1. **Problem statement –**

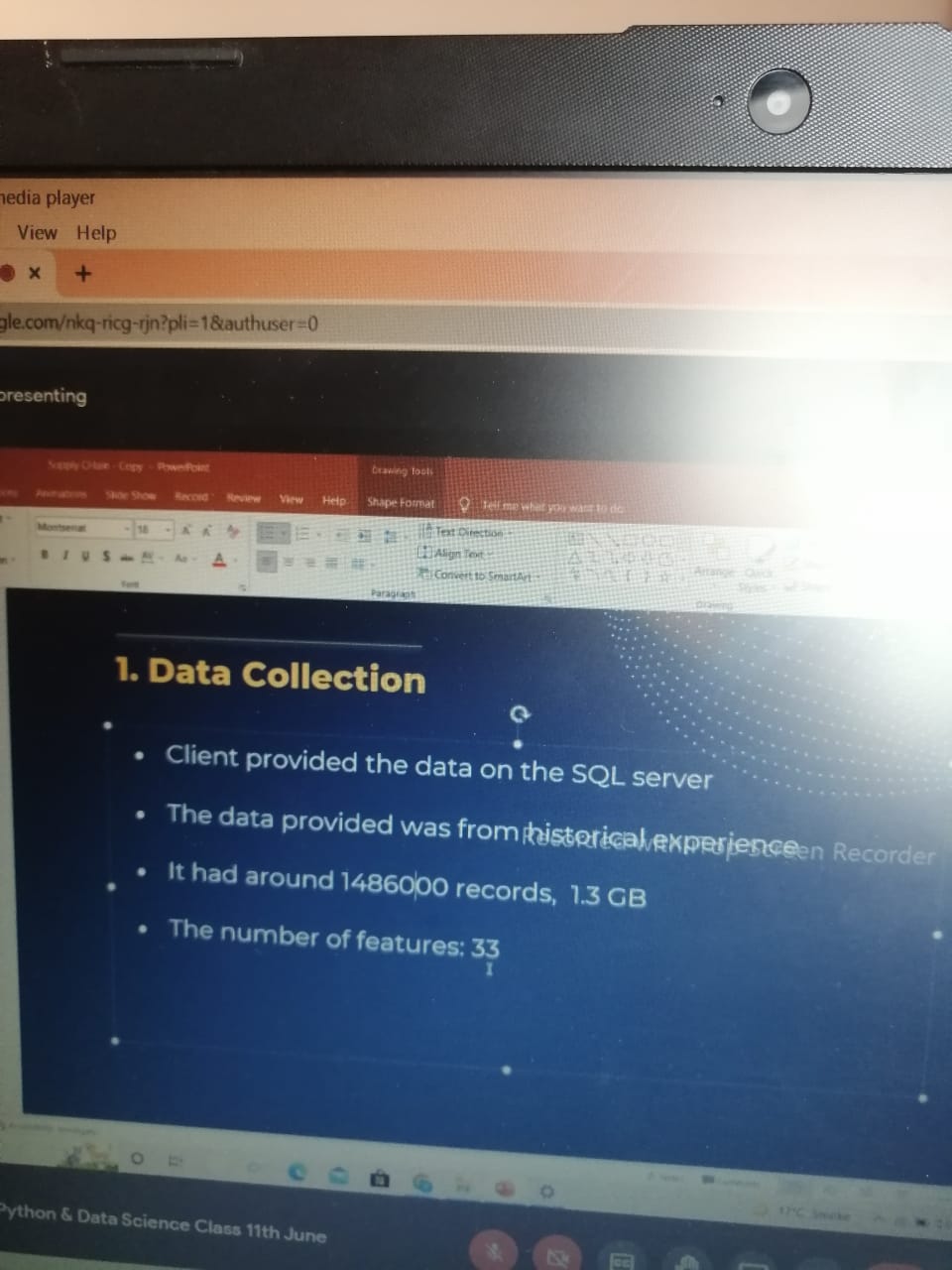
**Content writer team In our company made this problem statement**

It is desirable to develop a machine learning model that can predict customers who will leave the company. You are expected to perform the necessary data analysis and feature engineering steps before developing the model.

**POC** - proof of concept

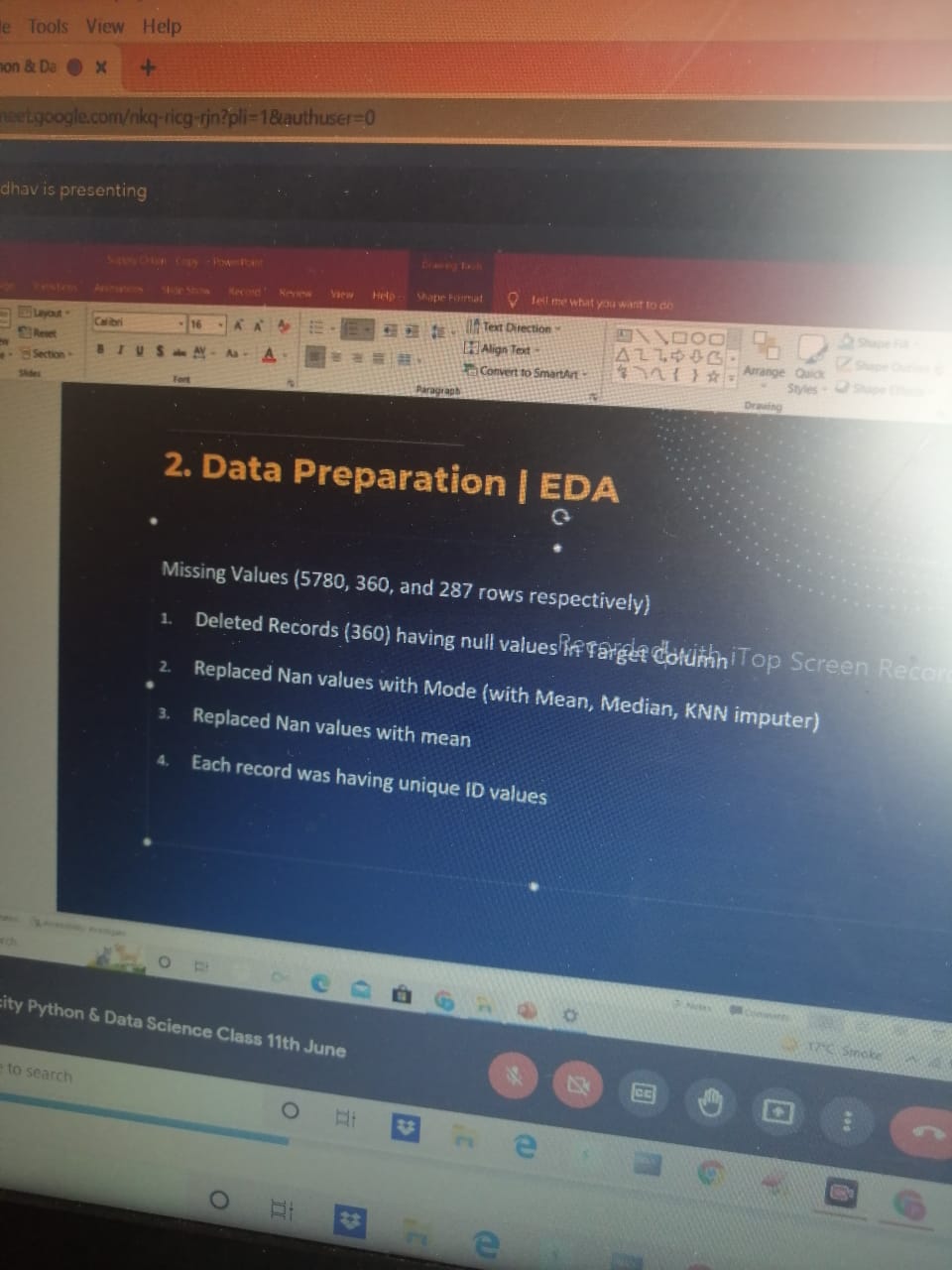
* A proof of concept (POC) is a demonstration of a product, service or solution in a sales context.
* Number of records = around 4.2 lacs and initial attributes = 48 ,finalized attributes = 25
* While POC training model we had total 48 numbers of attributes from that 17 are object type and int type = 15 and float type = 16
* It gives the names of categorical, numerical and categorical but cardinal variables in the data set.
  + Cardinal - cardinal variable (plural cardinal variables) **A variable whose values are ordered, that can be multiplied by a scalar, and for which the magnitude of differences in values is meaningful**.
* Note: Categorical variables with numerical appearance are also included in categorical variables.
* Checked Non null values replaced them with mode for categorical data and for continuous data replaced with median
* From data categorical= 14 , numerical = remaining , numerical but categorical = 5
* Attribute names
  + ['customer ID', 'gender', 'Senior Citizen', 'Partner', 'Dependents', 'tenure', 'Phone Service', 'Multiple Lines', 'Internet Service', 'Online Security', 'Online Backup', 'Device Protection', 'Tech Support', 'Streaming TV', 'Streaming Movies', 'Contract', 'Paperless Billing', 'Payment Method', 'Monthly Charges', 'Total Charges', 'Churn']
* Categorical columns -
  + ['gender', 'Partner', 'Dependents', 'Phone Service', 'Multiple Lines', 'Internet Service', 'Online Security', 'Online Backup', 'Device Protection', 'Tech Support', 'Streaming TV', 'Streaming Movies', 'Contract', 'Paperless Billing', 'Payment Method', 'Churn', 'Senior Citizen']
* Numerical cols:
  + ['tenure', 'Monthly Charges', 'Total Charges']
* Cardinal cols:
  + ['customer ID']

1. **Data Collection or Gathering –**

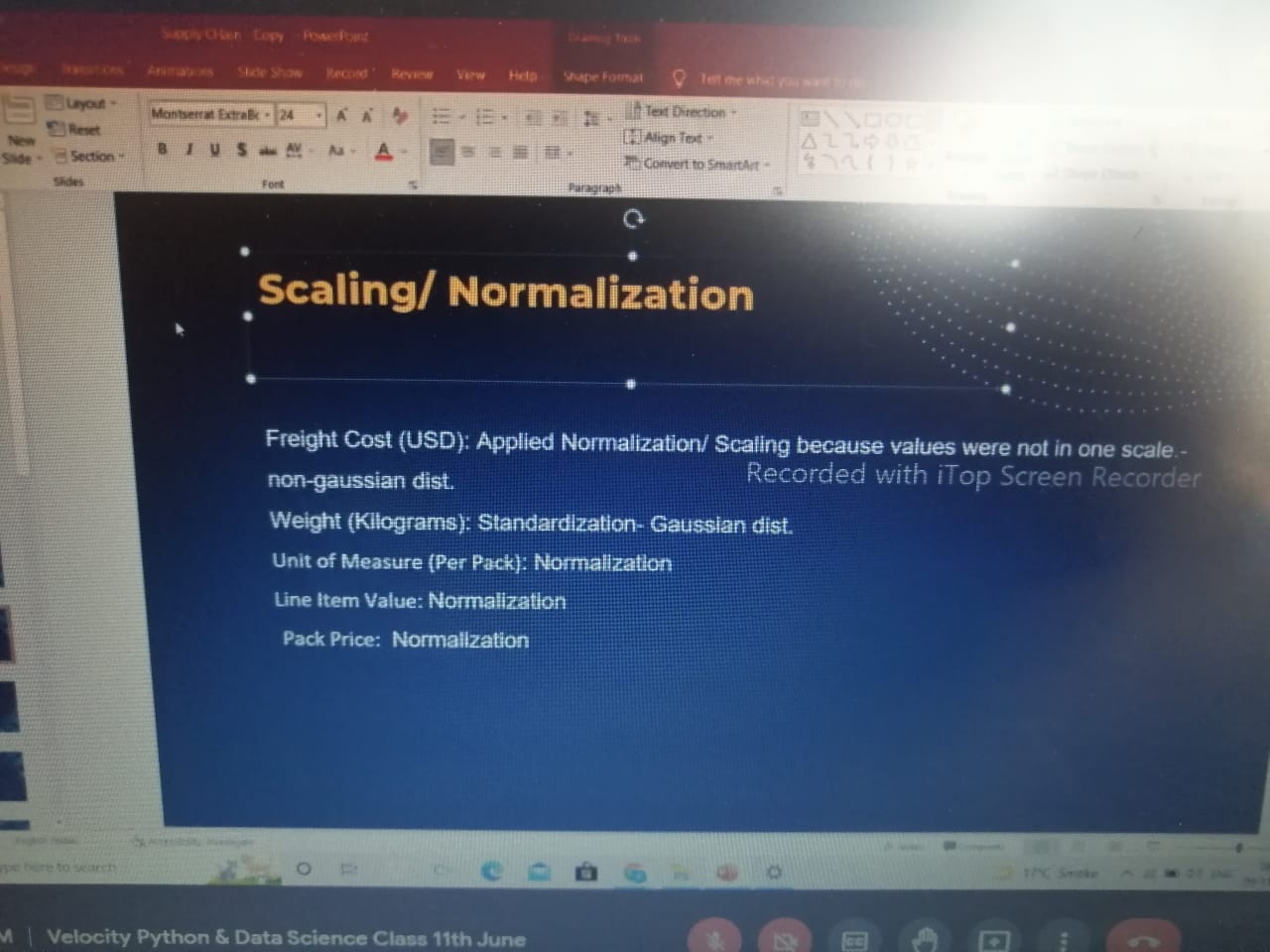
****

* **Client provided the data through SQL server**
* **I got data in CSV format from data handling team through sharepoints**
* **Data provided was from historical experience**
* **It is around 14,25,007 lacs of 1.5 GB**

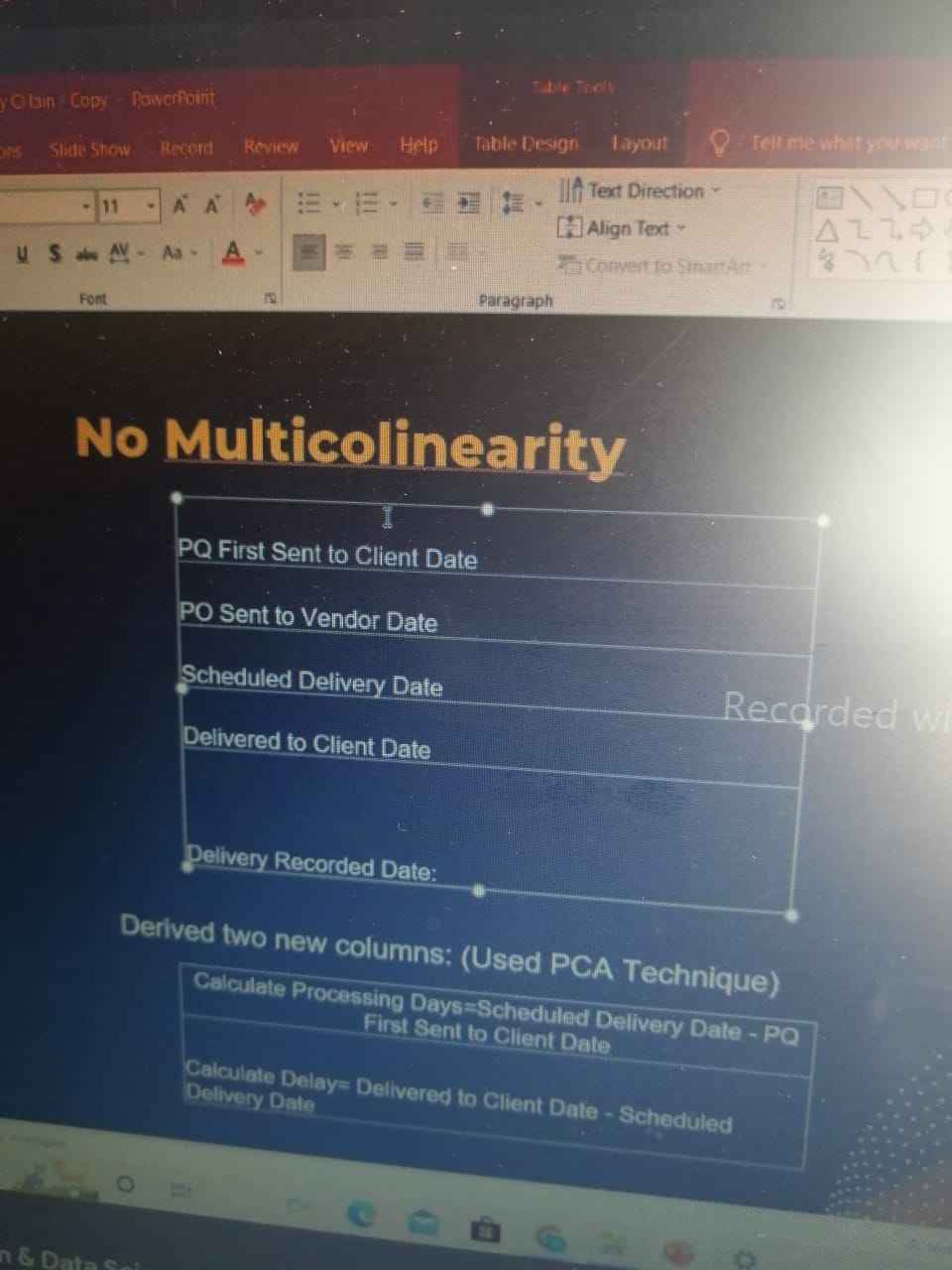
1. **EDA**

****

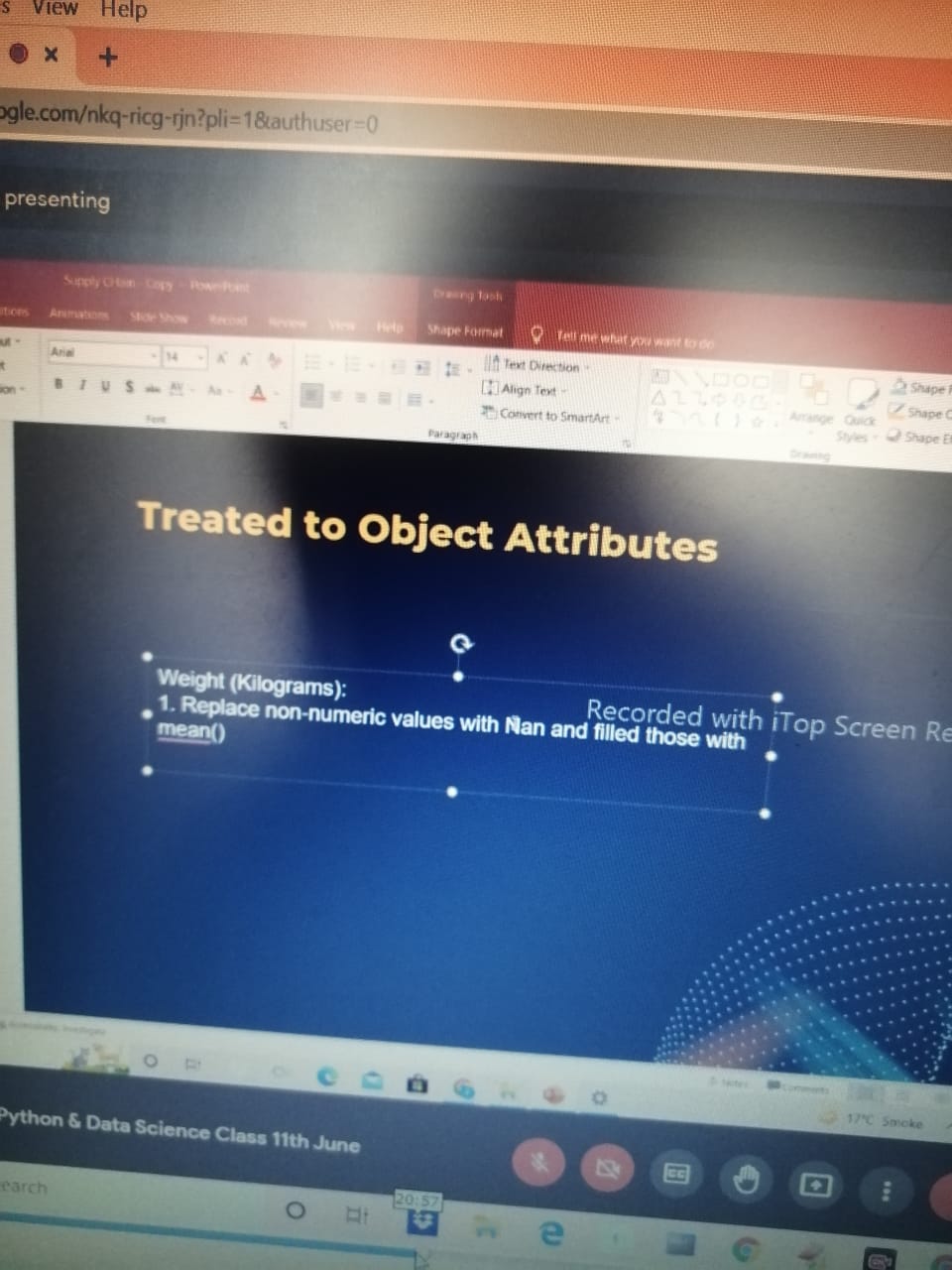
* 1. The dataset is unbalanced. So predictions will be biased towards Non Churn customers.
  2. Customers who have a Partner churned less as compared to those has'nt a Partner.
  3. Similarly churning is low who has Dependents
  4. Customers without telephone service have a lower churn rate
  5. Customers with dsl internet service have a lower churn rate
  6. Customers with protection service (Online Backup, Online Security, Device Protection), the churning rate is lower.
  7. Customers who pay monthly have a higher churn rate.
  8. Most of the customers whose payment method is electronic check have been los
* **I mostly worked in EDA part where we** 
  + **Checked variables and data types of variables** 
    - **Total 48 attributes**
    - **Object data types = 27**
    - **With high cardinality (more unique non numeric values)**
    - **One target column with unique 4 modes , imbalanced data managed by SMOTE technique**
  + **Data preparation** 
    - **Missing values handled (5775, 363, 267 rows resp.)**
    - **Deleted Records (363 having null values in Target columns )**
    - **Replaces null values with mode (mean , mode ,median , KNN imputer)**
    - **Replace nan values with mean**
    - **Each records was having unique ID values**
  + **Derived new 2 columns (using PCA technique)** 
    - **Create any two new attributes -** 
      * **1**
      * **2**
* **SCALING / NORMALIZATION**

****

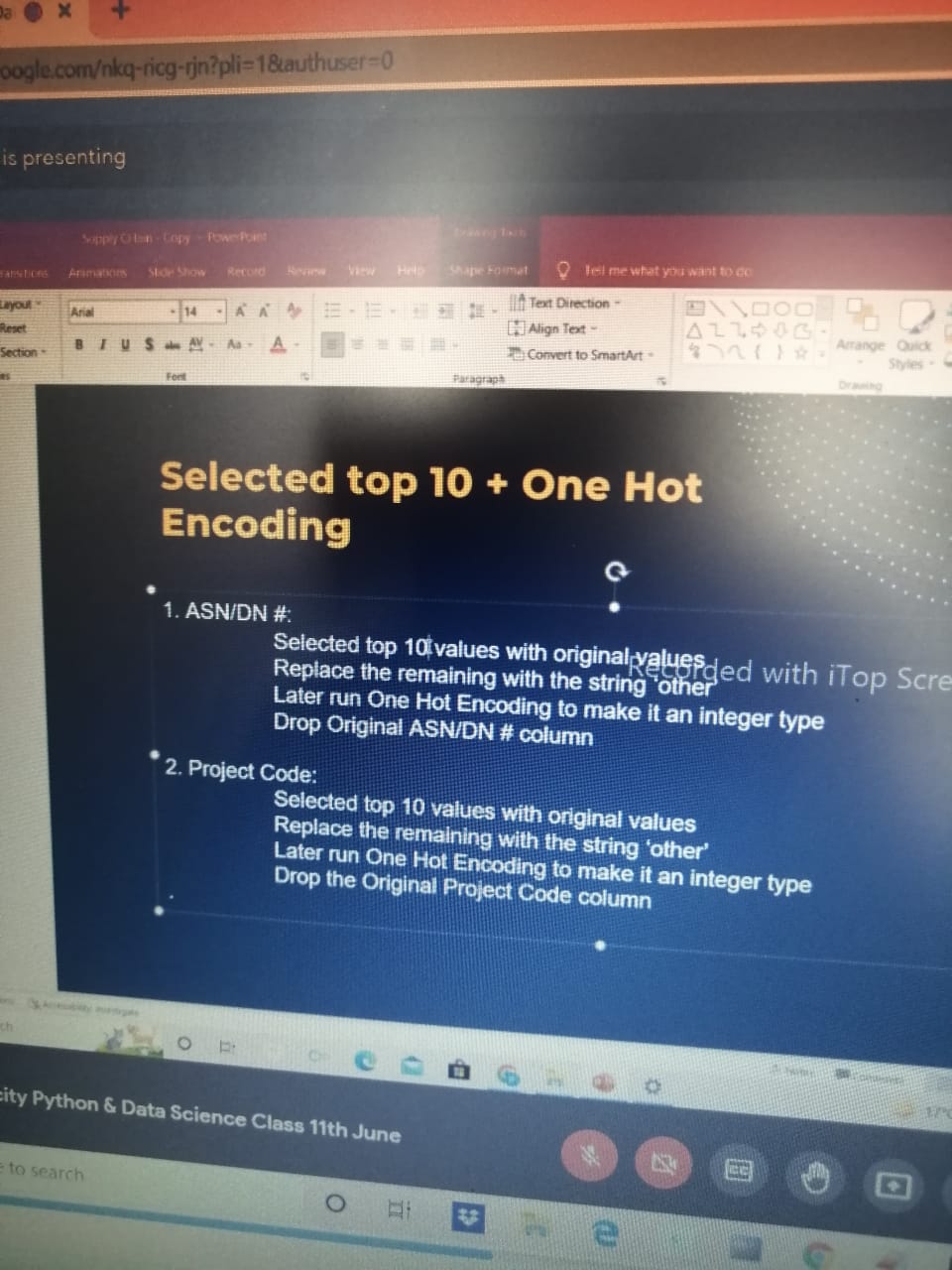
* + **Any attribute name from project (applied normalization/scaling bcz values are not in one scale)**
  + **Standardization – Gaussian Distribution –** 
    - **(Monthly Charges, New\_ Customer\_ Commitment)**
  + **Normalization -** 
    - **Choose any attribute**
* **NO multi co linearity**

****

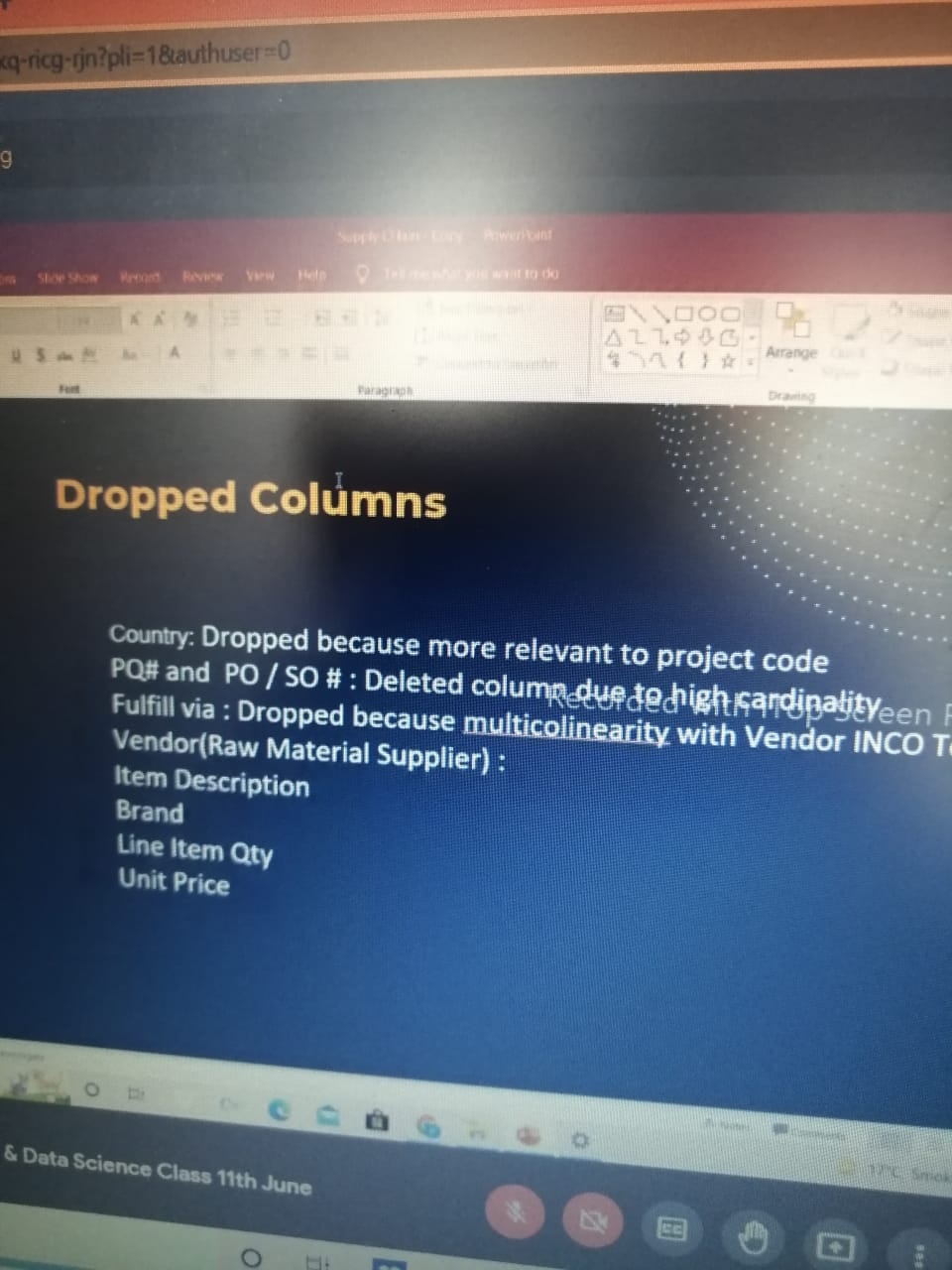
* **Treated to Object Attributes**
  + - **Replace non numeric values with nan and filled those with mode() ------------ contract**
    - **Replaced month- month ,one year ,two year with numeric values 0,1,2**

****

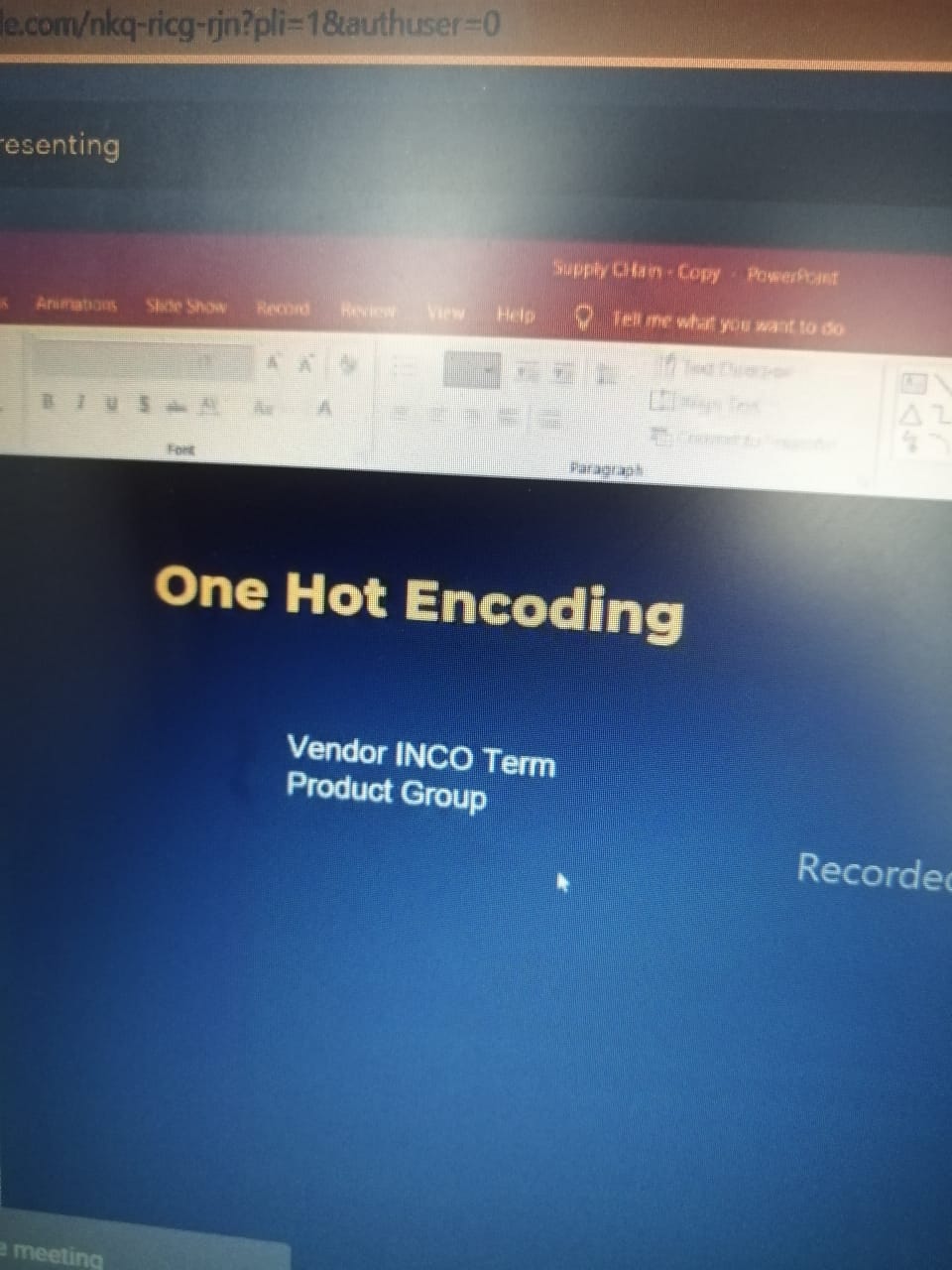
* **Selected top 10+ one hot encoding –**

****

* + **By using one hot encoding added more attributes with column name as unique values from attributes** 
    - **Columns name** 
      * **Internet service**
      * **Payment method**
* **1. Internet Service** 
  + **Selected top 10 values with original values**
  + **Replace the remaining with the string ‘other’ later run one hot encoding to make it an integer type**
  + **Drop original column internet service**
* **2.Payment method –** 
  + **Selected top 10 values with original values replace the remaining with string ‘other’ later run one hot encoding to make it an integer type**
  + **Drop original column payment method**
* **Dropped Column –**

****

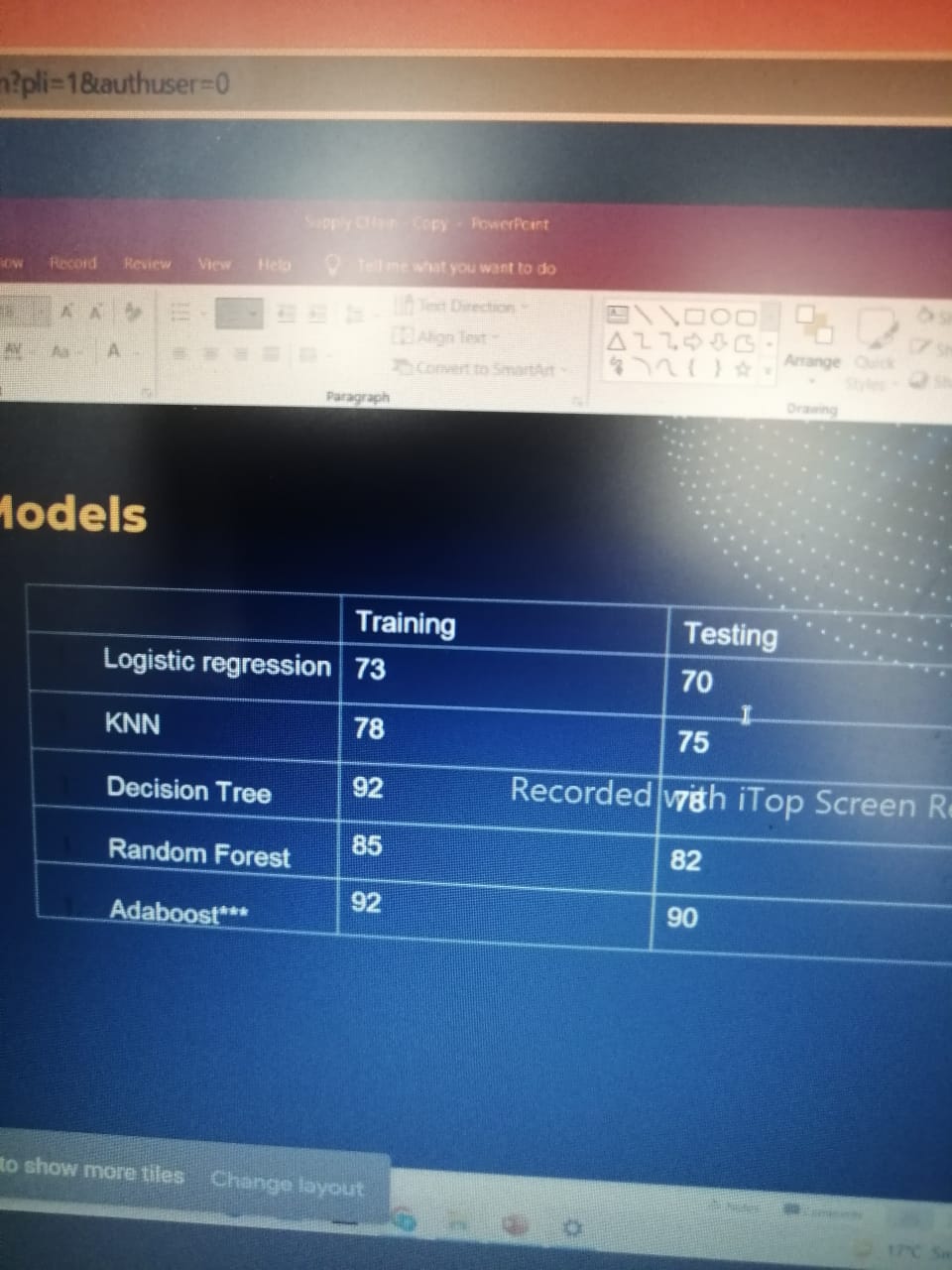
* + **Costumer ID – it has all unique values and it does not affect the prediction of target that why to reduce dimensionality we have delete costumer ID**
* **One hot encoding** 
  + **Contract**
  + **Multiple Lines**
  + **Internet Service**
  + **Online Security**

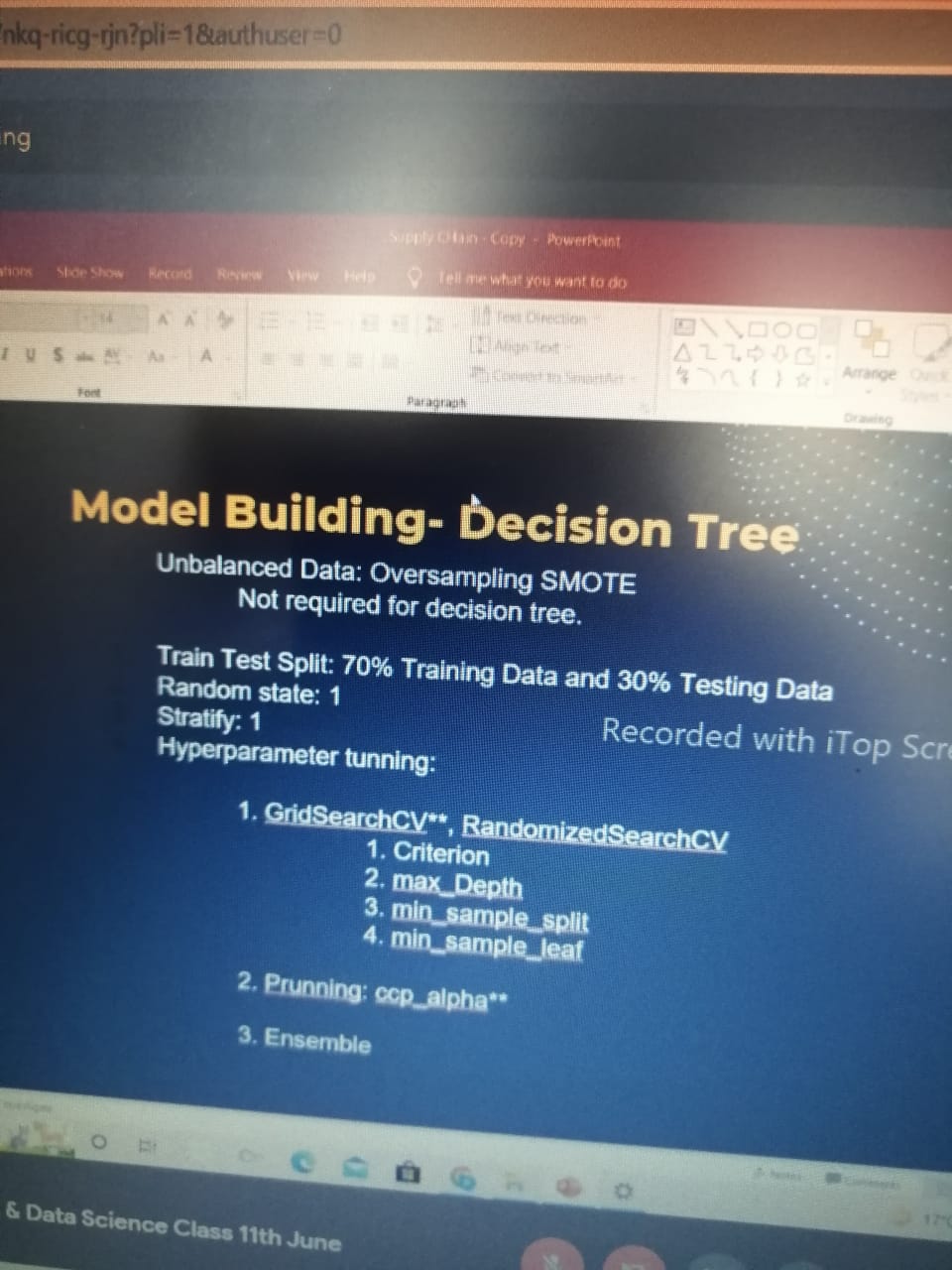
****

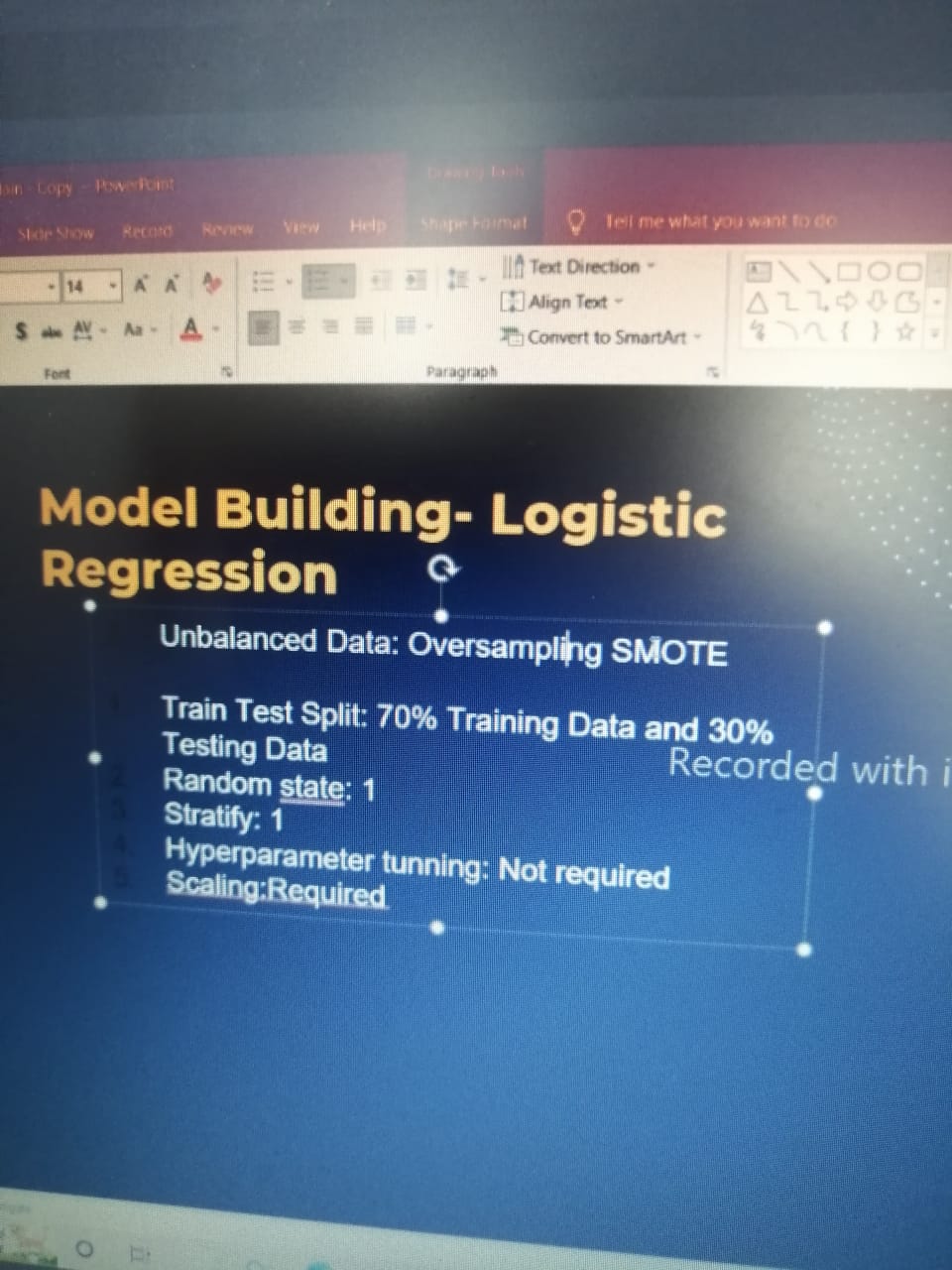
* **Models –**

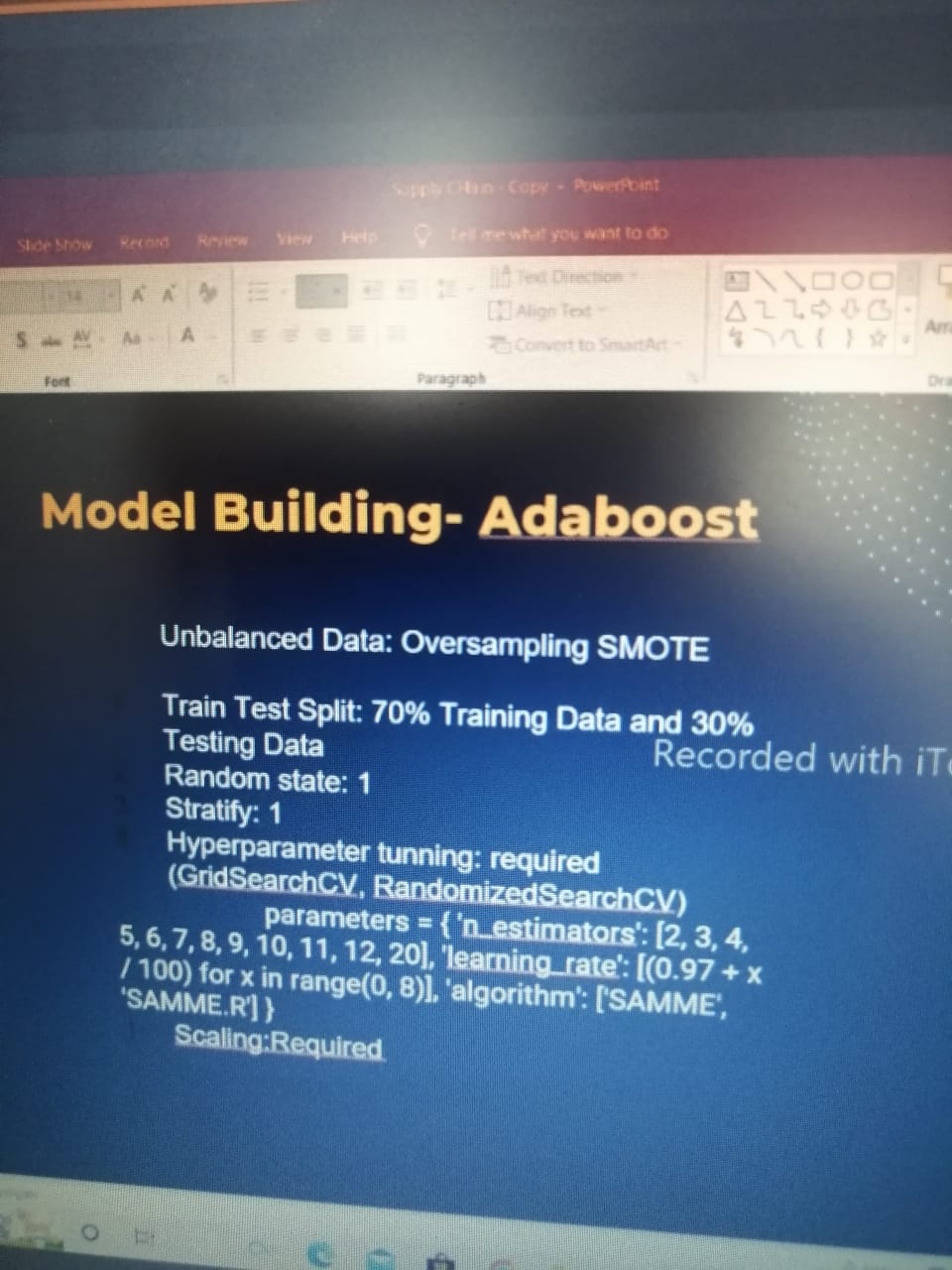
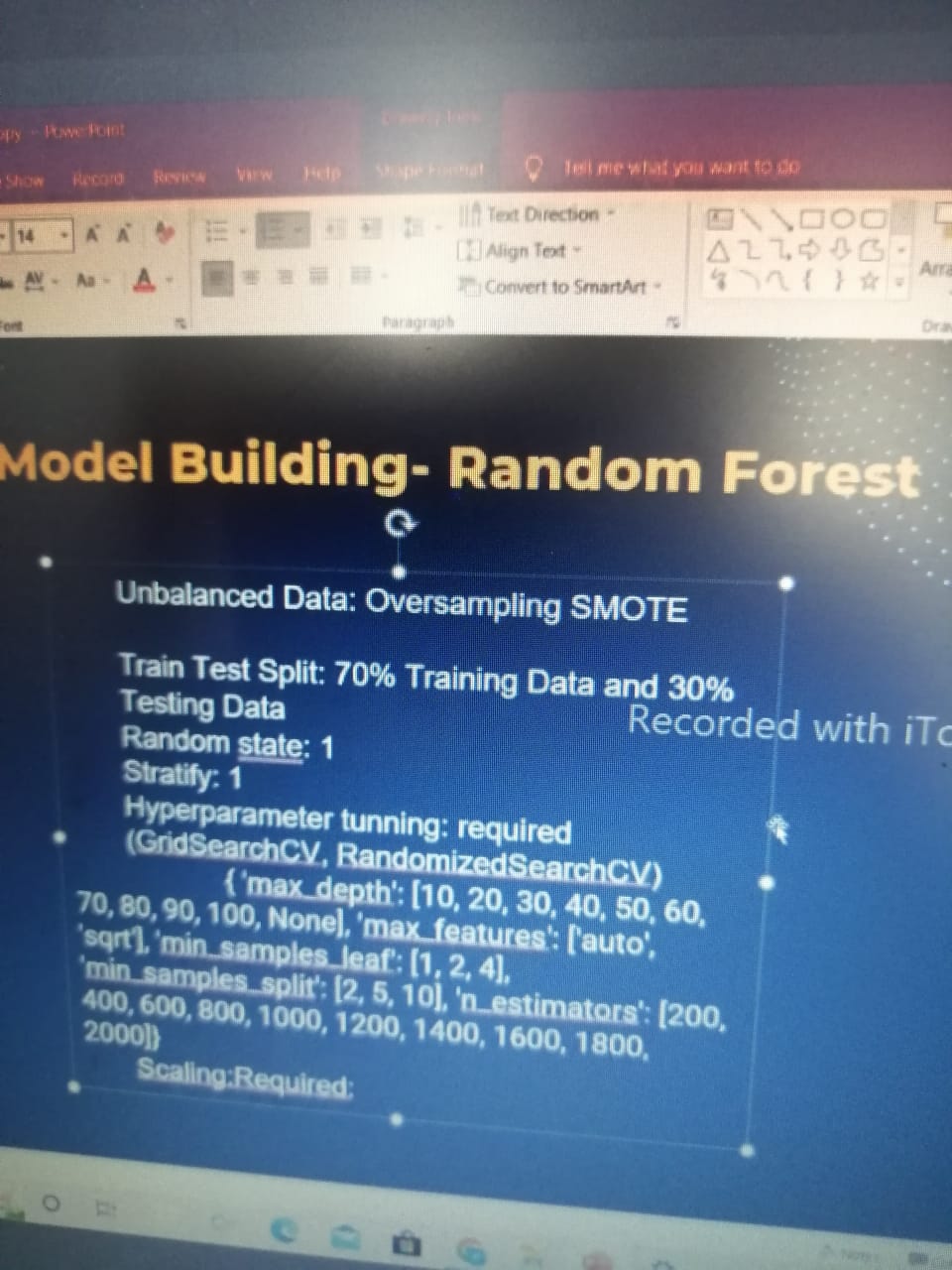
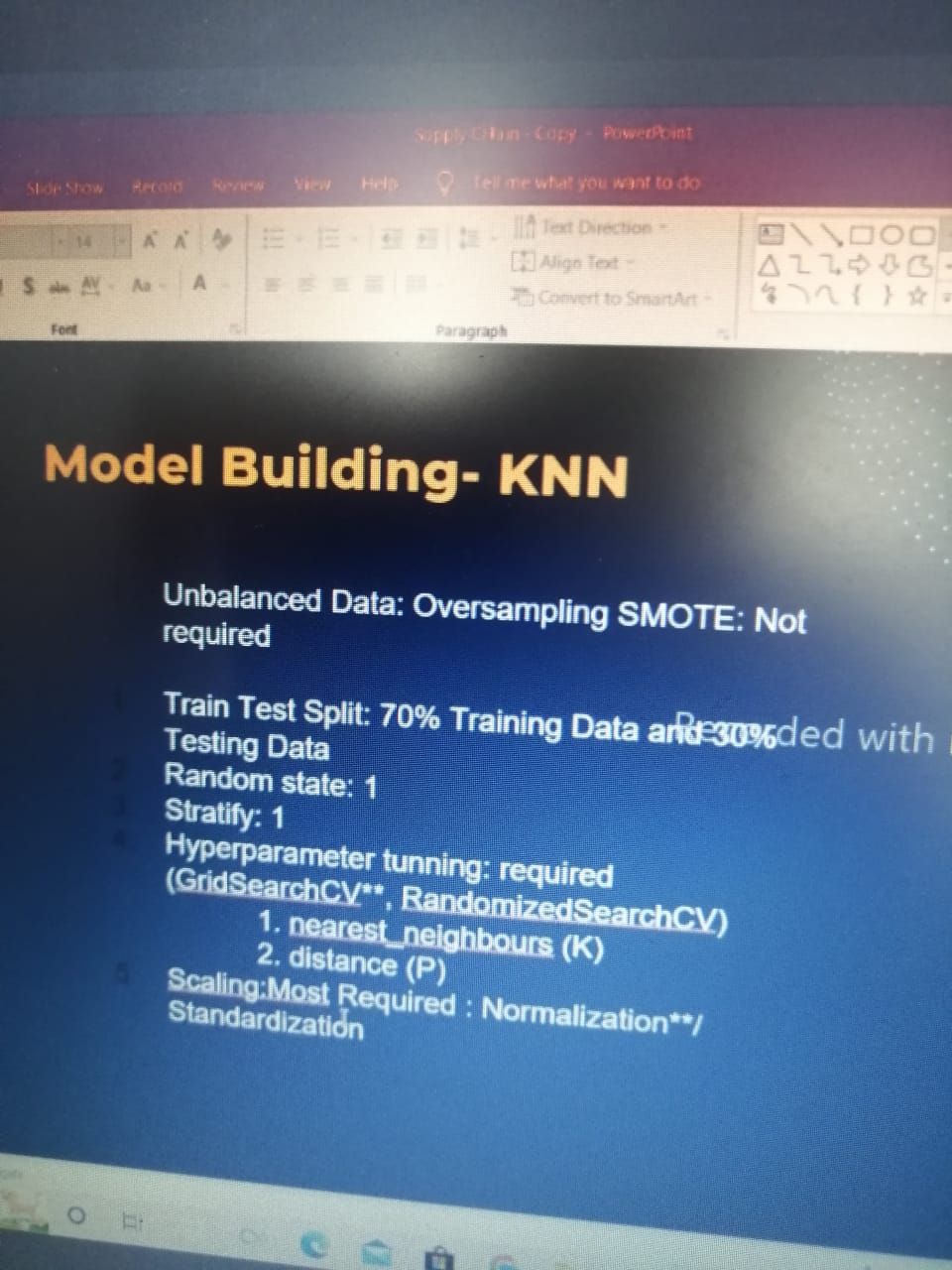
|  |  |  |
| --- | --- | --- |
| * + **Logistic regression** | **78** | * + **75** |
| * + **K nearest Neighbor** | * + **84** | * + **79** |
| * + **Random Forest** | * + **92** | * + **86** |
| * + **Decision Tree** | * + **87** | * + **80** |

**From this model random forest has highest accuracy so we finalize this model**

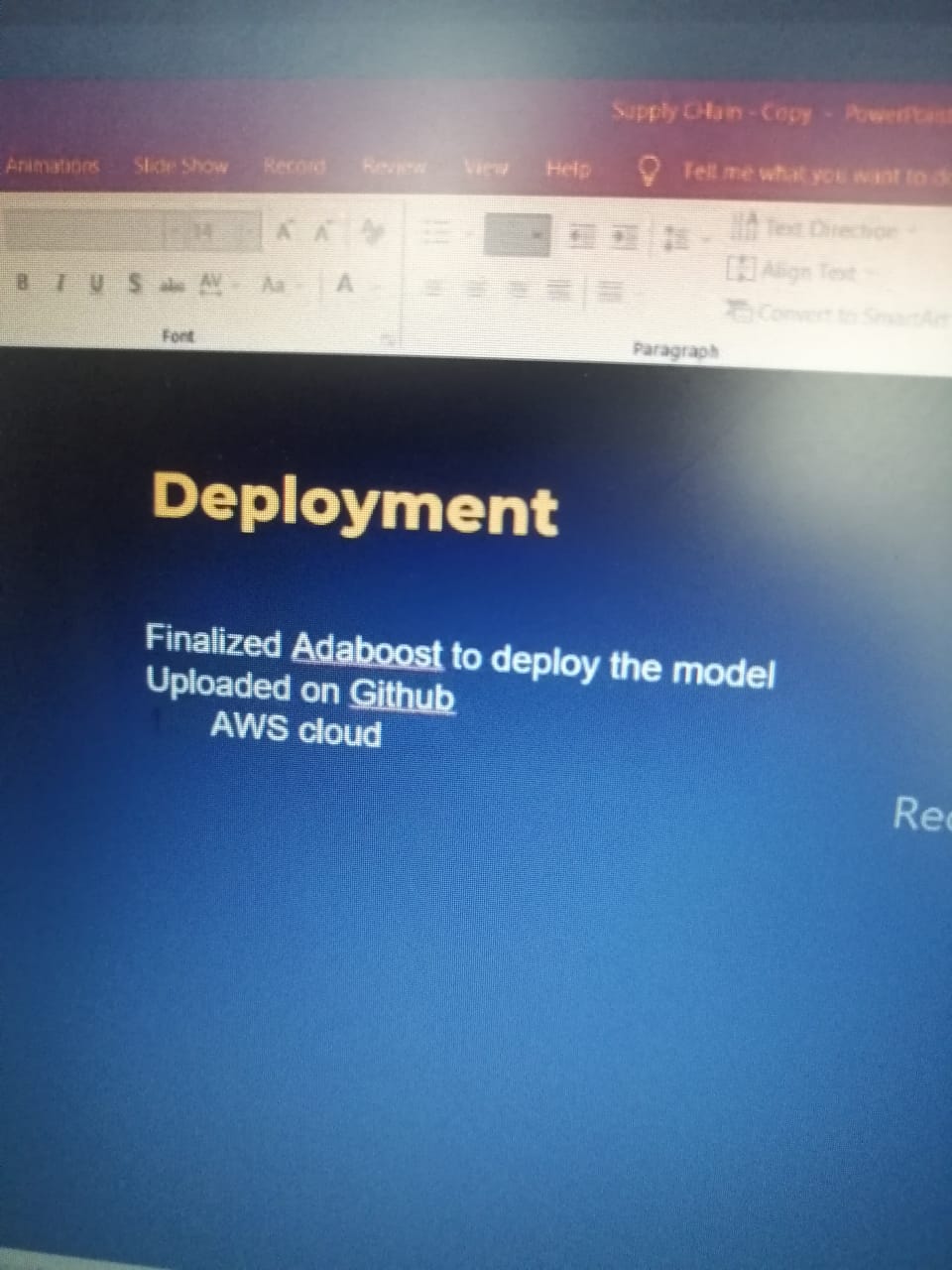
****

* ****

****

****

* **Project deployment** 
  + **Finalized Random Forest to deploy the model uploaded on Github**
  + **AWS cloud**

****