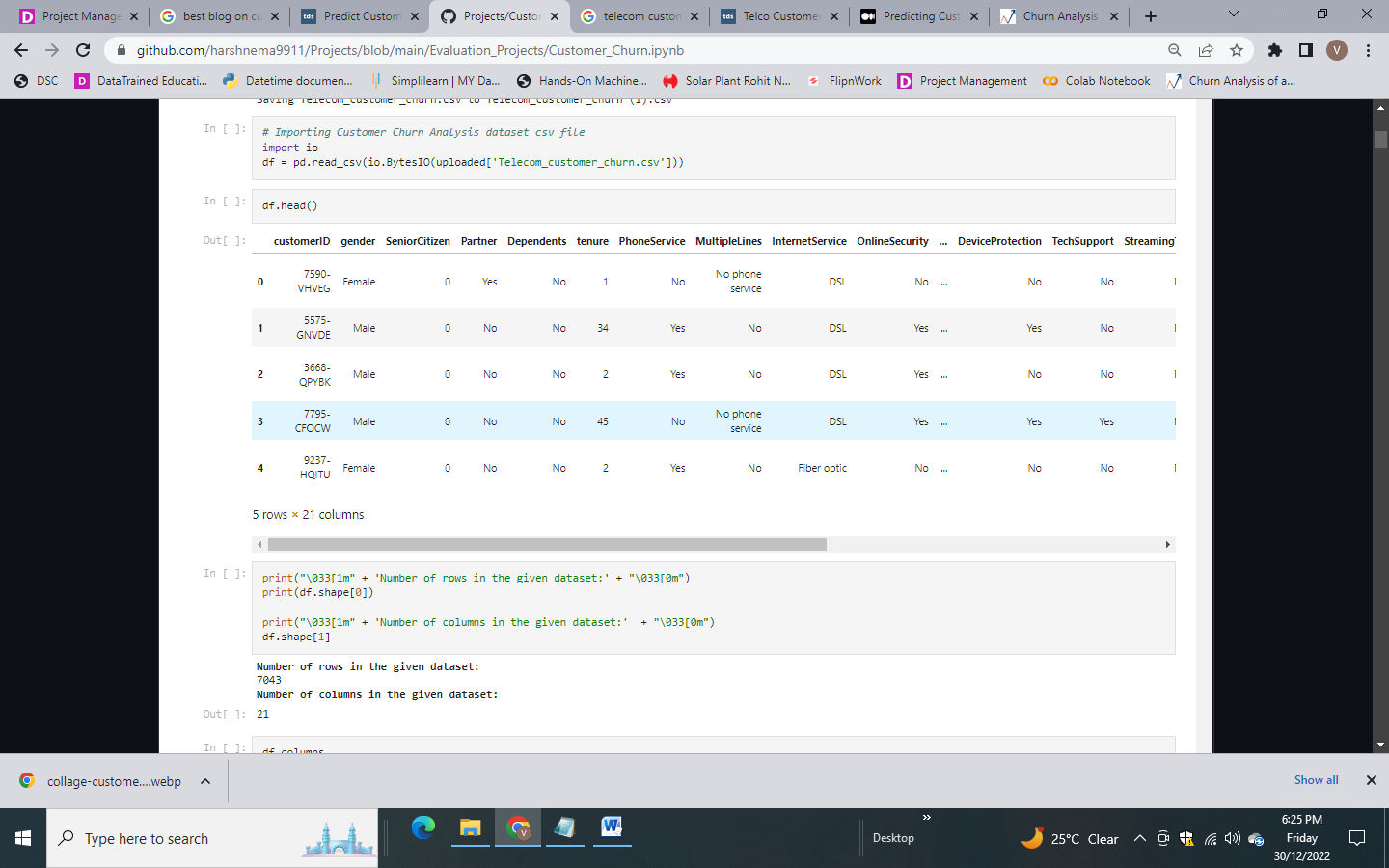
**What is attrition?**

**Churn** — or **customer attrition** — is simply defined as the number or percentage of customers lost within a specific period of time, netting off the new acquisitions during that time. In other words, the metric tracks how successful or not you have been at keeping your customers engaged.

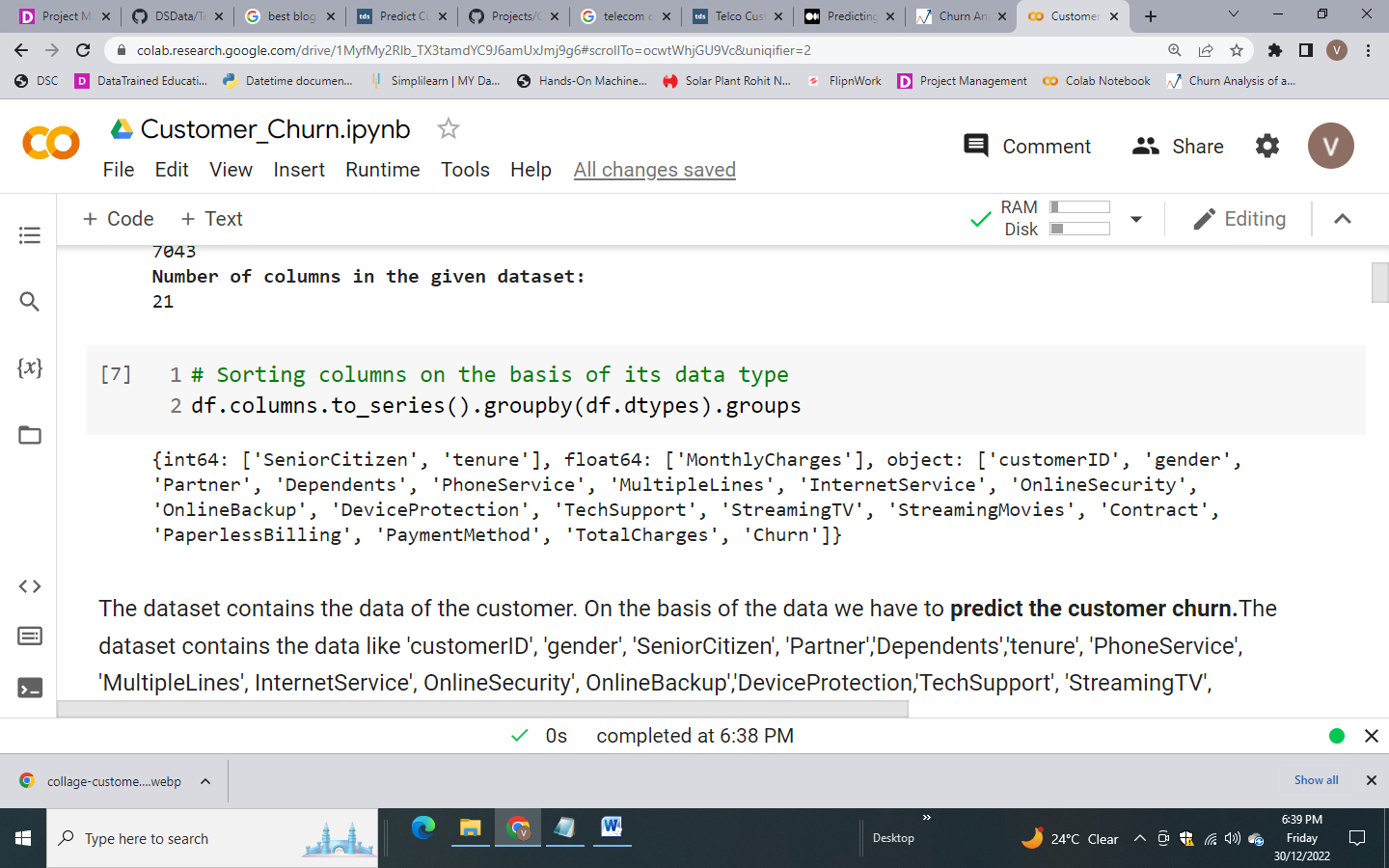
For a telecom company you may identify a customer as ‘attrited’ is he/she hasn’t used your network for a month. Most companies today, design specific initiatives to address attrition. The heart of churn management lies in being able to identify the early warning signals from potential attritors. If I know early enough that a specific customer is likely to leave my business, I can take proactive steps to prevent it from happening. And this is where analytics can play a transformational role.

# Problem Definition

Studying historical data around negative customer experiences and how customers of different types have responded to them can help develop a robust model for predicting reactive churn (Customers react to specific negative experiences that trigger their move away from your business is known as ‘Reactive’ churn). In this blog, we will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models. The company provides home and internet services to **7043 customers** in California. Our challenge is to help the company predict behavior to retain customers and analyze all relevant customer data to develop focused customer retention programs.

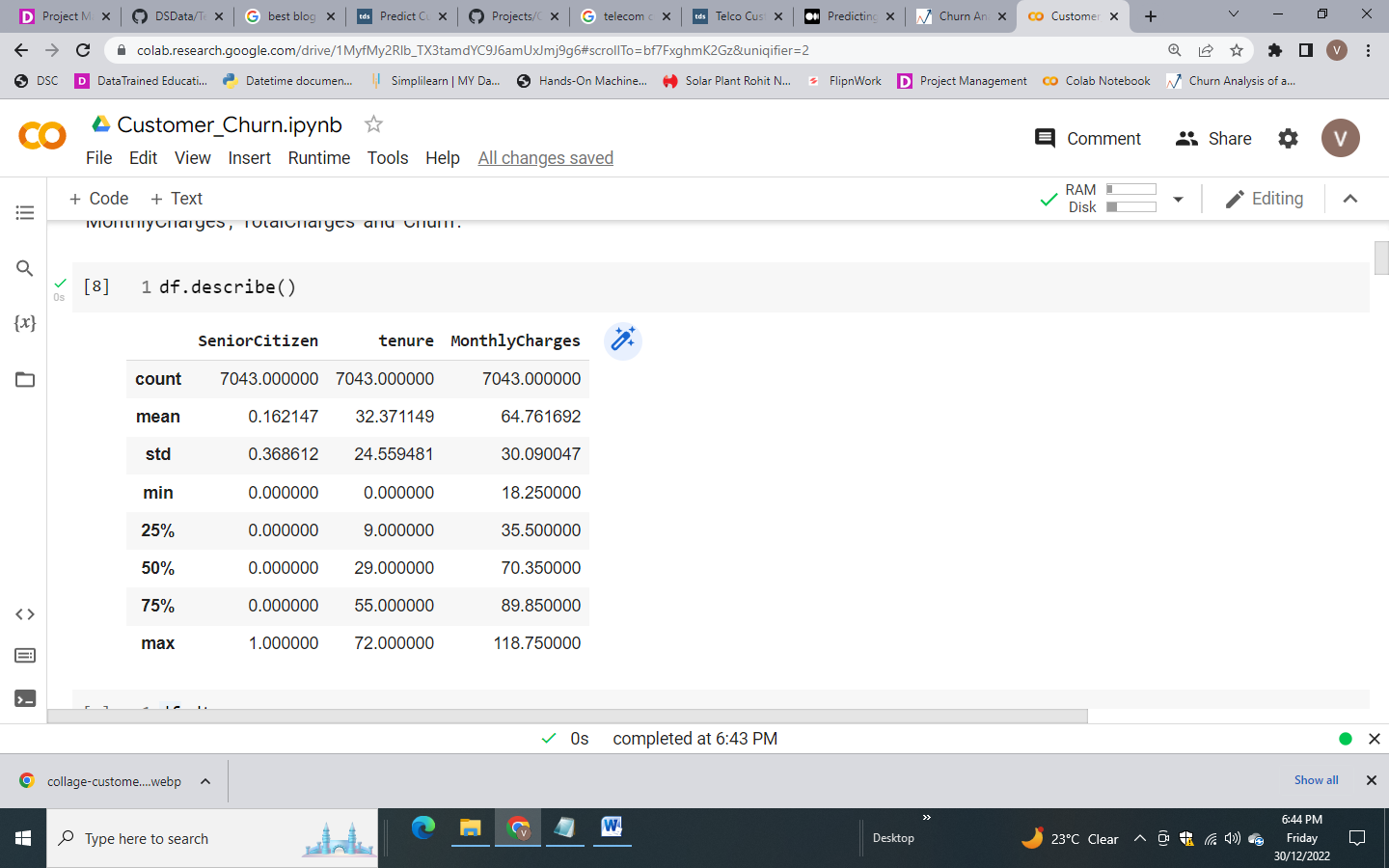


Now, let’s have a look at the **21 columns** present in the dataset:



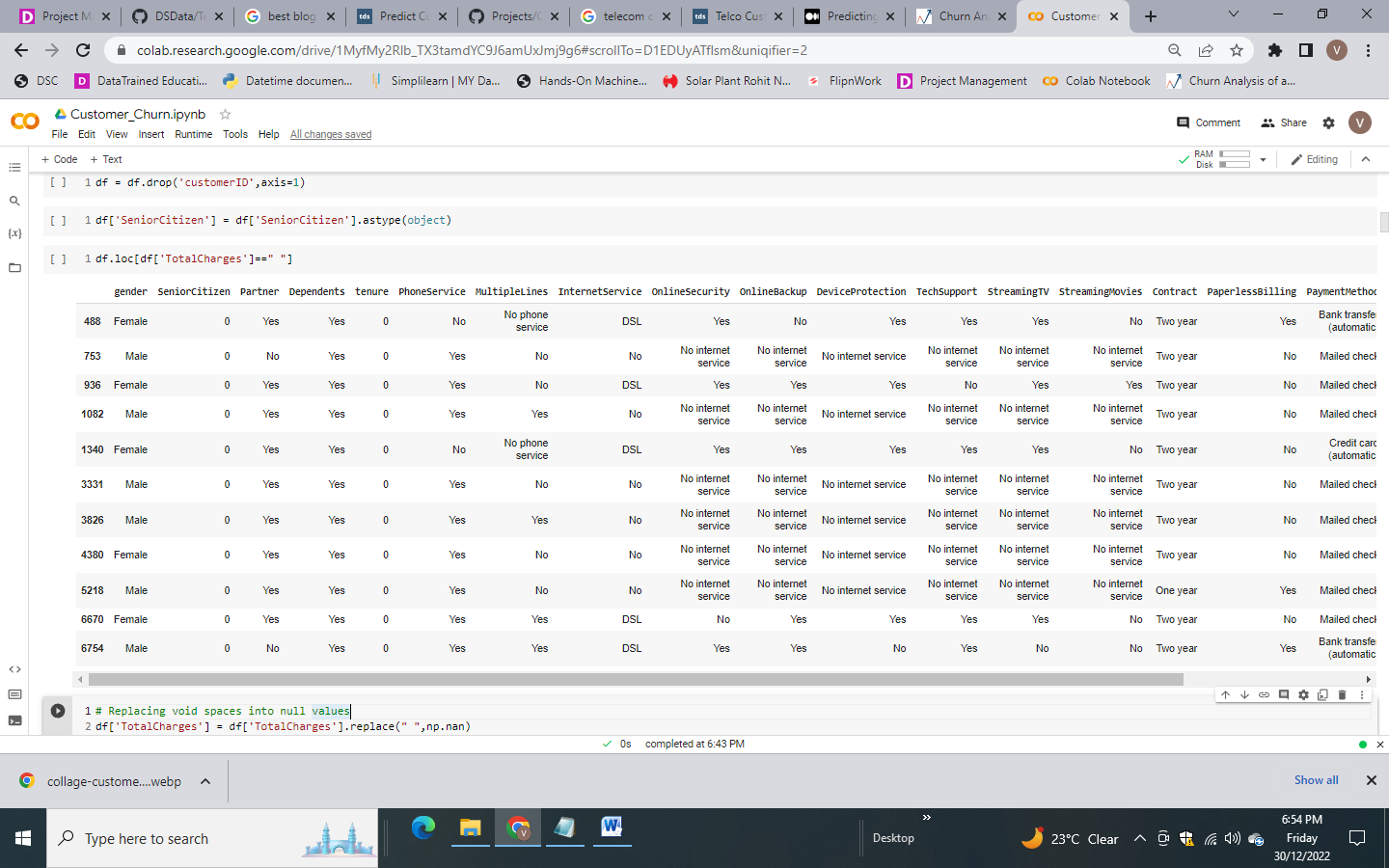
## Data Analysis

Checking statistics part of numerical data:

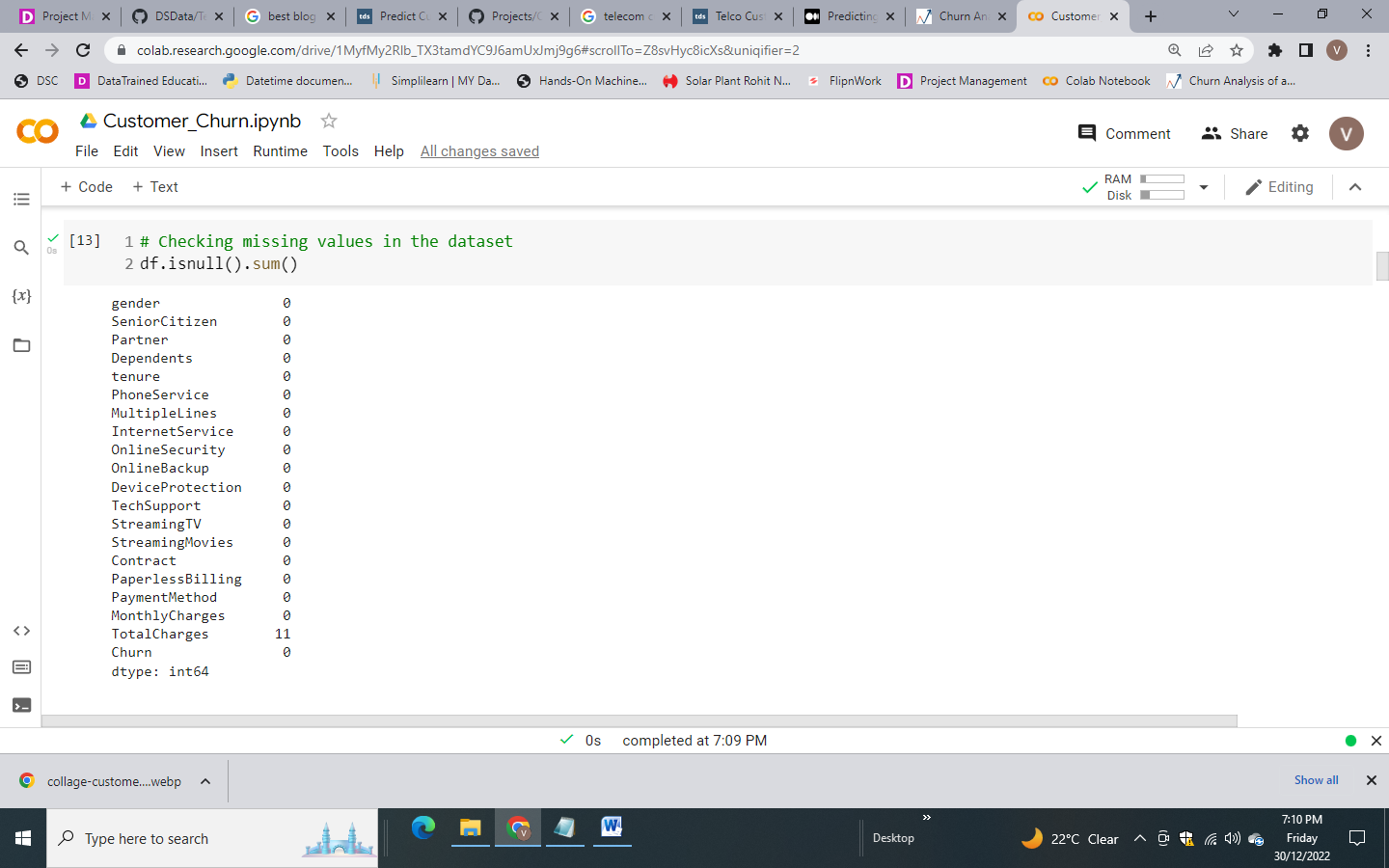


**Observation:**

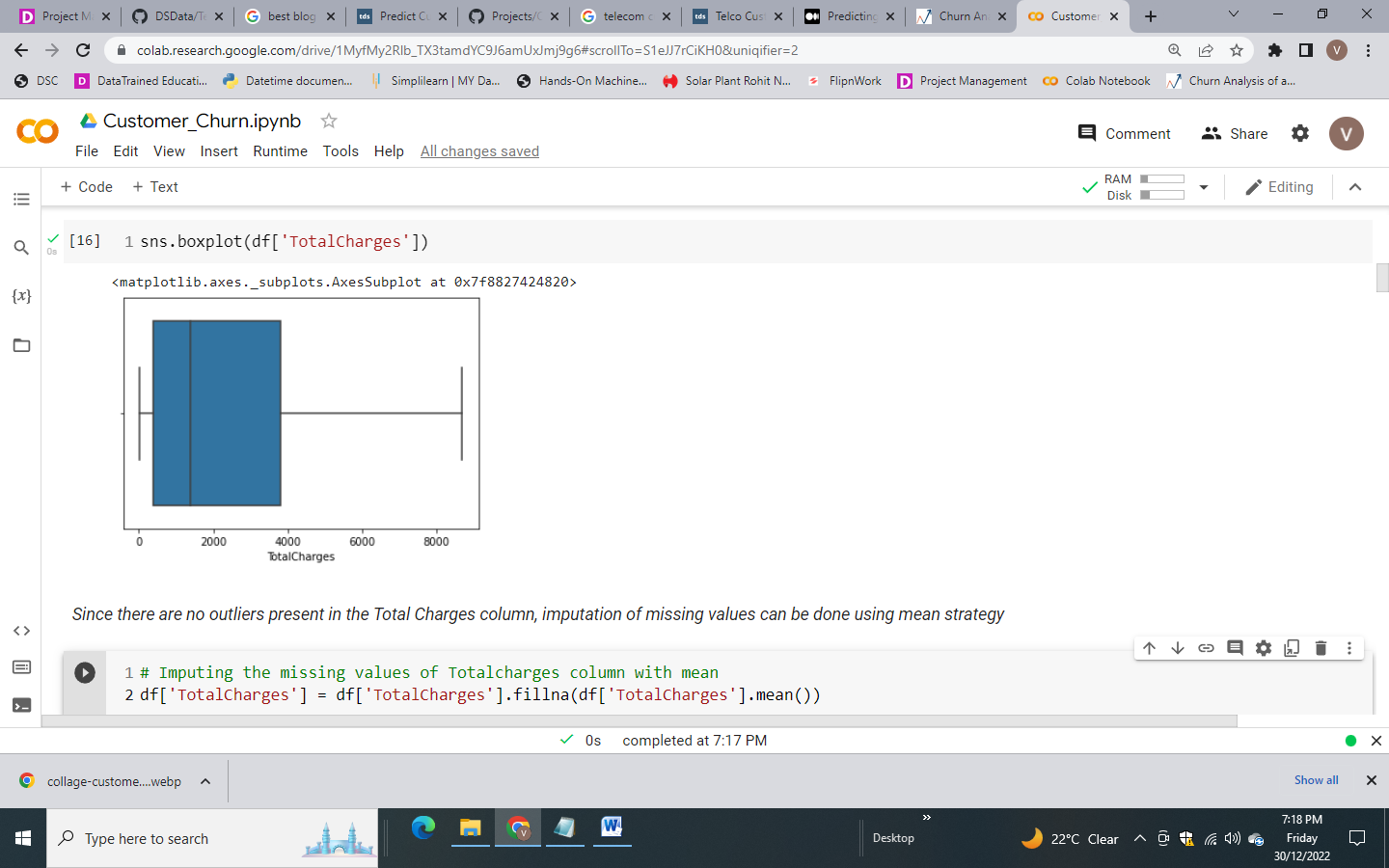
* Senior Citizen is a categorical variable but it is listed as int64,need to change its data type to object
* Total Charges feature is numerical in nature but categories as object data types. This implies that there **is presence of string variable** in this column or might be data error
* Customer ID is unnecessary variable from our analytical & modeling viewpoint and hence will drop 'Customer ID' column

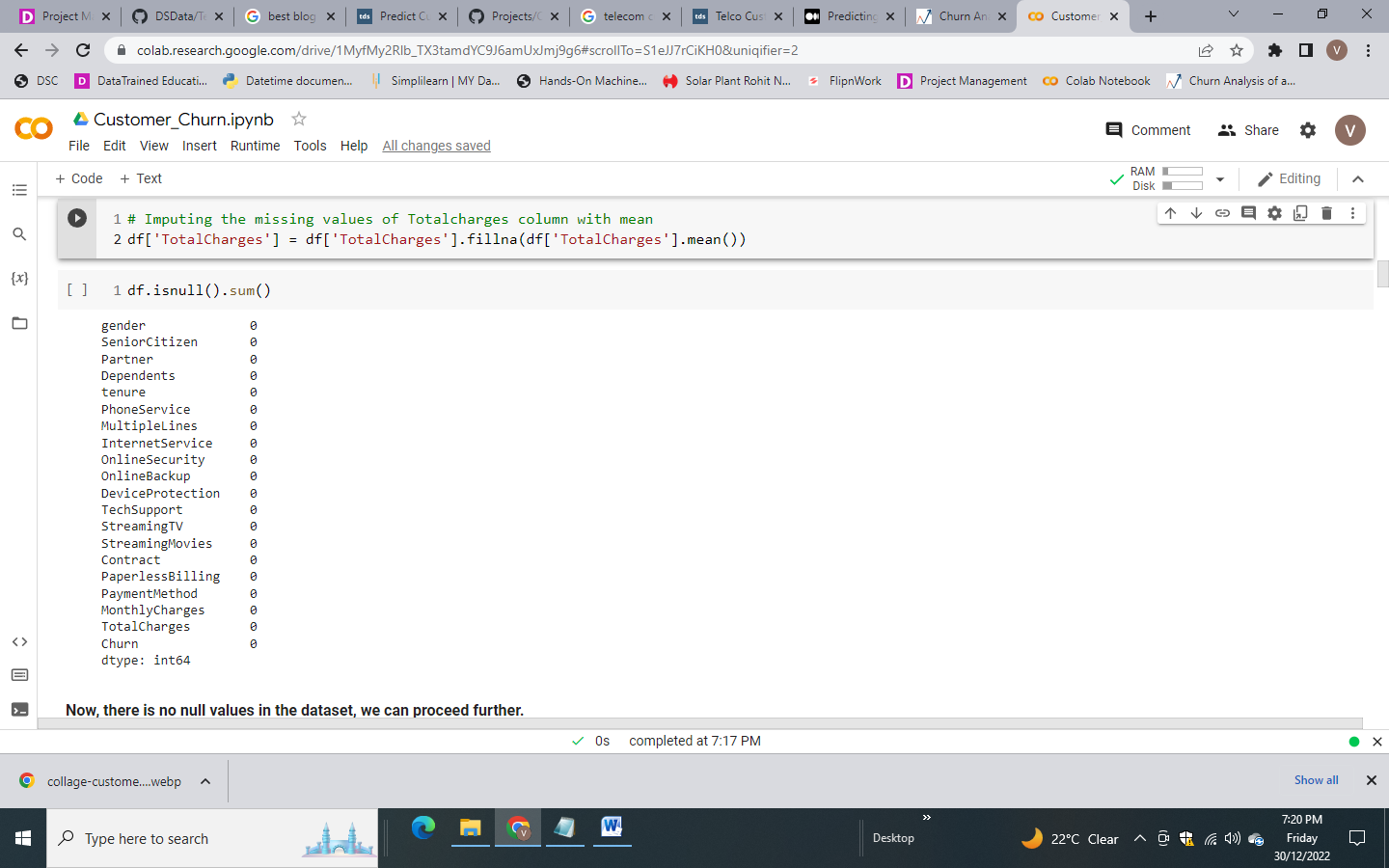


* Here we can see that there are 11 blank spaces in the Total charges column
* Converting blank spaces into null values

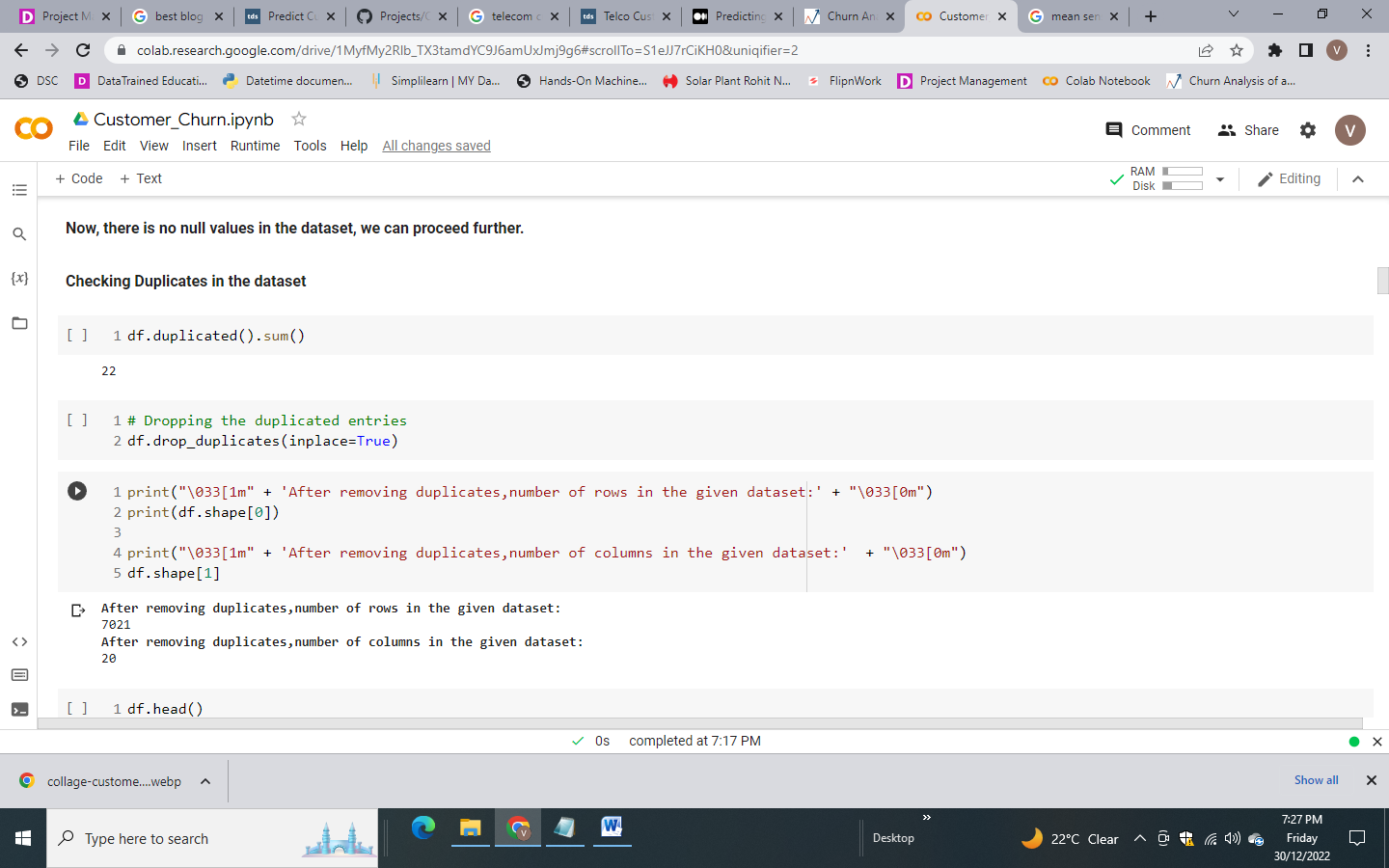


* Now we need to impute these null values in the **Total Charges** column
* Converting Total Charges column from object to float data type
* Checking outliers in Total Charges column, as mean is more sensitive to the existence of outliers than the median





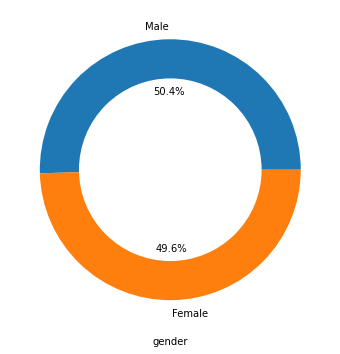
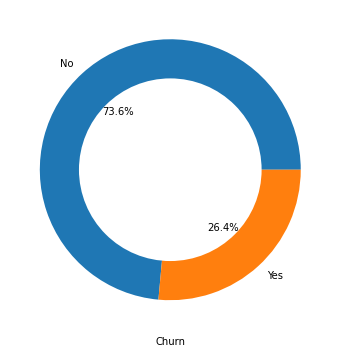
Checking duplicated data present in the dataset:

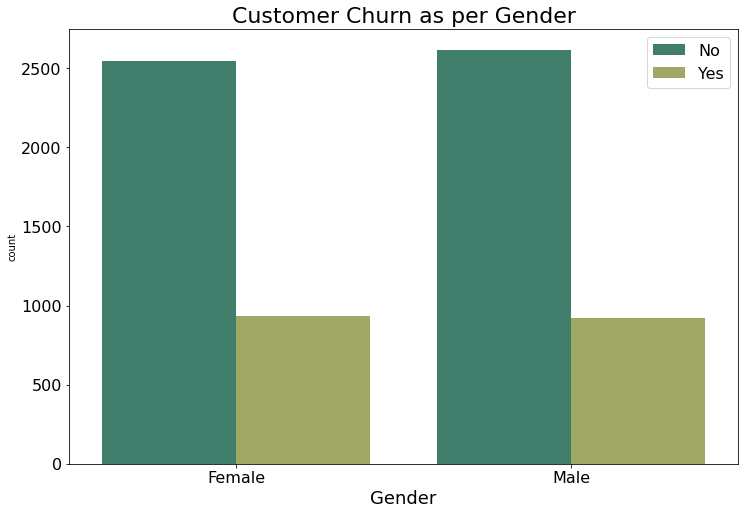


* Removing the duplicate data from the dataset and making it clean for further investigation
* The shape of the dataset after removing duplicates becomes **7021 rows and 20 columns**

### Exploratory Data Analysis

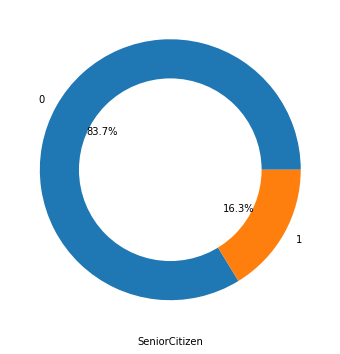
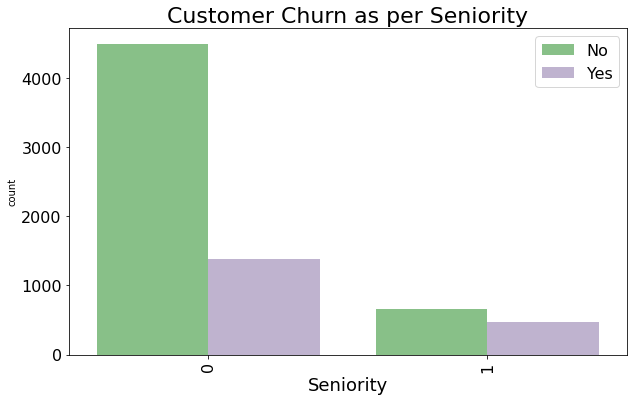
1. **Customer Churn as per Gender**



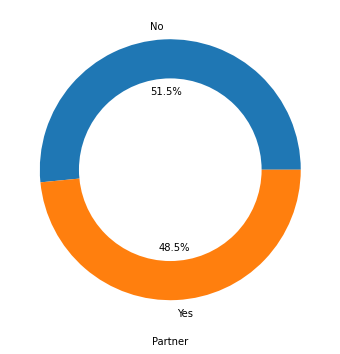
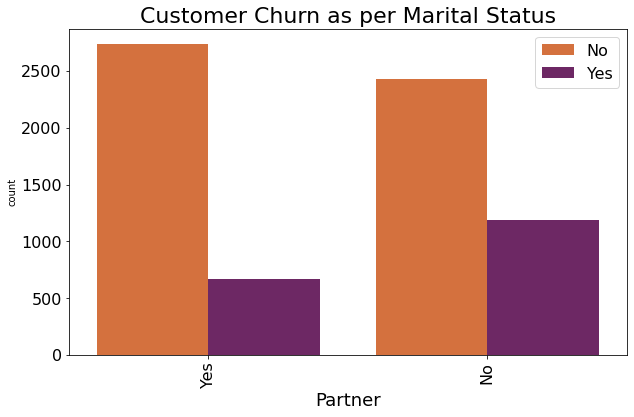
* From the above plots, equal number of male and female looks dissatisfied with the telecom company

1. **Customer Churn as per Seniority**

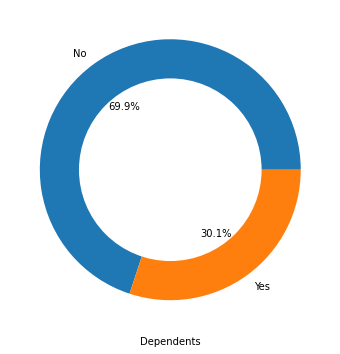
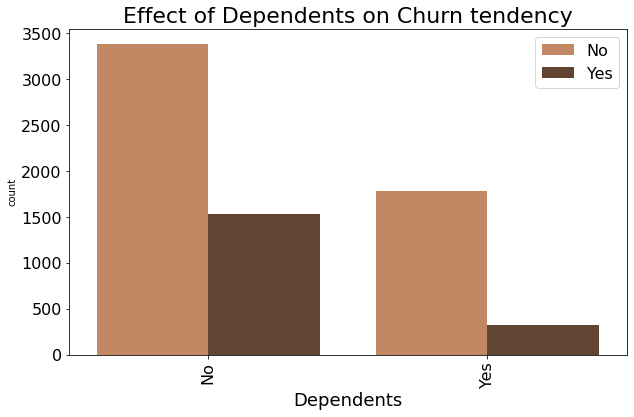
* More number of young people are dissatisfied and switching one telecom operator to other but it is obvious as almost 84% of younger people were present in the observation
* Non-Senior Citizens are high Churners

1. **Customer Churn as per Marital Status**

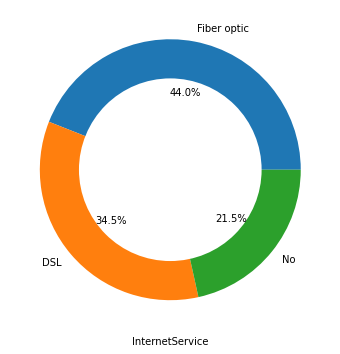
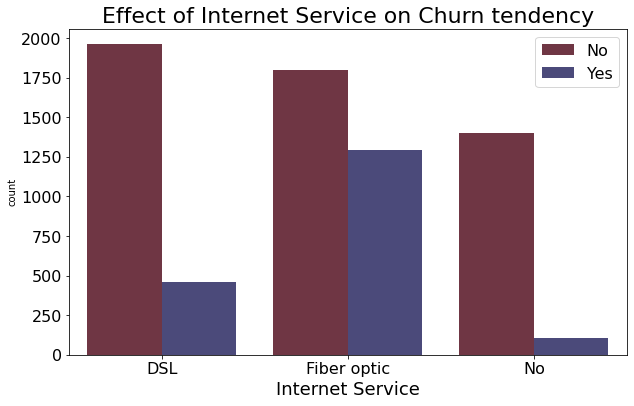
* There are 51.5% of married customer present in the dataset
* Bachelors are more reluctant to churn

1. **Effect of dependents on Churn tendency**

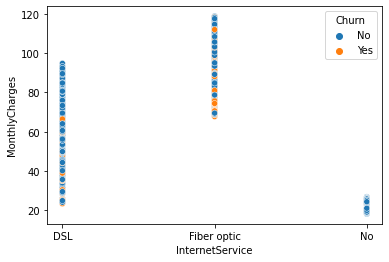
 

* Number of dependent customers is 2110 and number of independent customer is 4911
* Customer having depedents have less tendency to Churn

1. **Effect of Internet Service on Churn tendency**

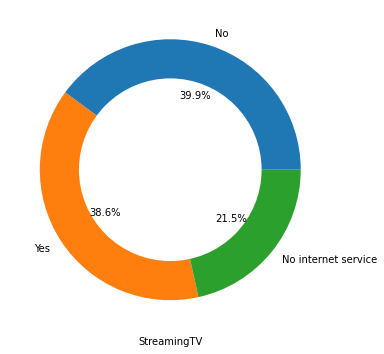
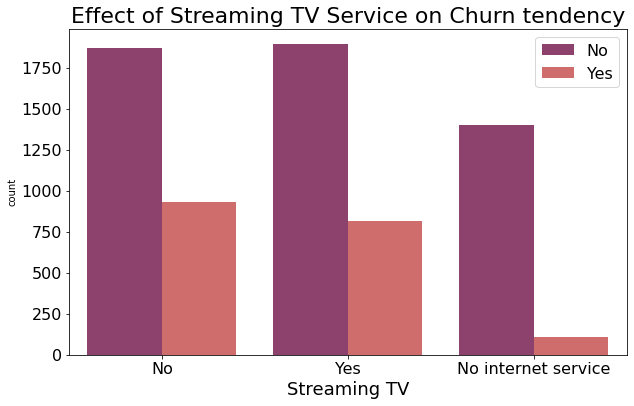
 

* Number of customers using Internet service is 5509 and customers with no internet service is 1512
* Out of 5509 internet users, 56% of customers uses Fiber optic as an Internet service
* 44% of Internet users opt for DSL service
* Customers using Fiber optic Internet service have more tendency to churn when compared with DSL and No Internet service



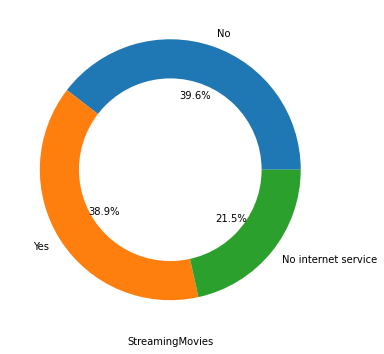
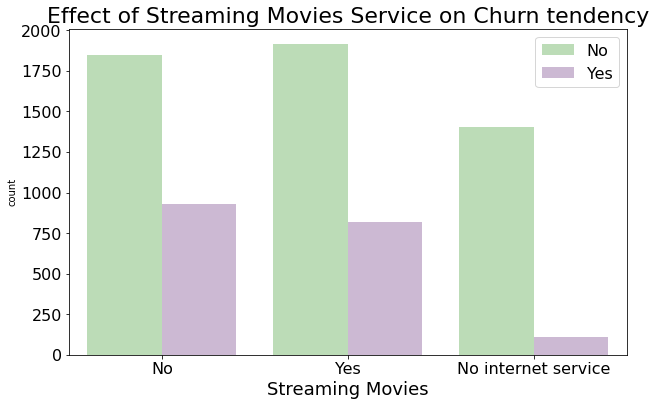
* High monthly charges can be seen among customers using fiber optic compare to DSL. So, high charges might be the reason of customer churn

1. **Effect of Streaming TV Service on Churn tendency**

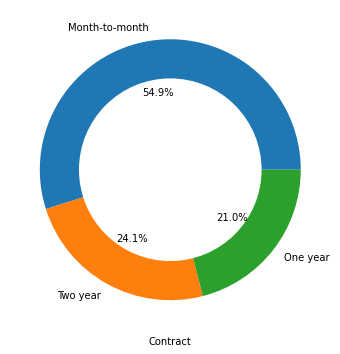
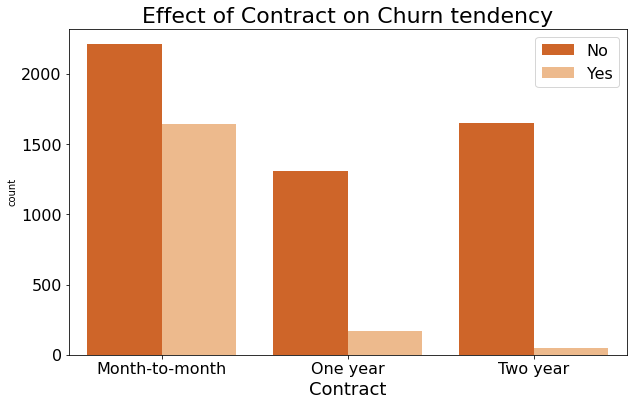
* Number of customers who don't use Internet service is 1512 and customers using Streaming TV service is 2707 and who don't use Streaming TV service is 2802
* Equal amount of churning can be seen for both Streaming TV users and non-users

1. **Effect of Streaming Movies on Churn tendency**

** **

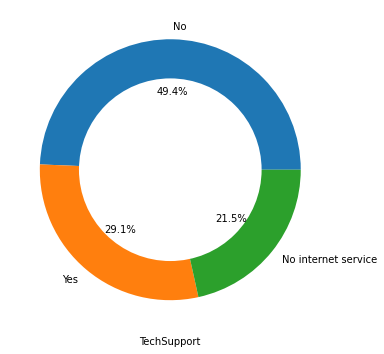
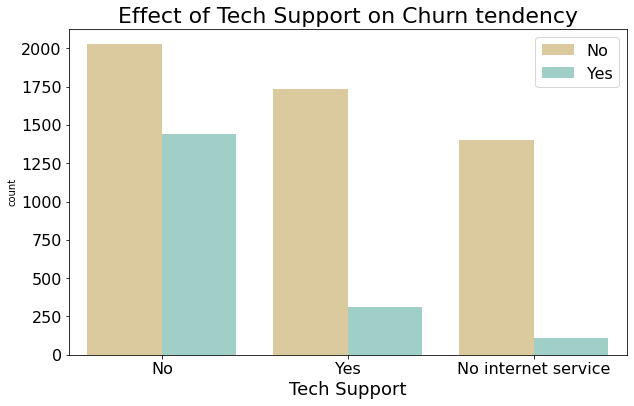
* Number of customers who don't use Internet service is 1512 and customers using Streaming Movies service is 2732 and who don't use Streaming Movies service is 2777
* Nothing much to conclude from Streaming movies service as equal amount of churning and non-churning can be seen for users and non-users

1. **Effect of Contract on Churn tendency**

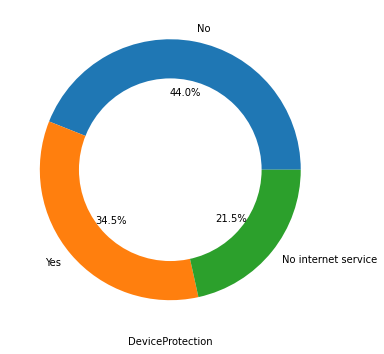
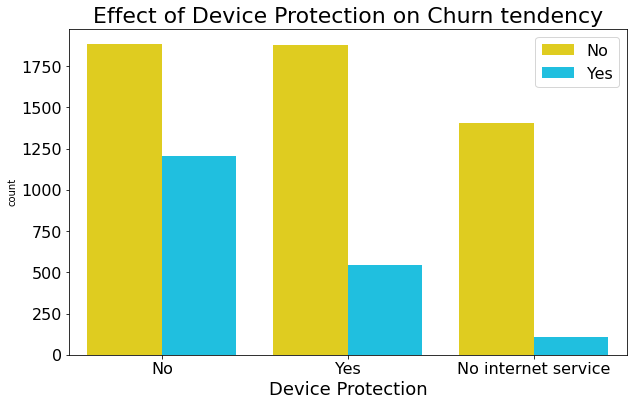
* Almost 55% customer prefer month to month contract compare to others
* Monthly contract customers are more likely to churn as they are free-to-go customers

1. **Effect of Tech Support on Churn tendency**

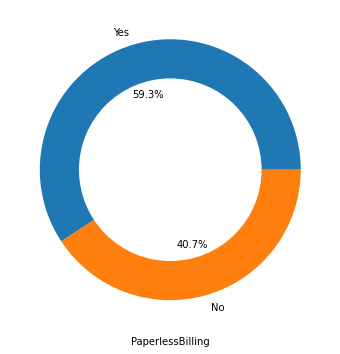
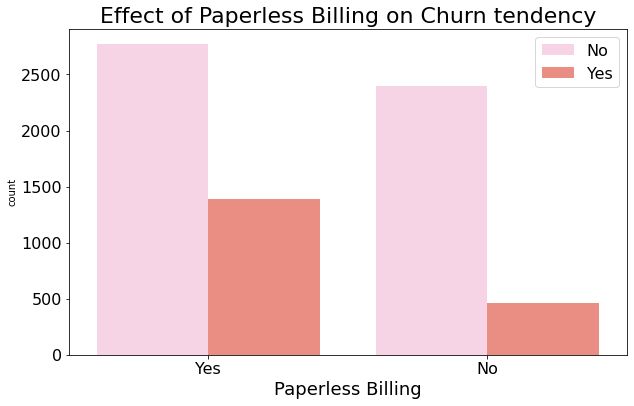
* Number of customers who don't use Internet service is 1512 and customers with tech support is 2044 and who don't use tech support is 3465
* Customers with no Technical support are more reluctant to churn

1. **Effect of Device Protection on Churn tendency**

* Only 34.5 % customers uses Device Protection service
* Customers with no device protection have more tendency to churn

1. **Effect of Paperless Billing on Churn tendency**

** **

* 60% of customers prefer paperless billing
* Customers using paperless billing has high tendency to churn

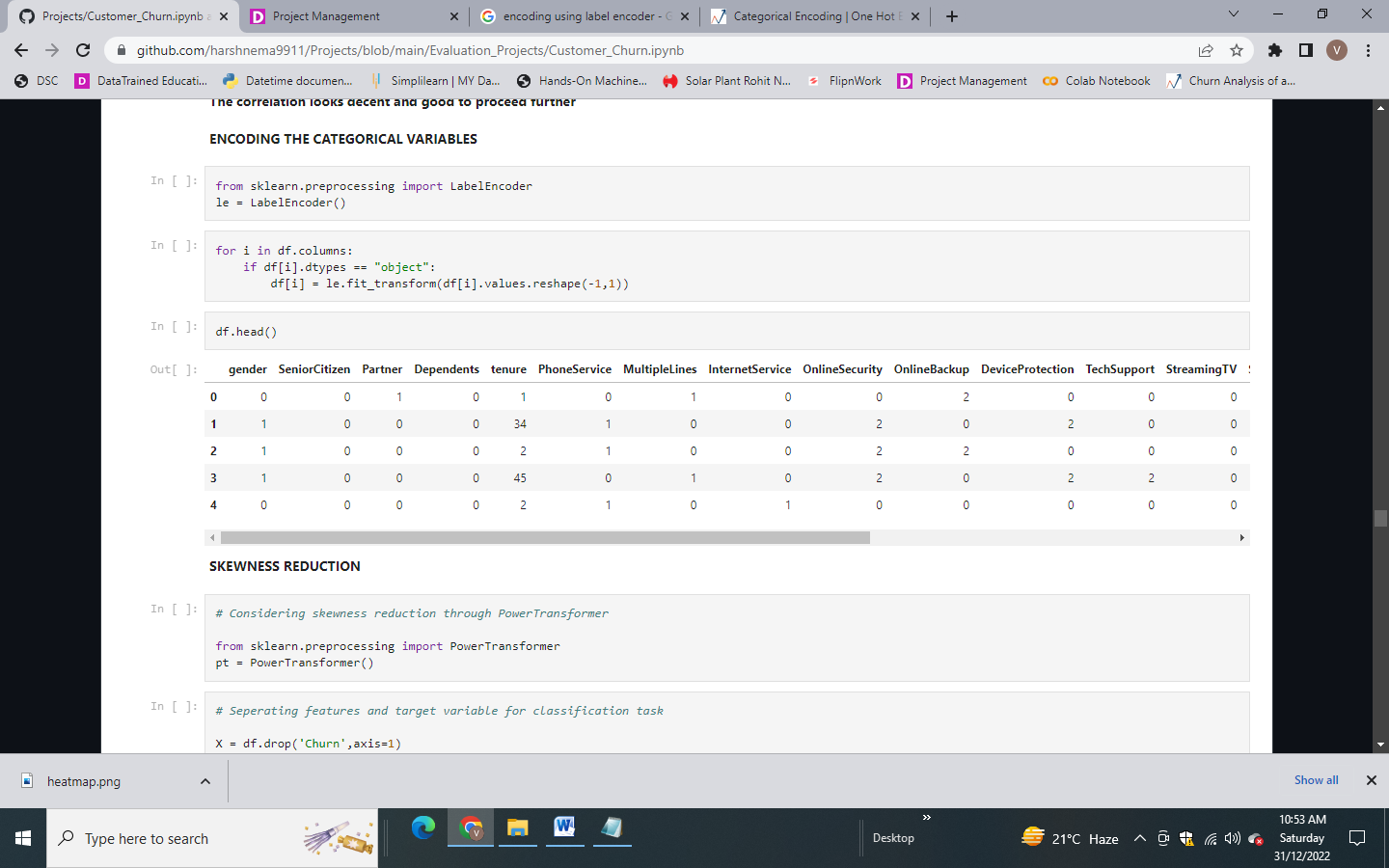
**Concluding remarks after EDA**

* Non-Senior Citizens are high Churners
* Bachelors are more reluctant to churn
* Customer having depedents have less tendency to Churn
* High monthly charges of Fiber Optic Internet might be the reason of customer churn
* Monthly contract customers are more likely to churn as they are free-to-go customers
* Customers with no Technical support are more reluctant to churn
* Customers with no device protection have more tendency to churn
* Customers using paperless billing has high tendency to churn

### Pre-Processing

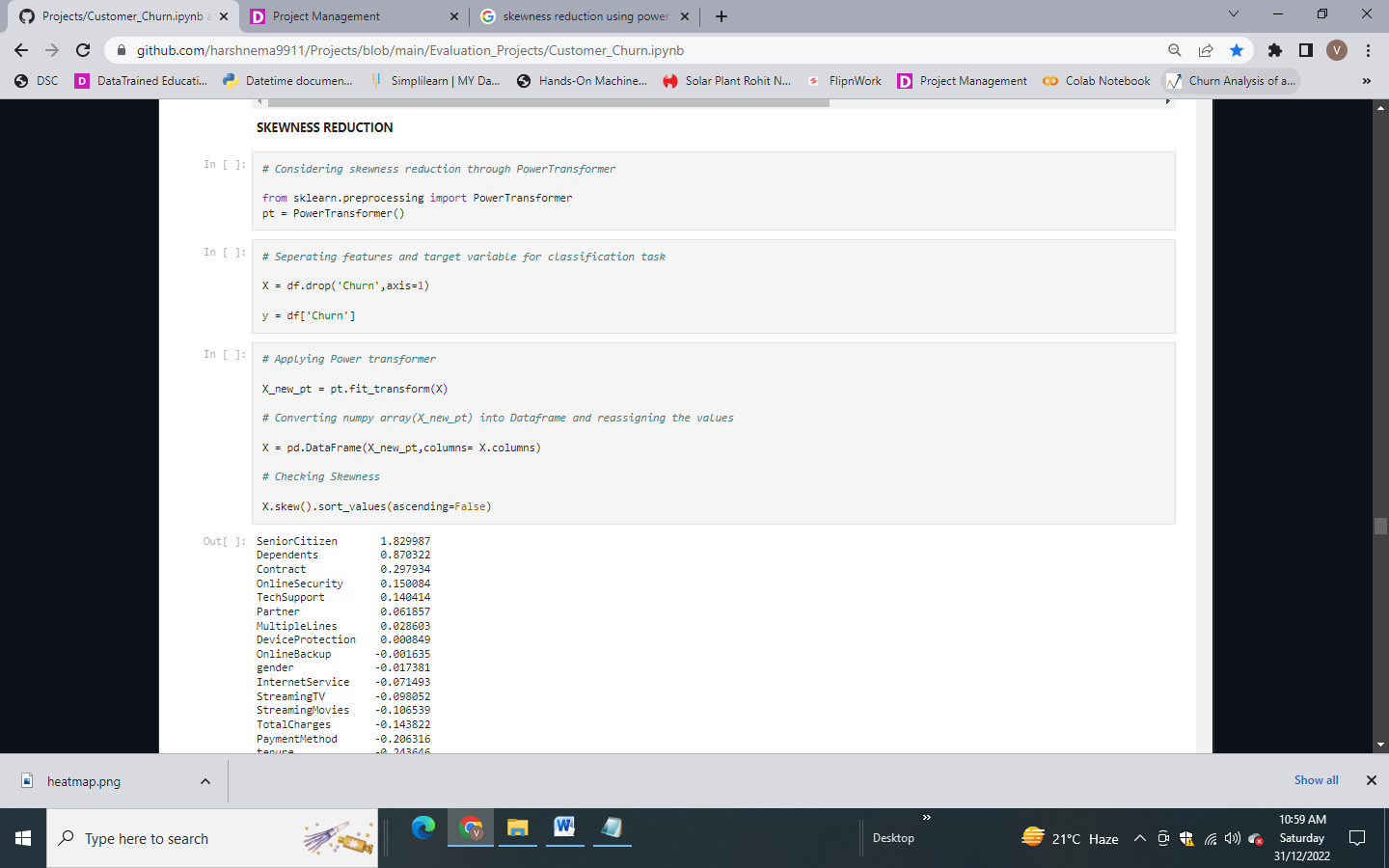
1. **Encoding the categorical columns**

Label Encoder is used for encoding the categorical variable. Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.



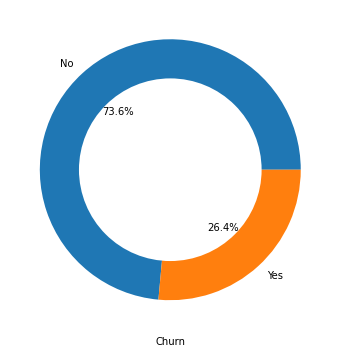
1. **Skewness Reduction**

A power transform will make the probability distribution of a variable more Gaussian.This is often described as removing a skew in the distribution, although more generally is described as stabilizing the variance of the distribution. Power transformer is used to reduce the skewness in the numerical variables.

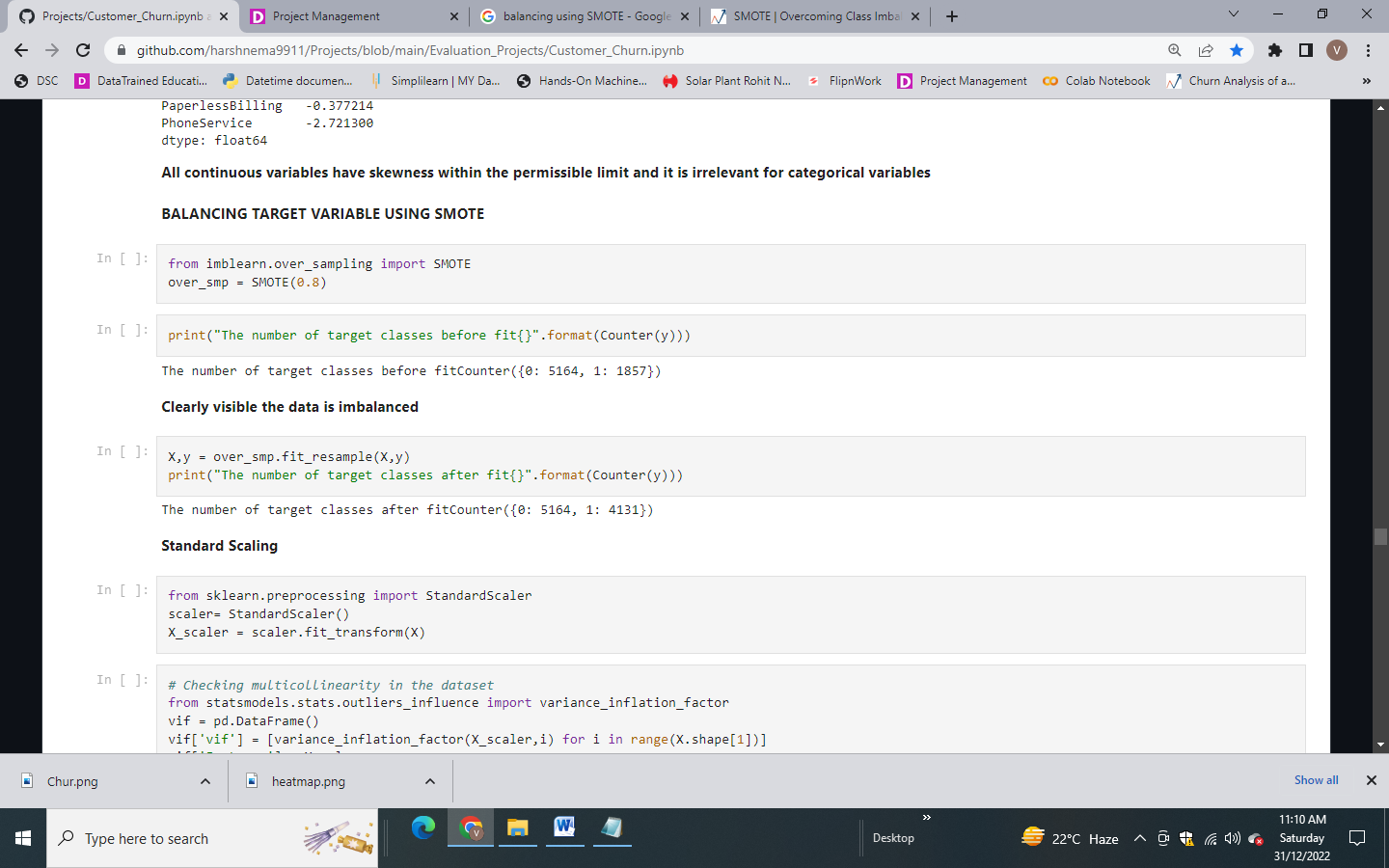


1. **Balancing the target variable using SMOTE**

SMOTE (Synthetic Minority Oversampling Technique) is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

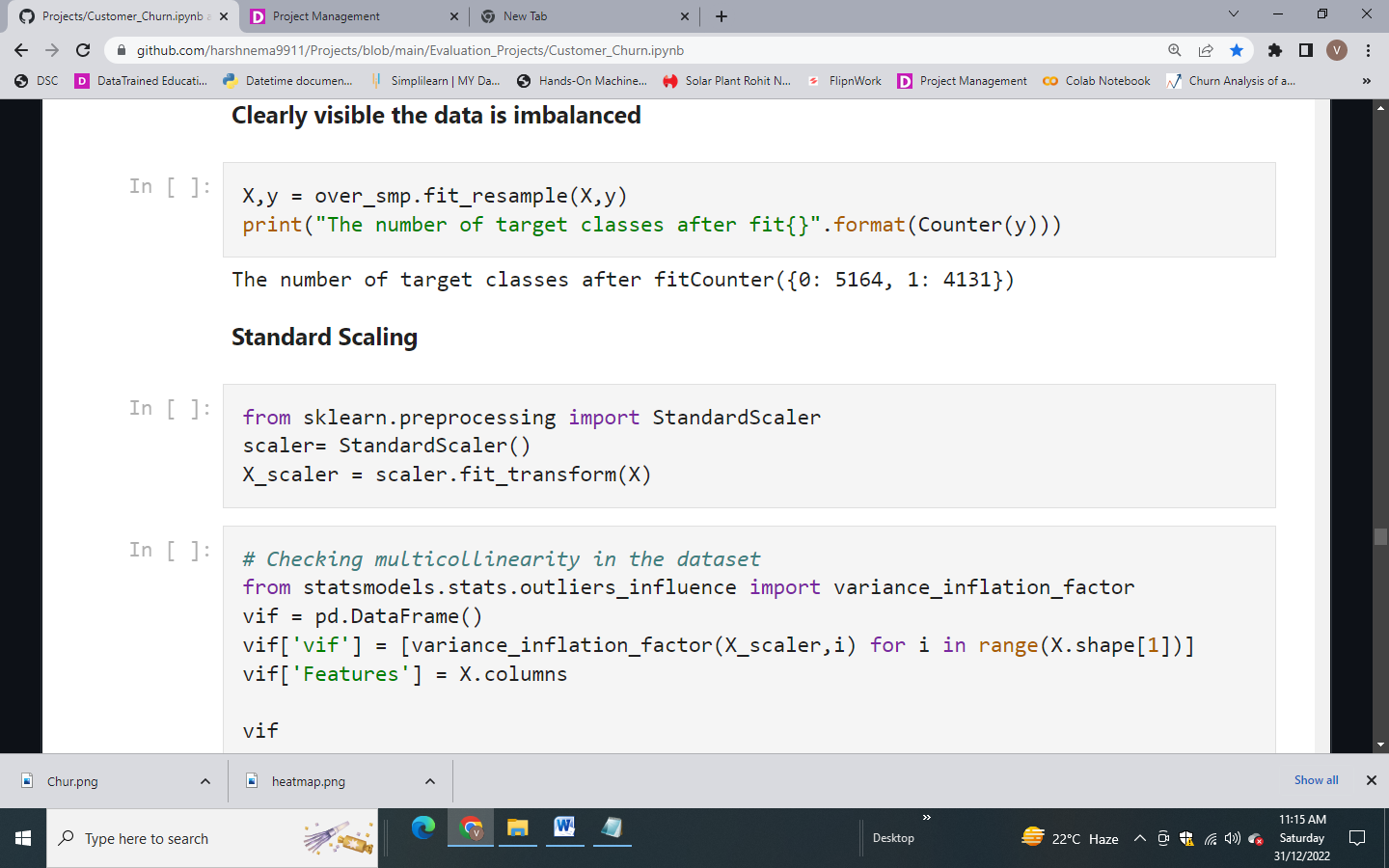


73.6% customers are satisfied with their current telecom company. But as far as model building is concerned, we need a balanced dataset i.e no bias to any class. Hence we will be using oversampling technique to balance the target class of our dataset.



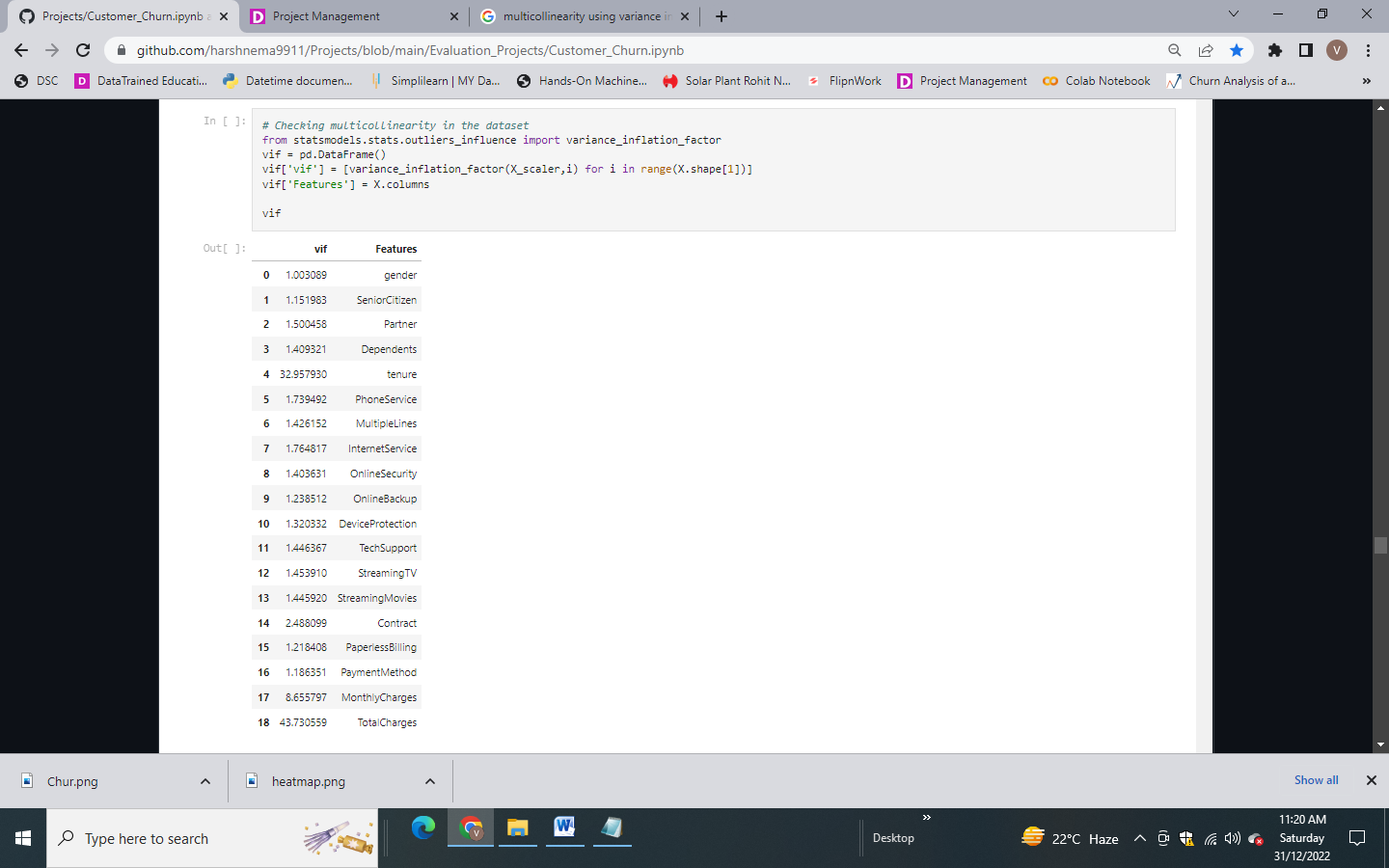
1. **Scaling of Data using Standard Scaler**

Standard Scaler standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation.



1. **Multicollinearity Check**

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.



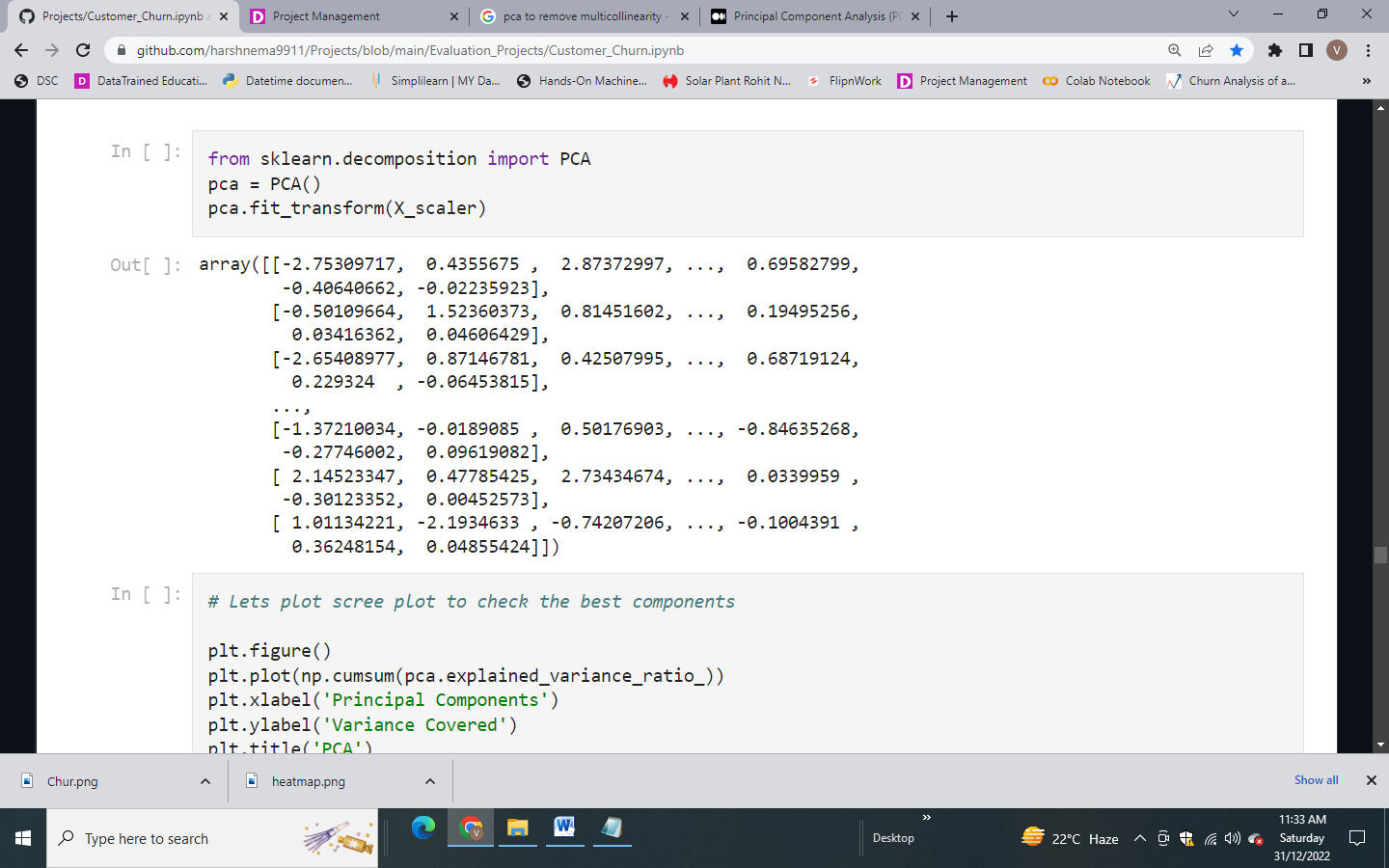
To deal with multicollinearity:

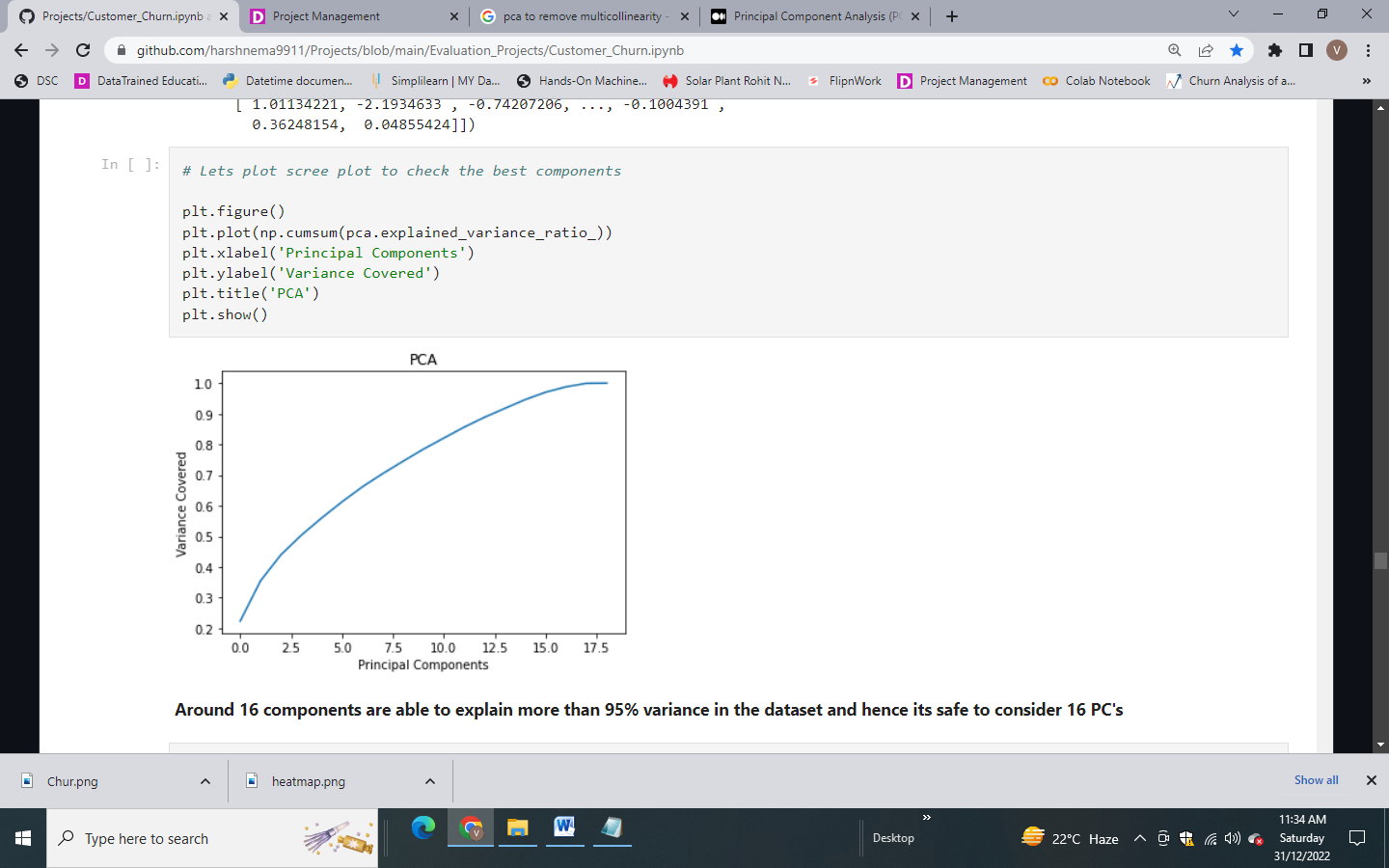
* Remove some of the highly correlated independent variables
* Linearly combine the independent variables, such as adding them together
* Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression

Tenure and TotalCharges showing vif value greater than 10. It means multicollinearity is present in the dataset.

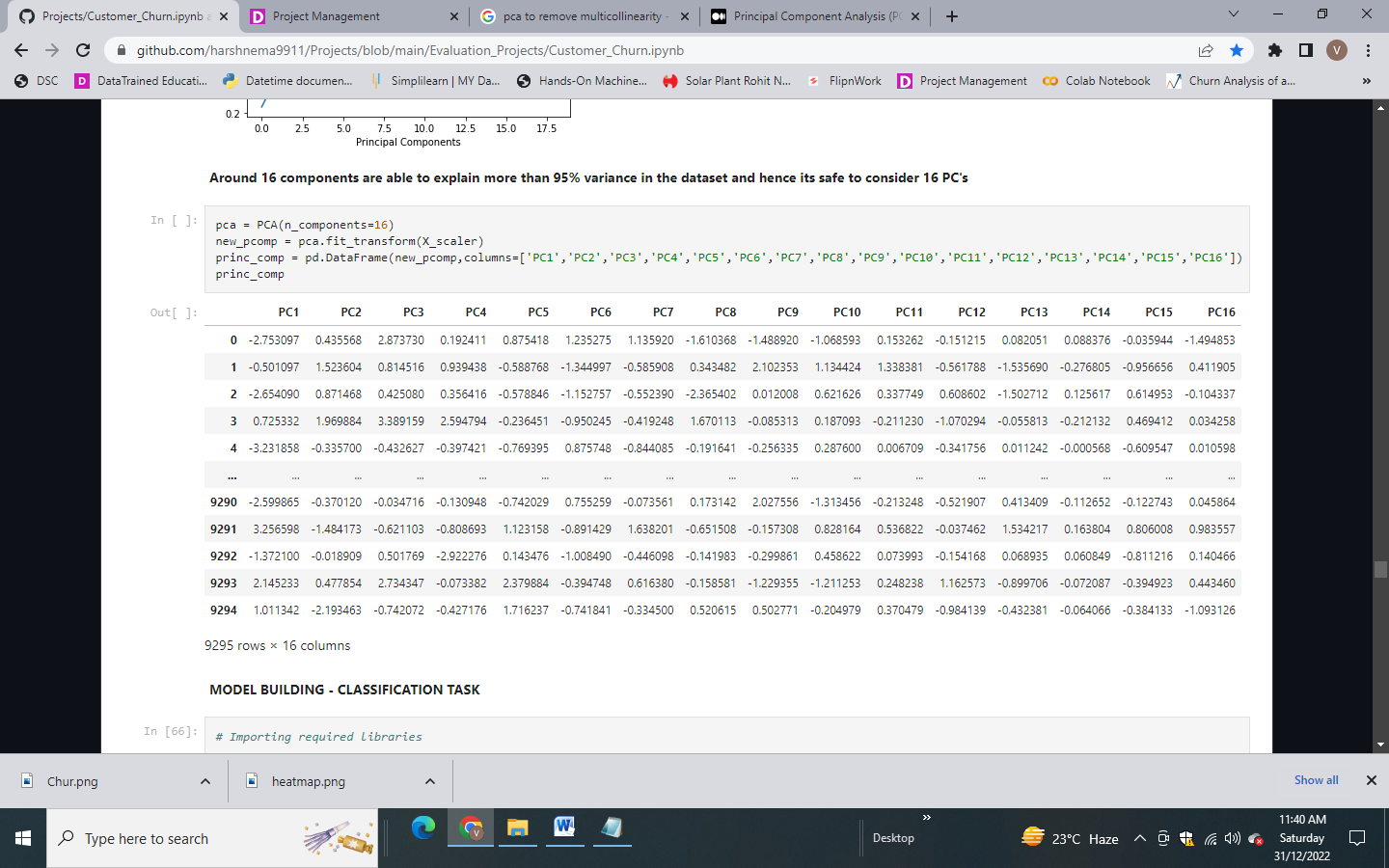
1. **Principal Component Analysis**

PCA (Principal Component Analysis) takes advantage of multicollinearity and combines the highly correlated variables into a set of uncorrelated variables called principal components. Therefore, PCA can effectively eliminate multicollinearity between features.PCA aims to reduce dimensionality in our dataset.



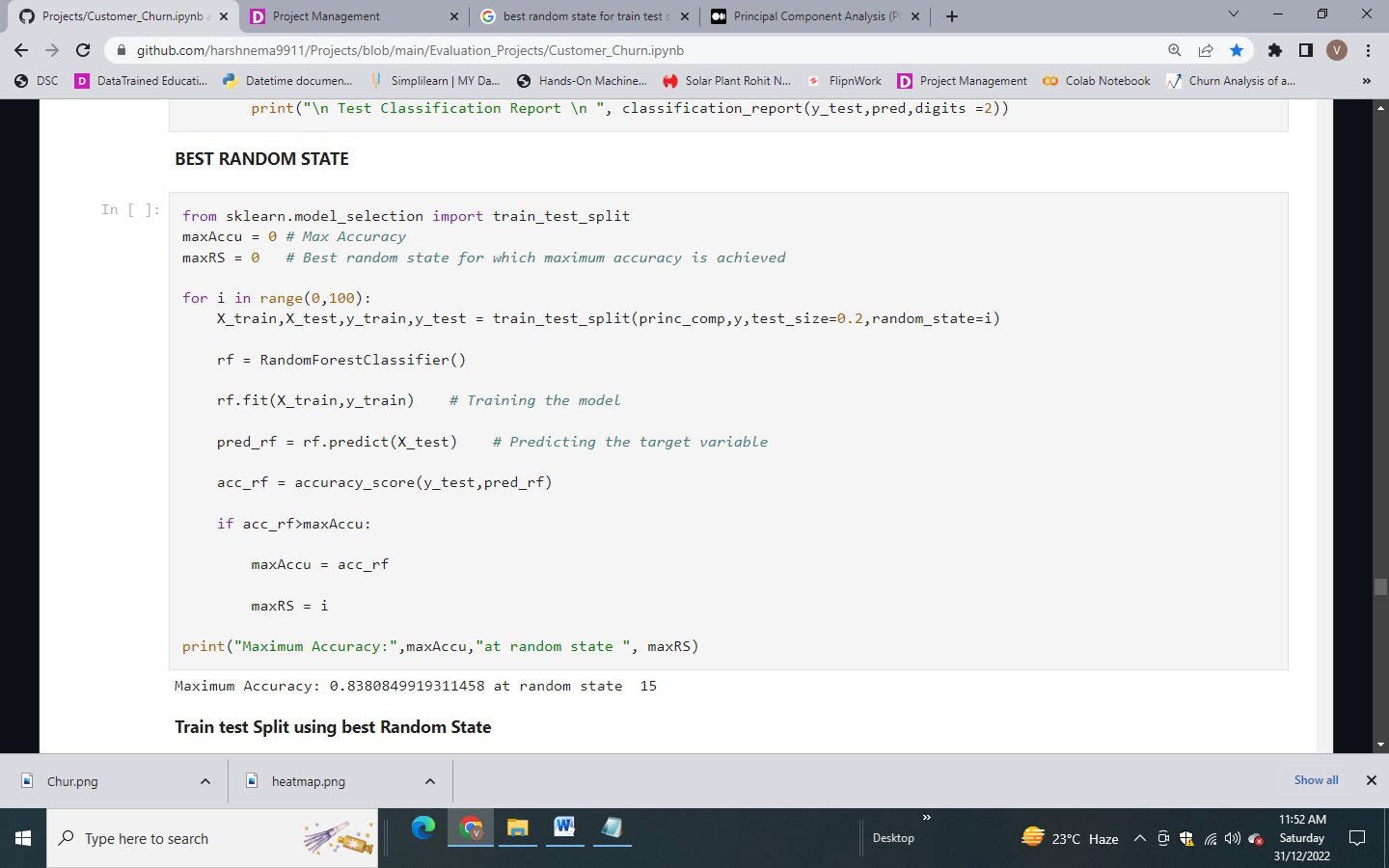


* Out of 19 independent features, we are getting 16 components explaining more than 95% of variance in the dataset and hence it is safe to consider 16 Principal components.



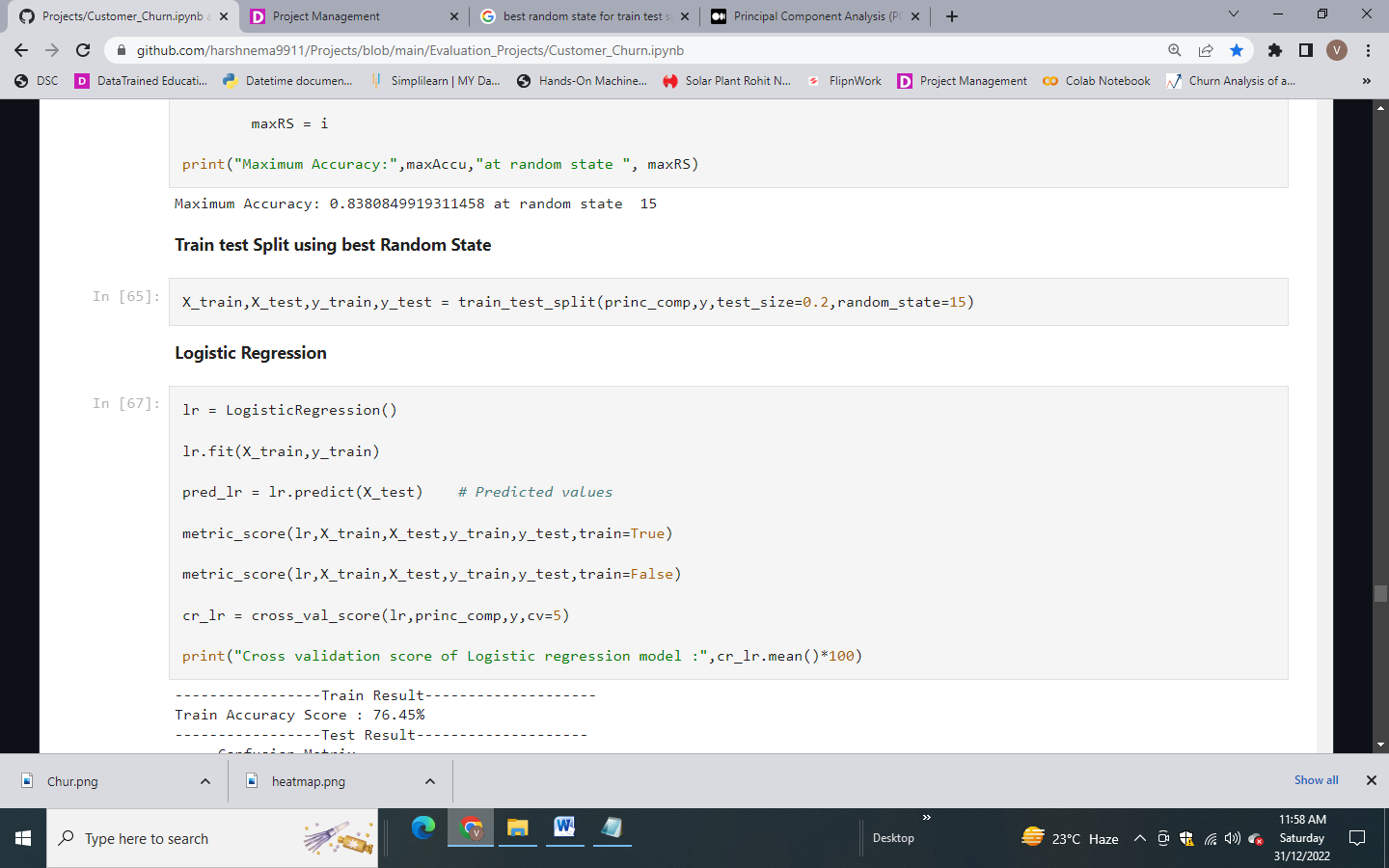
### Building Machine Learning Models

First, we will find the best random state. The random state hyper parameter in the train\_test\_split function controls the shuffling process. With random state=None , we get different train and test sets across different executions and the shuffling process is out of control.



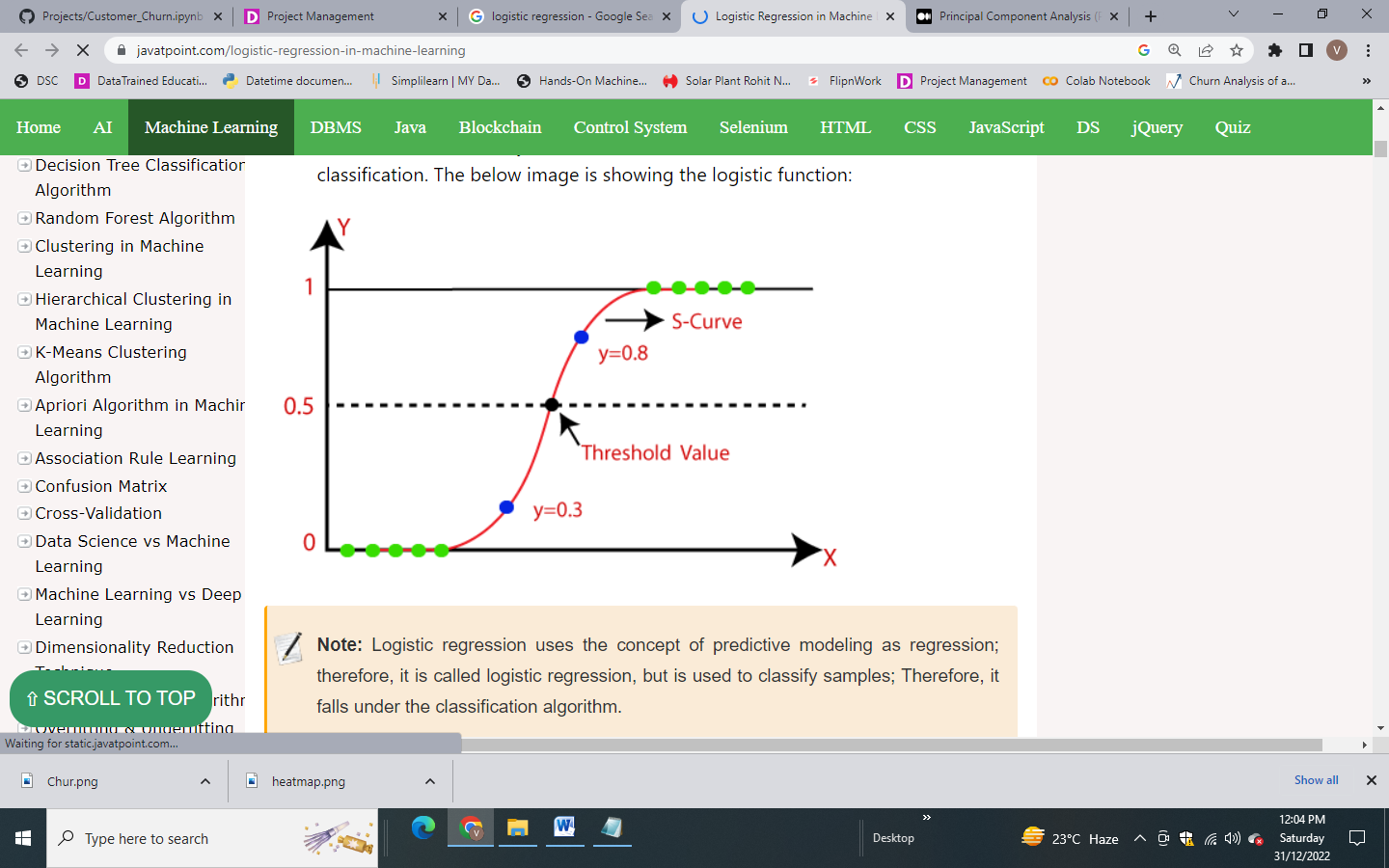
We get random state = 15 as we are getting the maximum R2 score for our base model (Random Forest Classifier).

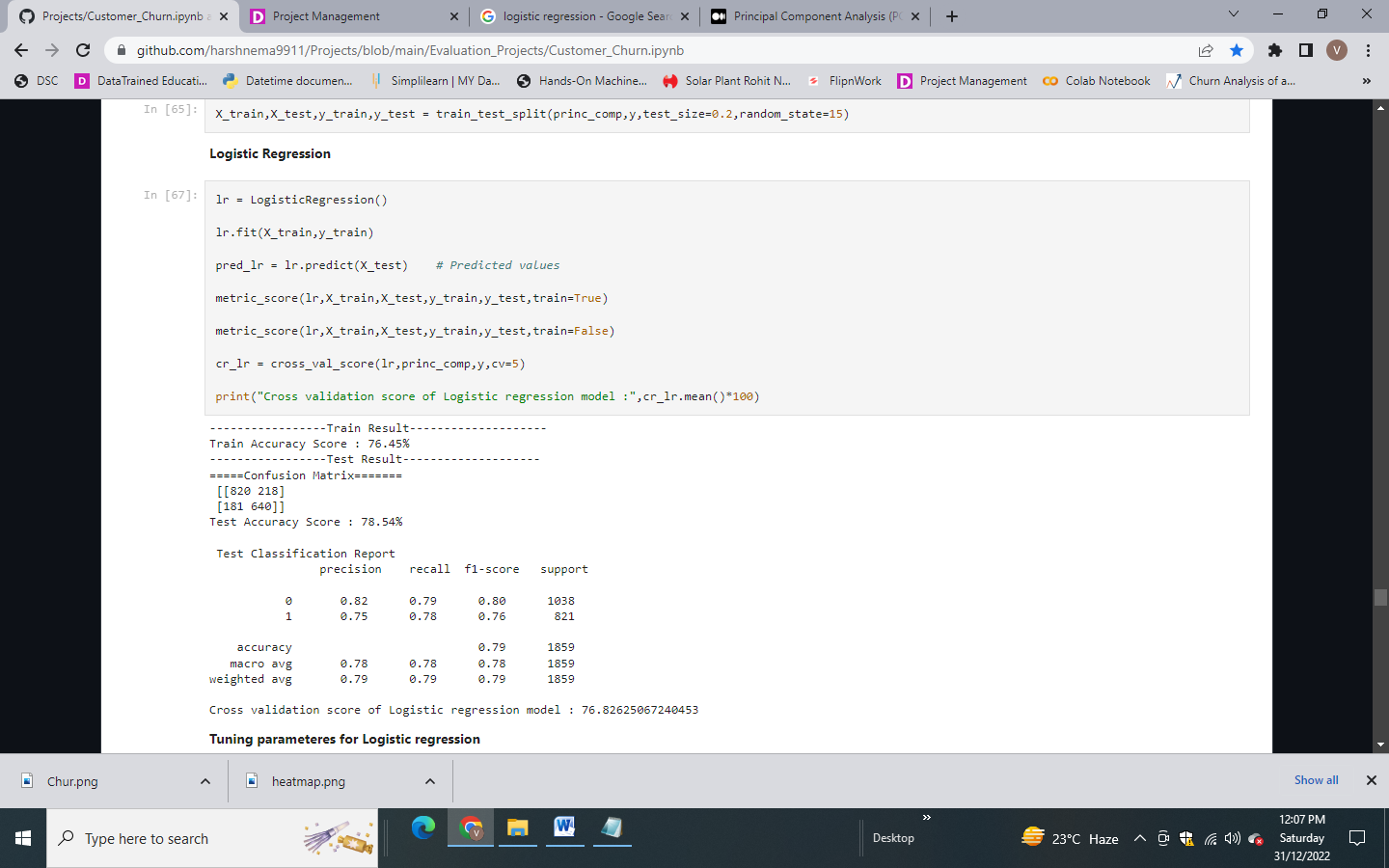
Then we will split the data into train and test dataset using the best random state.



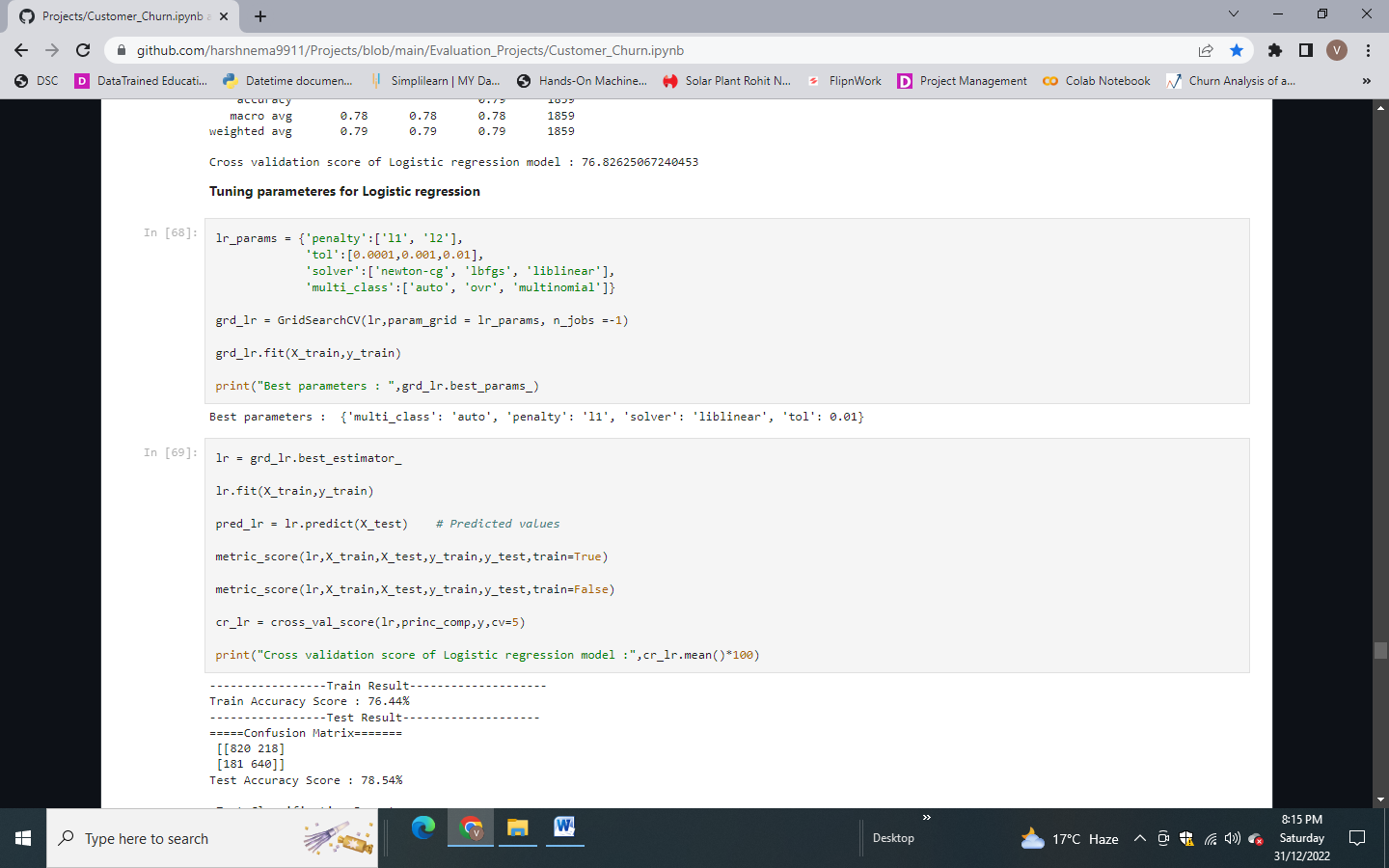
**Model 1 – Logistic Regression**

Logistic Regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.





The result looks good. We will try to improve the performance of the model by tuning the parameters.

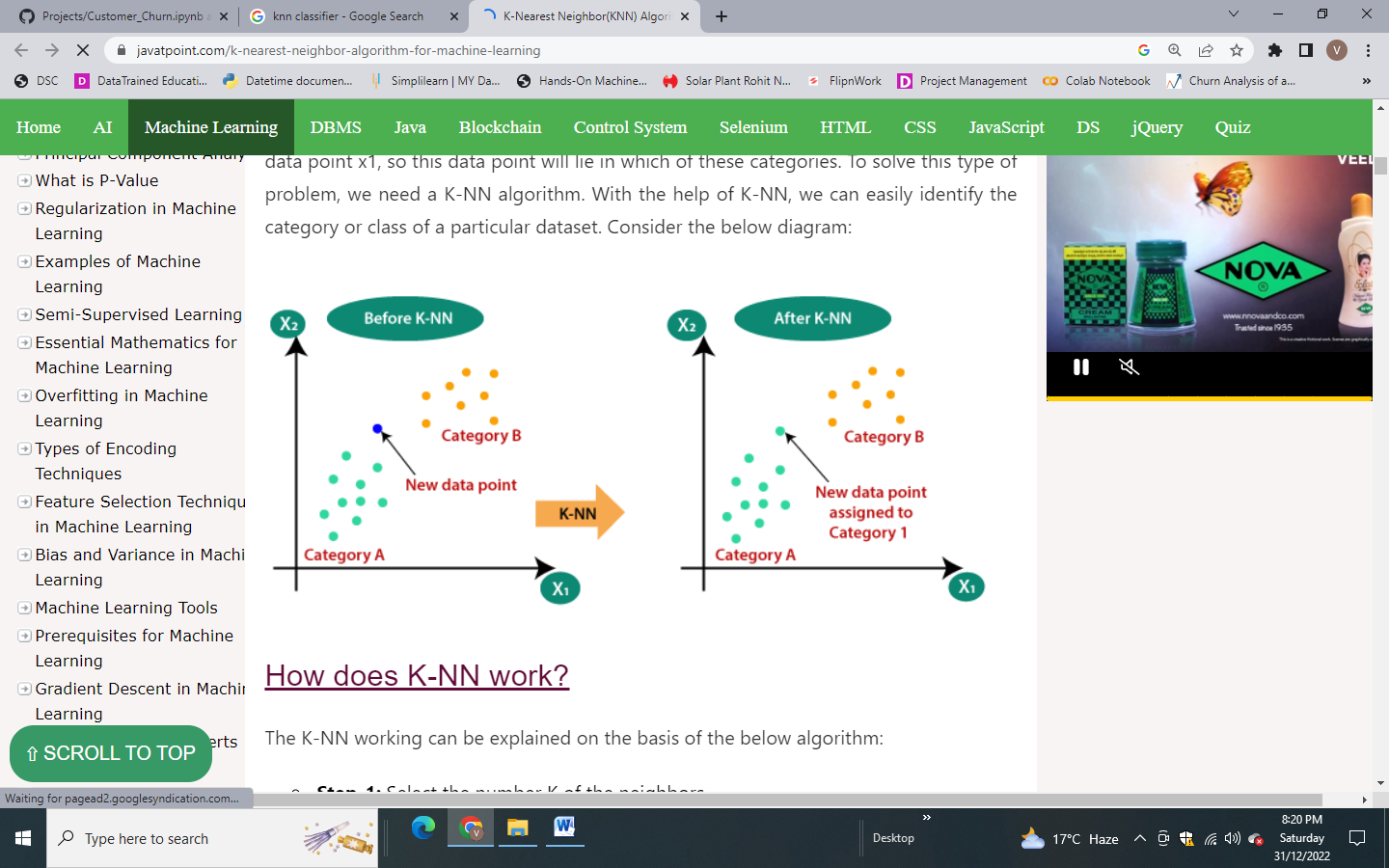


We get the following result:

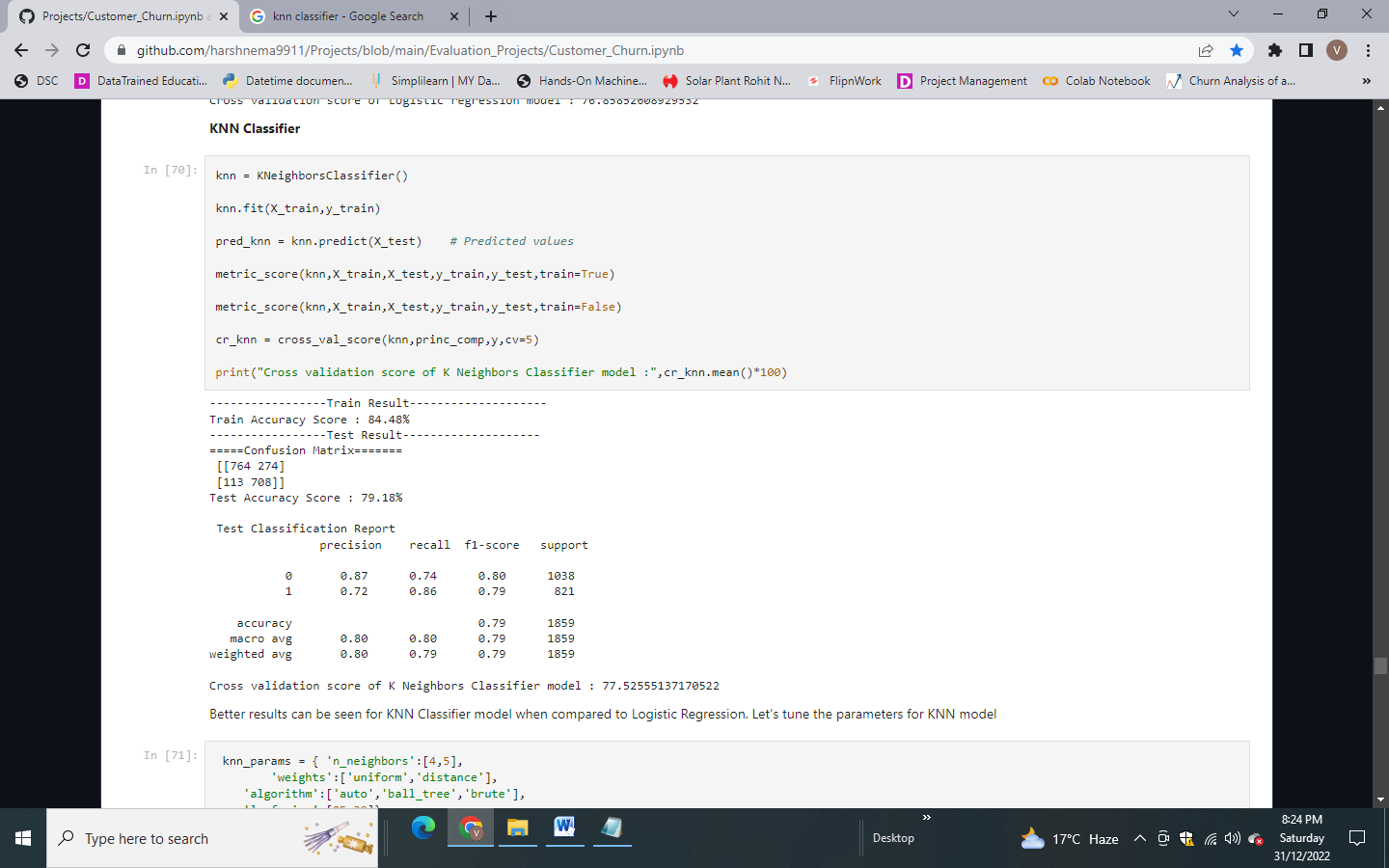


Not much difference after tuning the parameters. Lets go for another model.

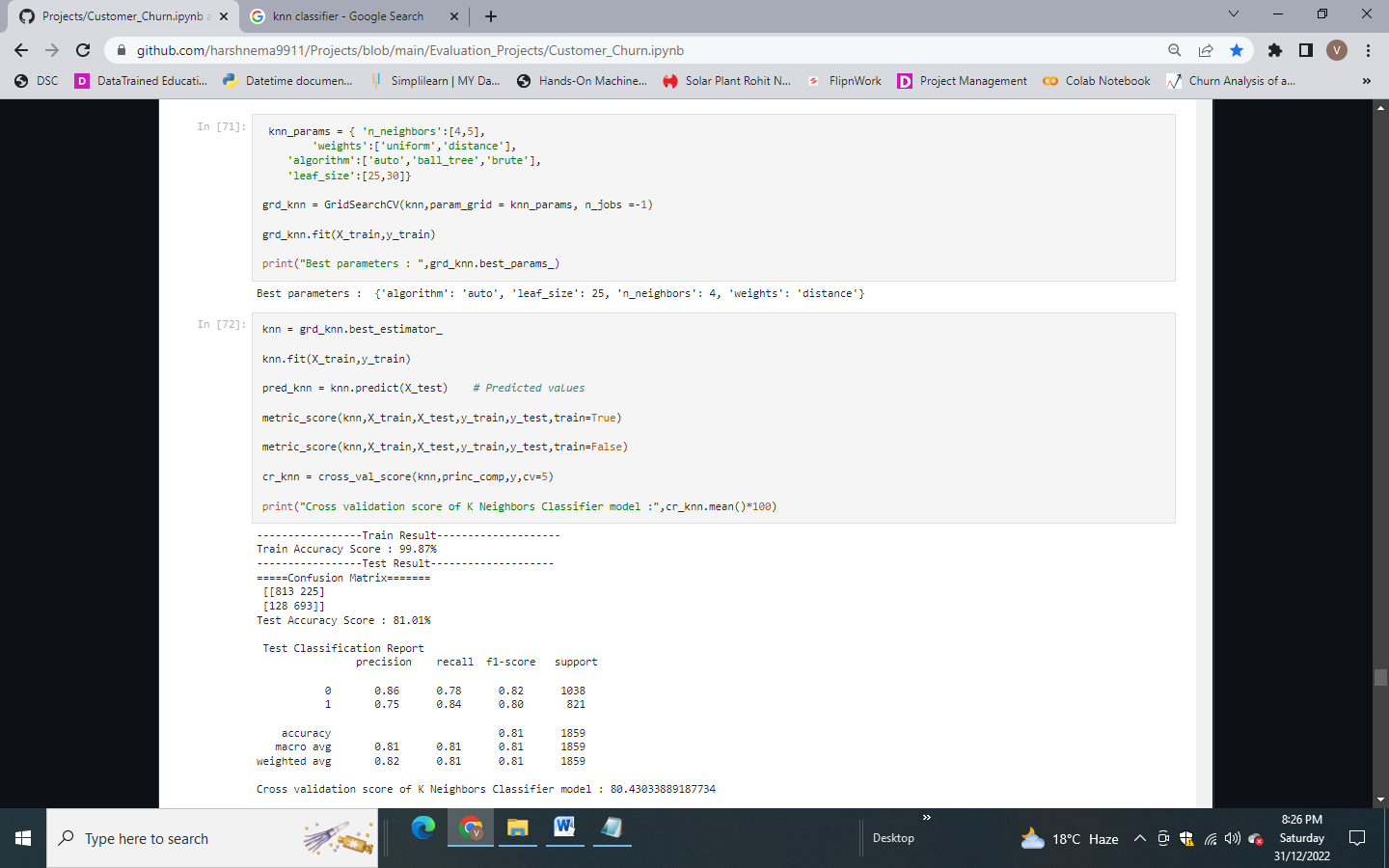
**Model 2 – K Nearest Neighbor Classifier**



K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

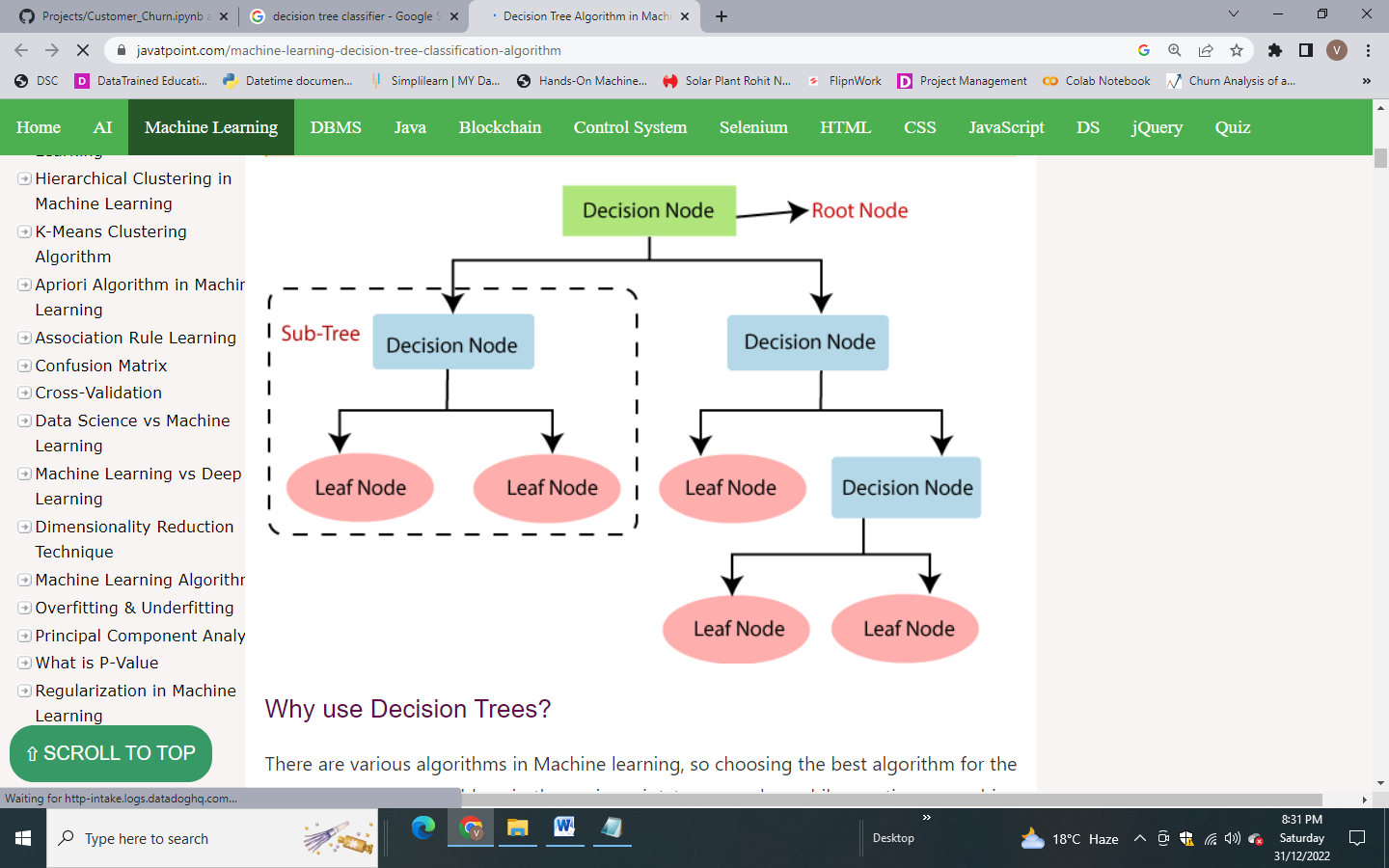


Hyperparameter tuning for KNN Classifier:

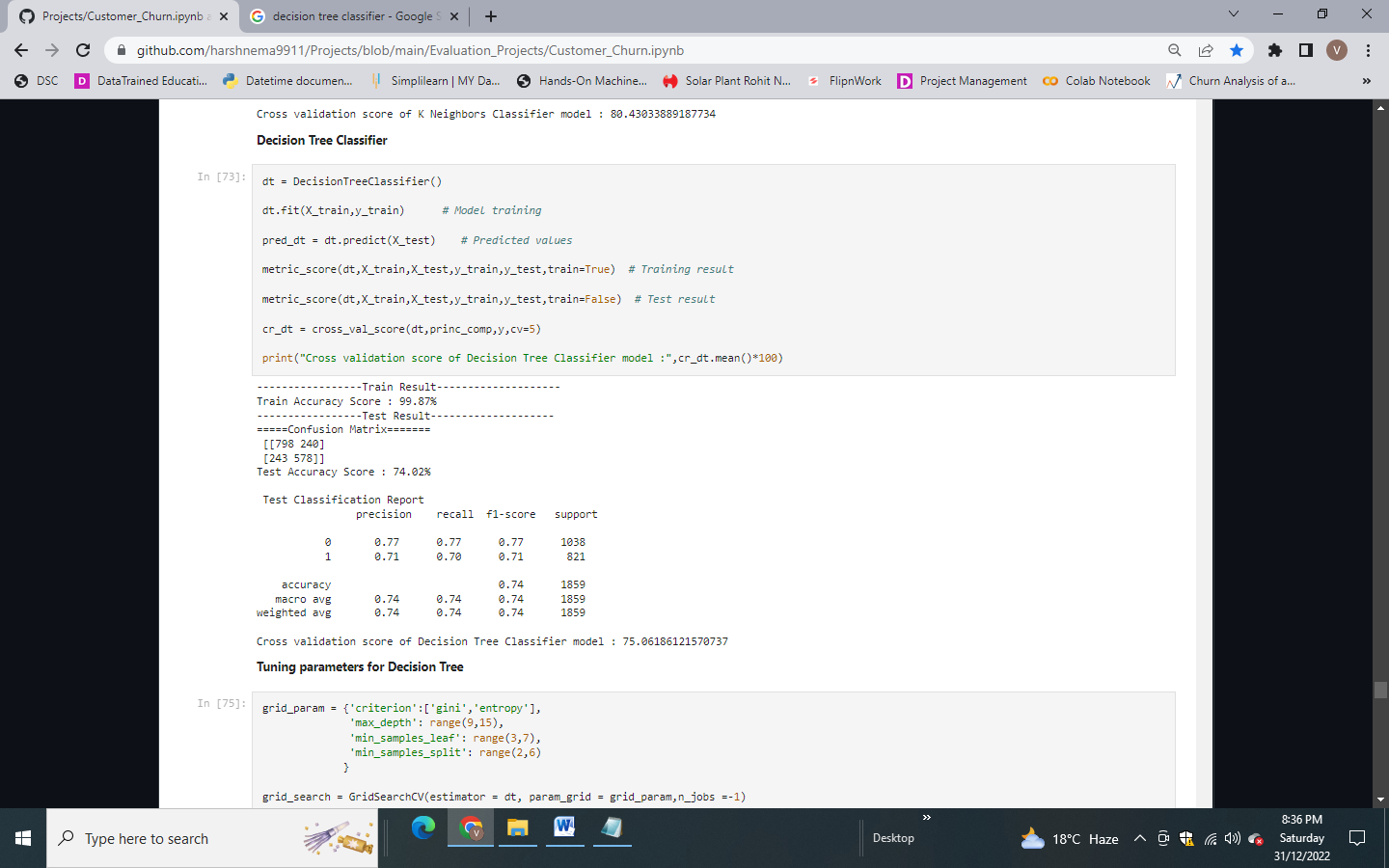


We get better result after hyperparameter tuning for this model. Let’s check another model.

**Model 3 – Decision Tree Classifier**



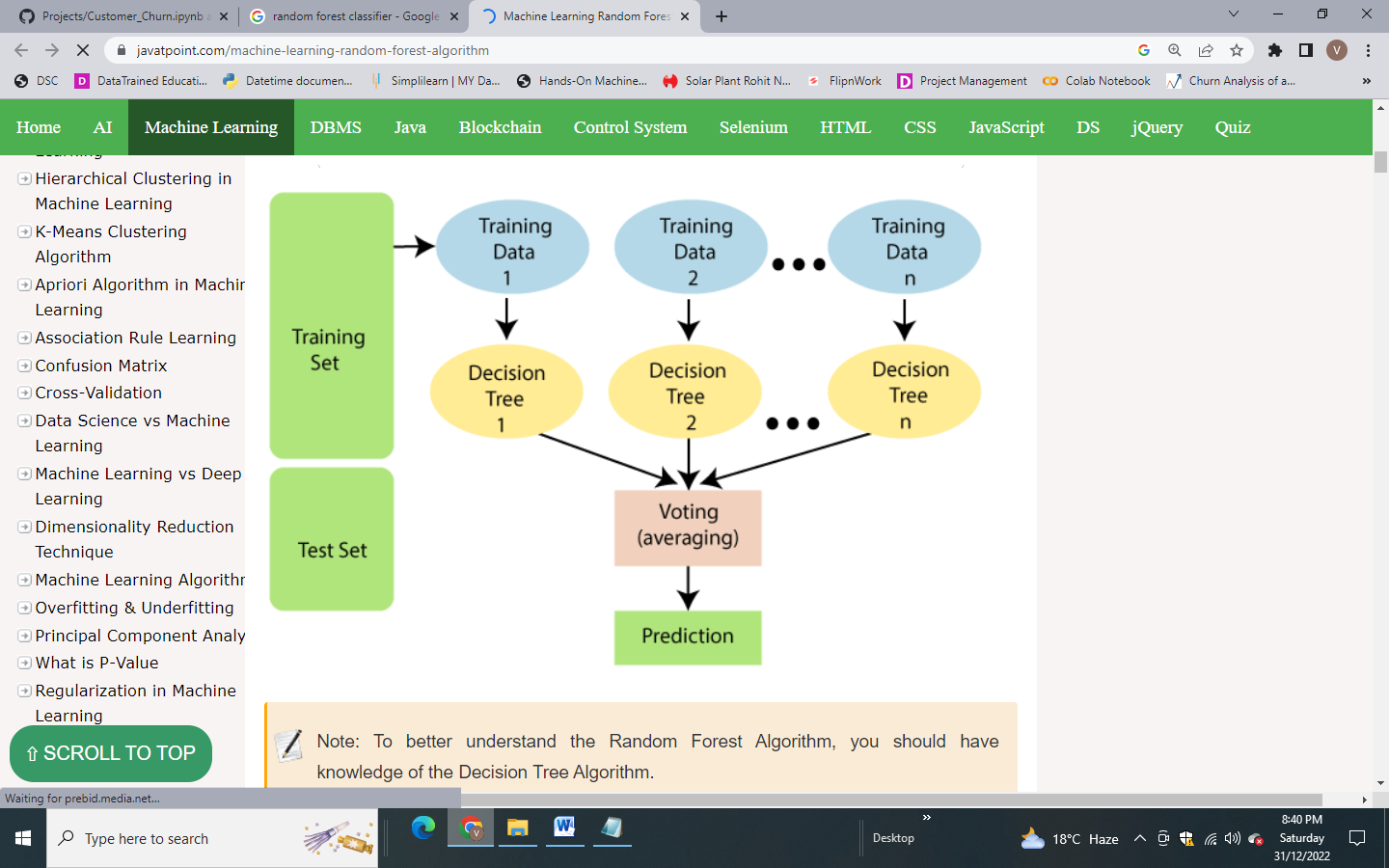
It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.



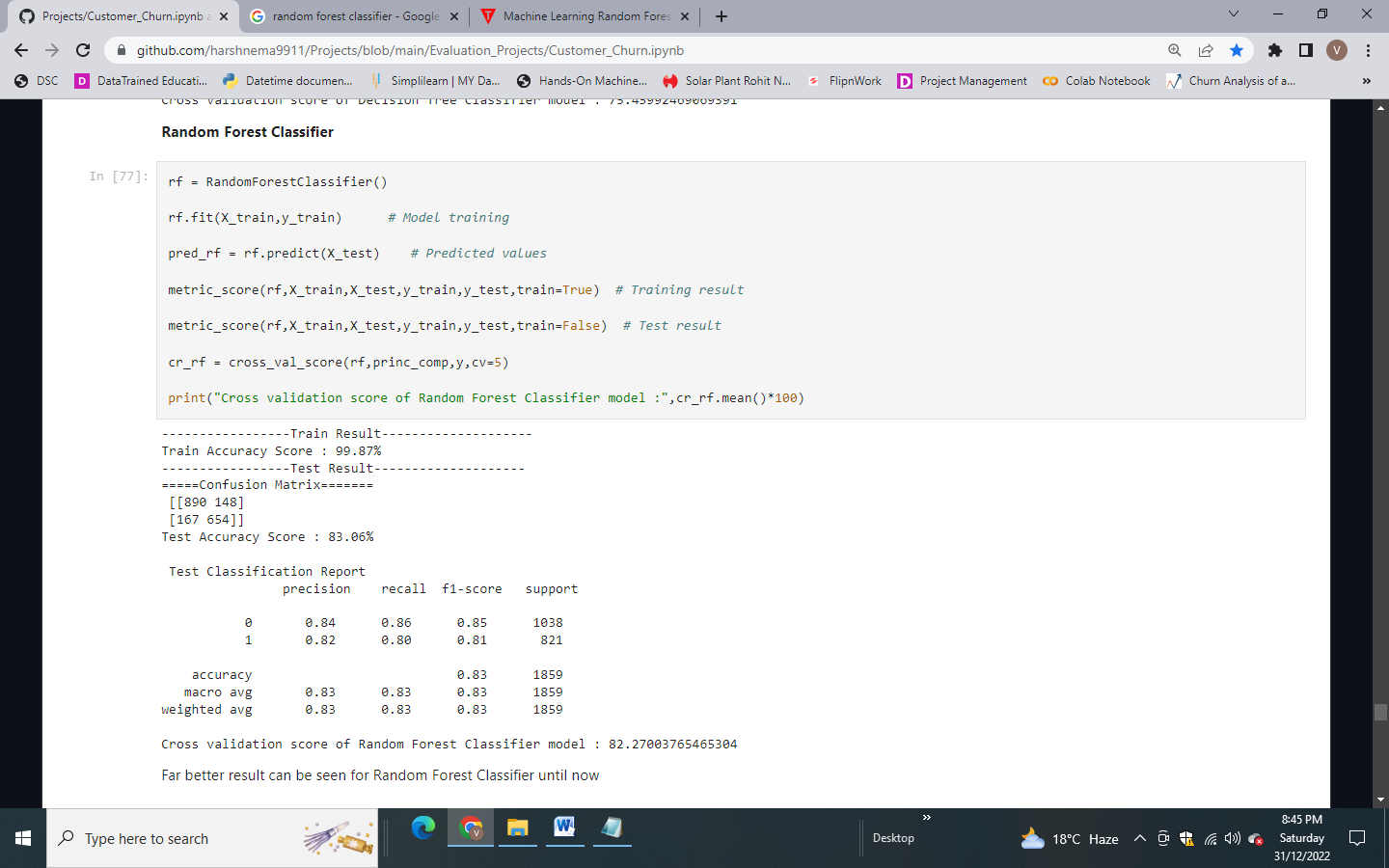
Performing hyper parameter tuning for Decision tree classifier:



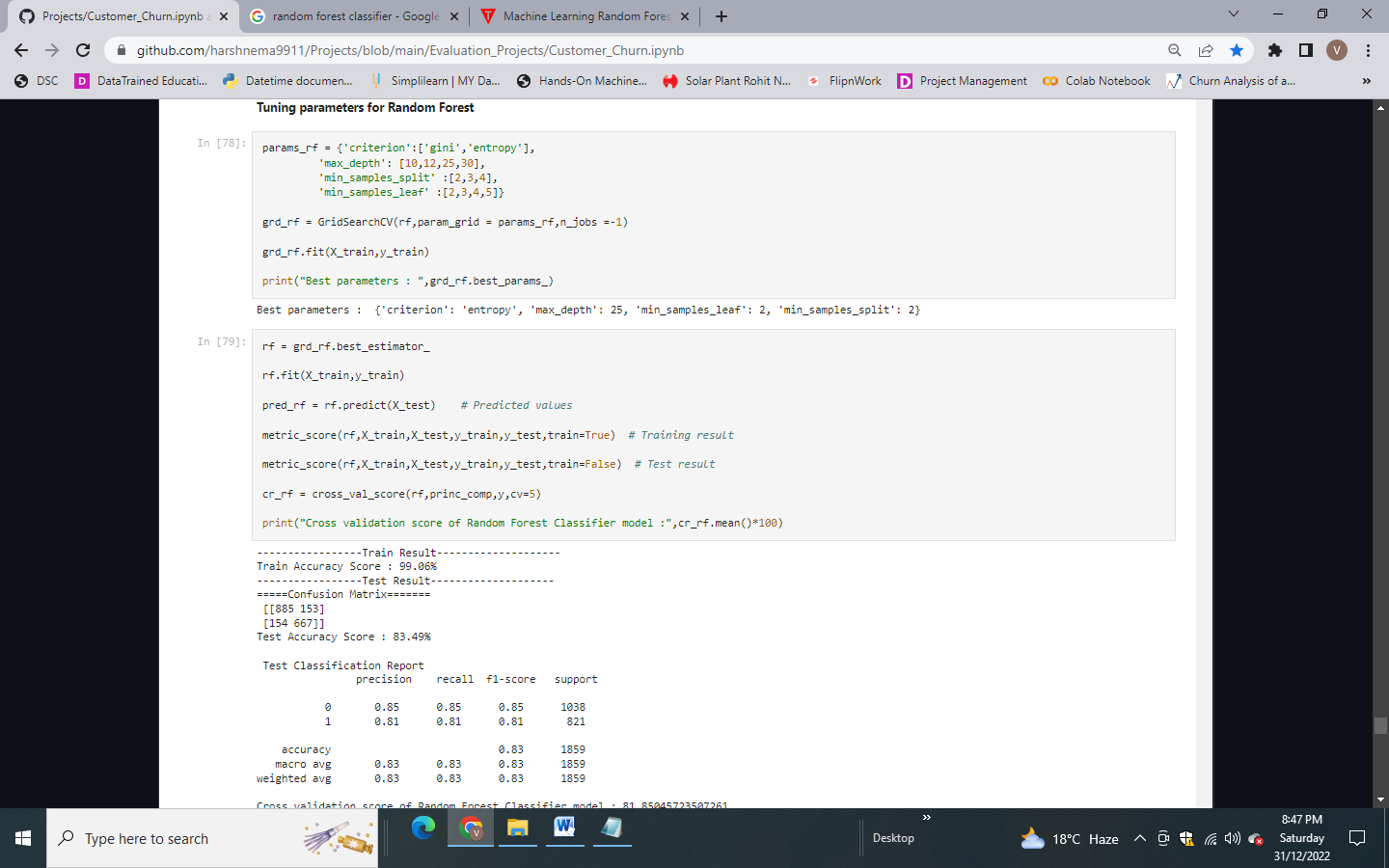
#### Model 4 – Random Forest Classifier



Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase. New data points are assigned to the category that wins the majority votes.

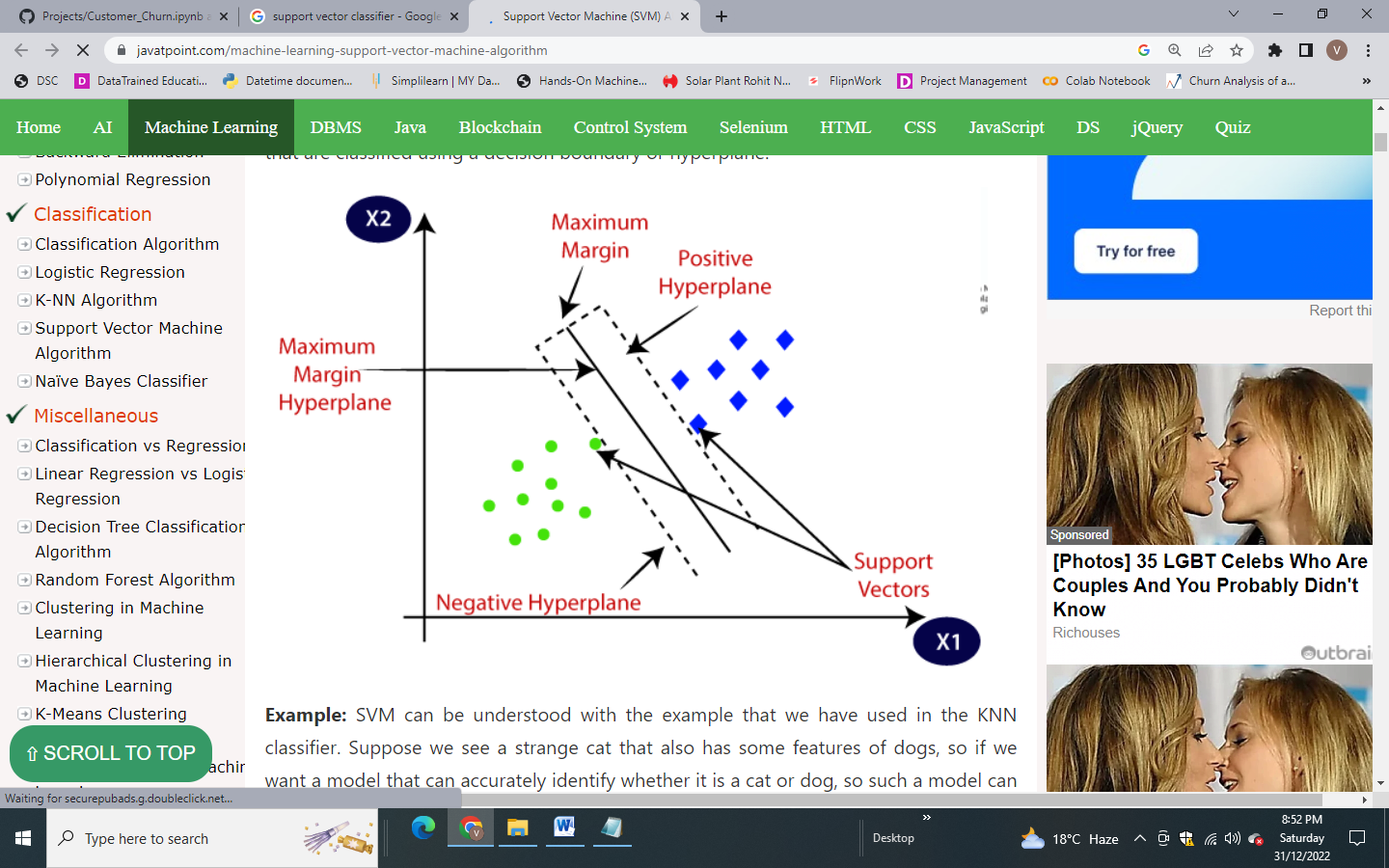


Hyper parameter tuning is performed on Random Forest:

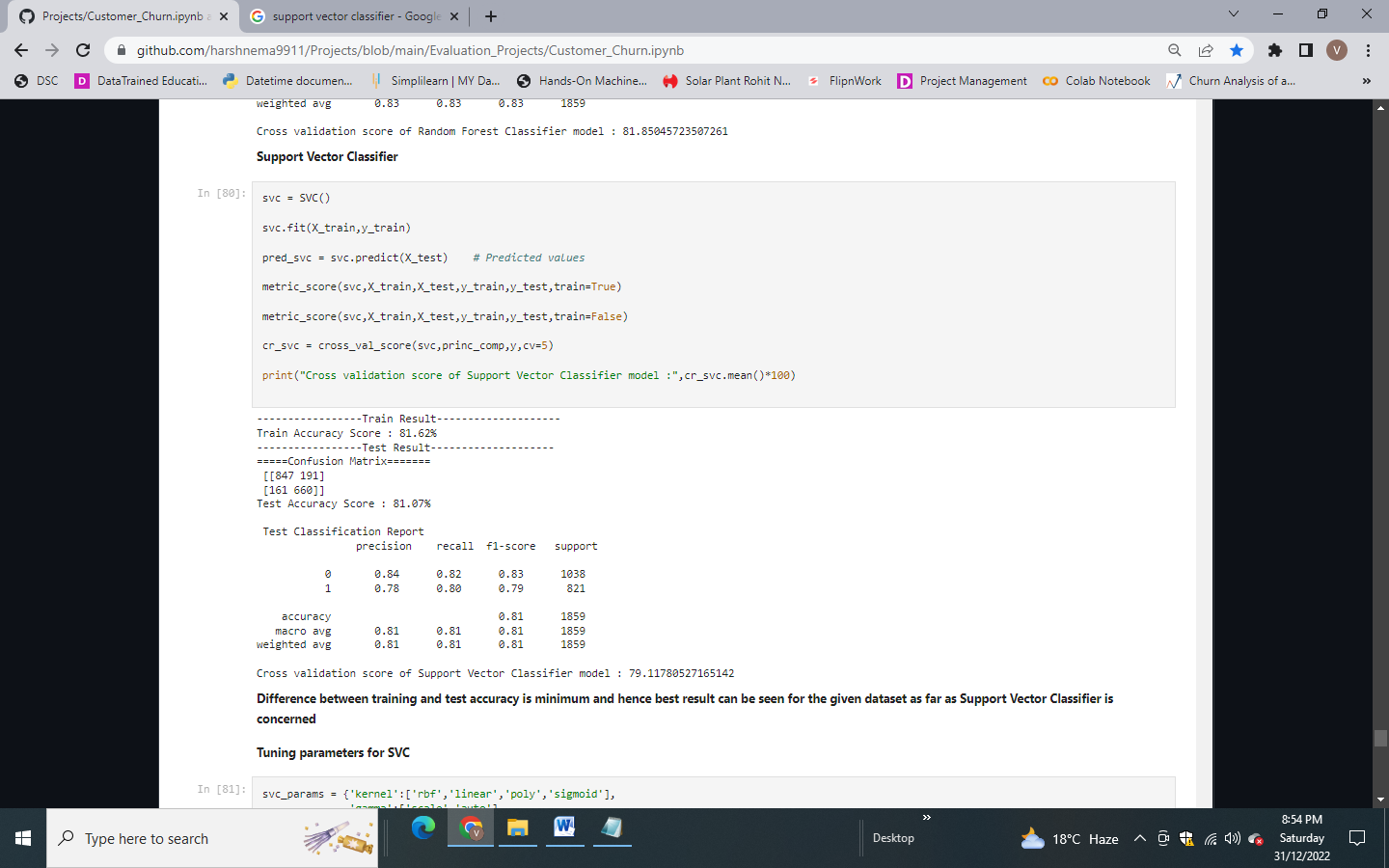


Not much difference was observed after hyper parameter tuning for Random forest model. But we get better training and testing accuracy and max test accuracy of 83.49%.

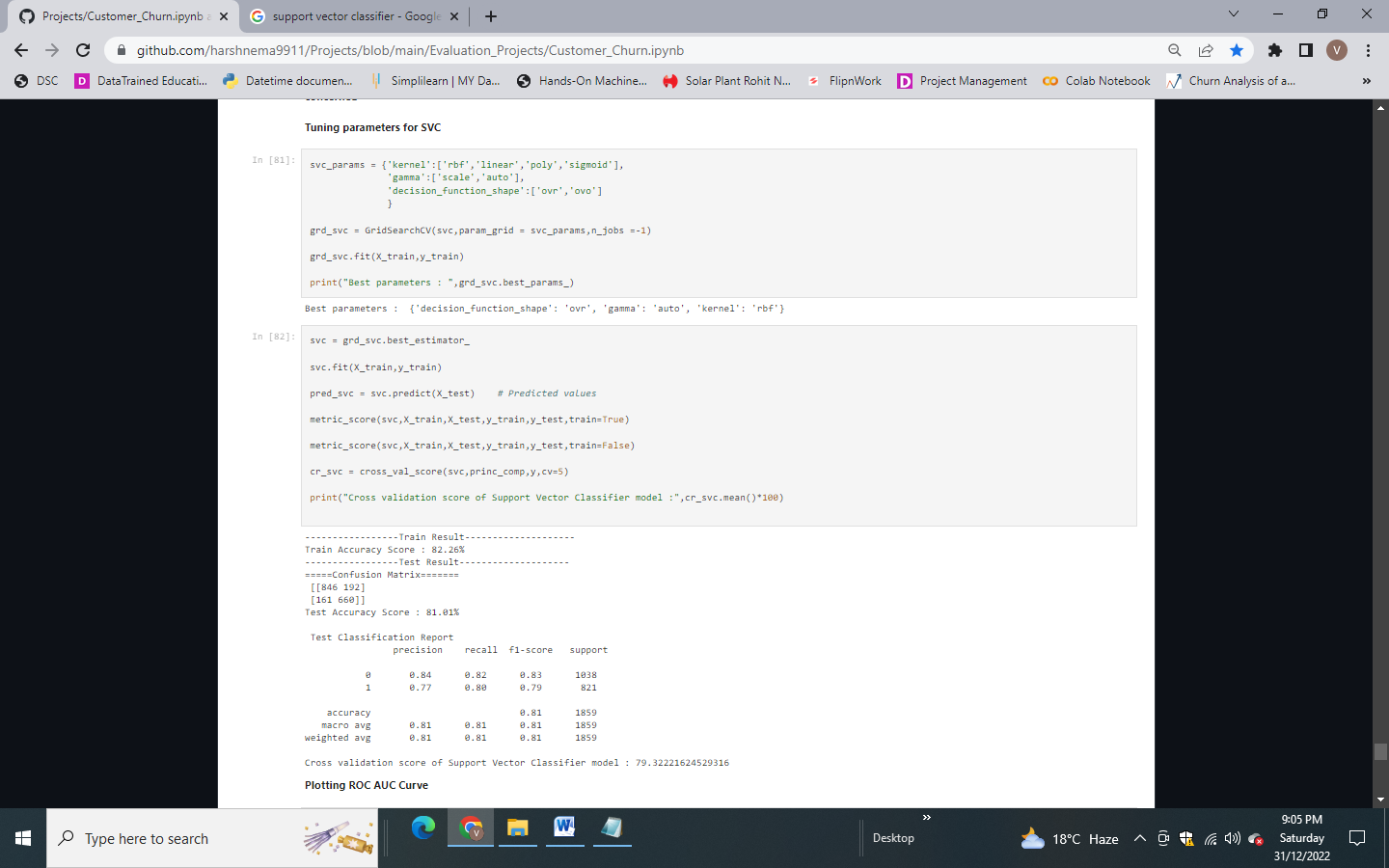
#### Model 5 – Support Vector Classifier



The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.



We get the result with minimum difference between training and test accuracy. Tuning parameters for Support Vector Classifier:

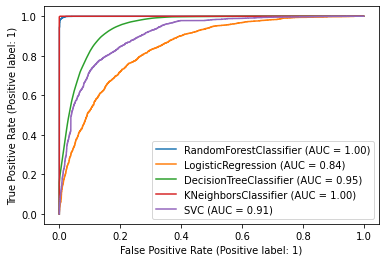


There is very little improvement in the result after tuning the parameters.

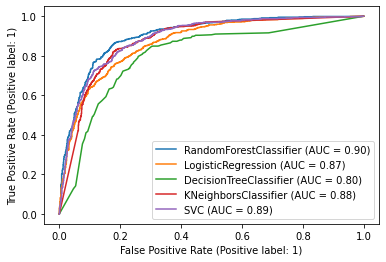
### Concluding Remark

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'.

ROC Curve for training dataset:



ROC Curve for test dataset:



As per ROC AUC Curve, **SVC model fits best for the given dataset as there is minimum difference between training and testing accuracy.**