# Deliverable 3

## 1. Team

**a) Members**

1. **Vishnu Vardhan Darimidi**
2. **Bharath Kumar Gompa**
3. **Koosha Sharifani**
4. **Tharun Abhinav Suraneni**
5. **Alex Miller**

**b) Communication plan to include project artifact repository:**

Methods of communication :- e-mail, Whatsapp(once a week)
Communication response times:- Same day

Meeting attendance:- After looking into everyone's schedule we decided to meet weekly once virtually to discuss the progress of the project and complete the required things and meet the deliverable requirements.

Running meetings:- Virtual.

Git Link:- https://github.com/Vishnu-Vardhan1809/Big-Data-Project

## 2. Selection of data to analyze from:

We looked at the information and data for our project. We'll be using a dataset called "2021 Kaggle Machine Learning and Data Science survey" from a website called Kaggle\competitions. We'll study this dataset to see how it's set up and what it contains, and then use that information to figure out some results.

Dataset -Kaggle link: https://www.kaggle.com/competitions/kaggle-survey-2021/data

## 3. Business Problem or Opportunity, Domain Knowledge (link to information on domain relative to data, problem or opportunity):

With the growing rate of data, the demand to hire data scientists in the market also increases. So to get a better understanding of the people that break into the field of data and the requirement to meet today's market's expectation to deliver the project as early as possible with high accuracy, we need predictive analysis of people who pursuing data science and machine learning job and what are the certain features that some of them have that make them as more significant potential in the job market.

The dataset contains a complete industry-wide survey that represents information about job seekers including the age, level of their education, skill sets, carrier, and origin. It has around 10,000 rows and 369 columns.

**Business Problem or Opportunity:**

1. How does the level of education of job seekers in the field of data science and machine learning correlate with their salary and job opportunities
2. What are the most common career backgrounds of people who transition into data science and machine learning jobs
3. Are there any specific skills or programming languages that are more desirable for data science and machine learning jobs in the current job market.

## Input Dataset:

There are 369 columns in this dataset. Below shown are few important columns in the dataset.

1. Age
2. Gender
3. Country
4. Education level
5. Job role
6. Years of coding Experience
7. Programming language used most often
8. Computing platform used most often
9. Type of most data used often
10. Annual Compensation

### 4. Research Objectives and Question(s)

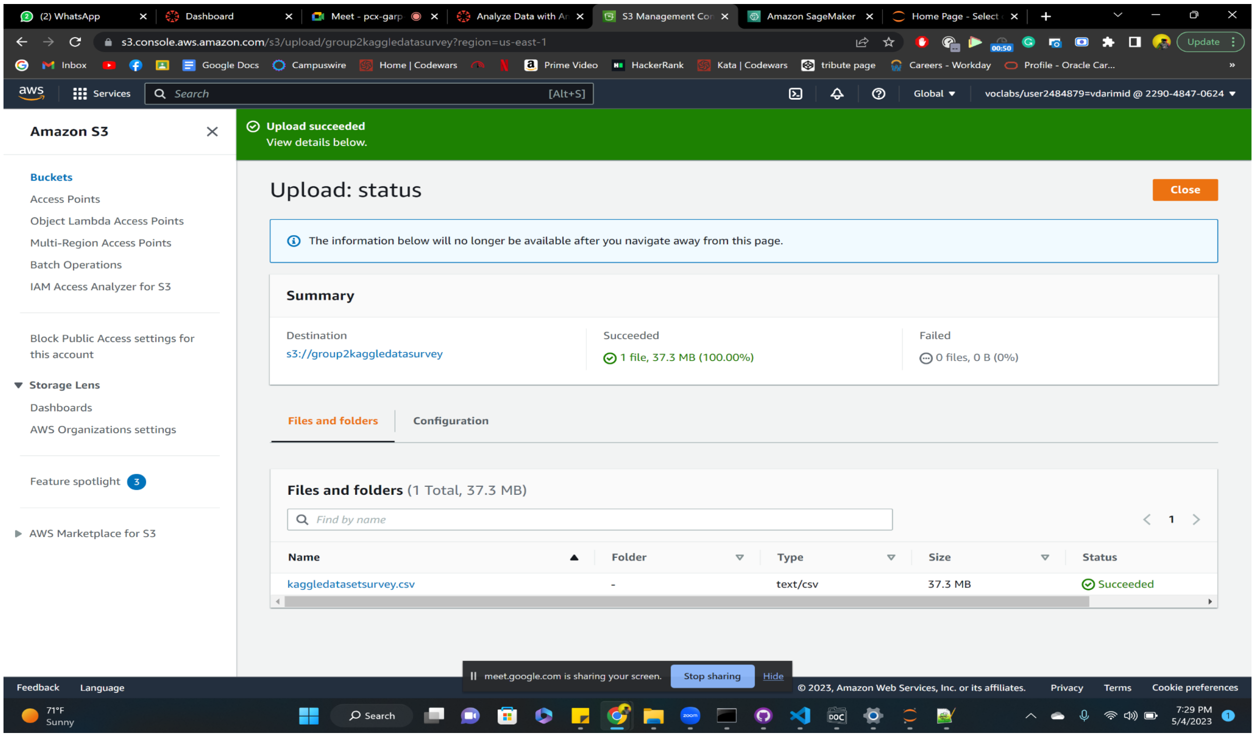
1. Analyzing the efficacy of a machine learning algorithm in forecasting job roles and selecting the algorithm that produces the most optimal results for the model.
2. Examine patterns and trends within the data science industry, and explore how various factors such as job roles, education, and compensation may be influencing these trends over time.
3. Explore the potential effects of emerging technologies and advancements in the data science industry on job roles, skills, and employment opportunities. Understand how the landscape of data science may shift in response to these changes over time.

### Analytics and Machine Learning

- Prior to developing any analytical model, the dataset was cleaned to remove missing values and special characters.
- We employed a classification technique to categorize the output column and, therefore, did not take outliers into account as they were deemed irrelevant.
- The dataset was split into training and testing sets with 60% and 40% of the data, respectively, to create two models for predicting job roles using random forest classification algorithm.
- Standard metrics such as accuracy and confusion matrix were used to evaluate the performance of these models.
- The  random forest classification algorithm. had an accuracy score of 0.5495.
- The statistical approaches utilized by these machine learning algorithms enable them to generate predictions based on the patterns and correlations they uncover within the data.

## Implementation:

The data was uploaded to an S3 bucket named group2kaggledatasurvey and was then loaded into a Jupyter notebook instance using the AWS S3 copy command. Subsequently, the data was combined into a single dataframe.



s3.console.aws.amazon.com/s3/upload/group2kaggledatasurvey?region=us-east-1

Upload succeeded
View details below.

Upload: status

The information below will no longer be available after you navigate away from this page.

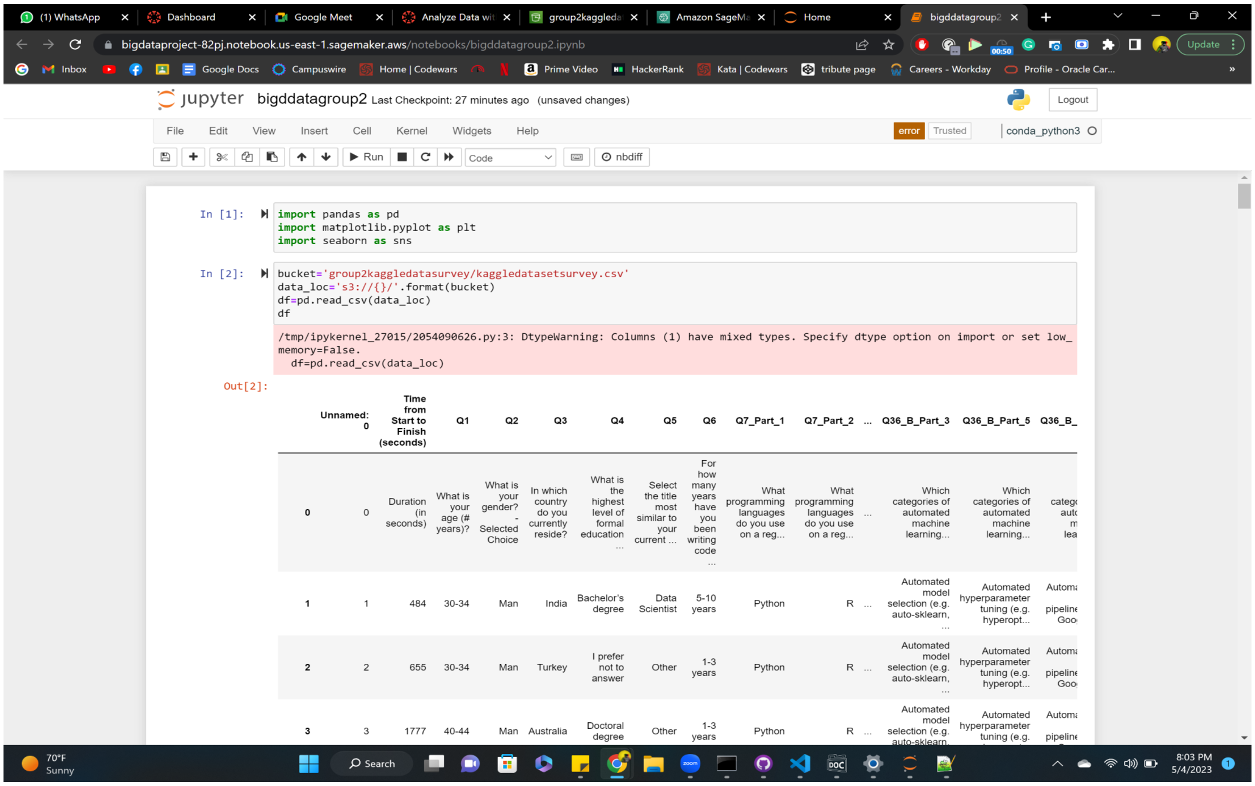
Summary

| Destination | Succeeded | Failed |
|---|---|---|
| s3://group2kaggledatasurvey | 1 file, 37.3 MB (100.00%) | 0 files, 0 B (0%) |

Files and folders (1 Total, 37.3 MB)

| Name | Folder | Type | Size | Status |
|---|---|---|---|---|
| kaggledatasetsurvey.csv | - | text/csv | 37.3 MB | Succeeded |

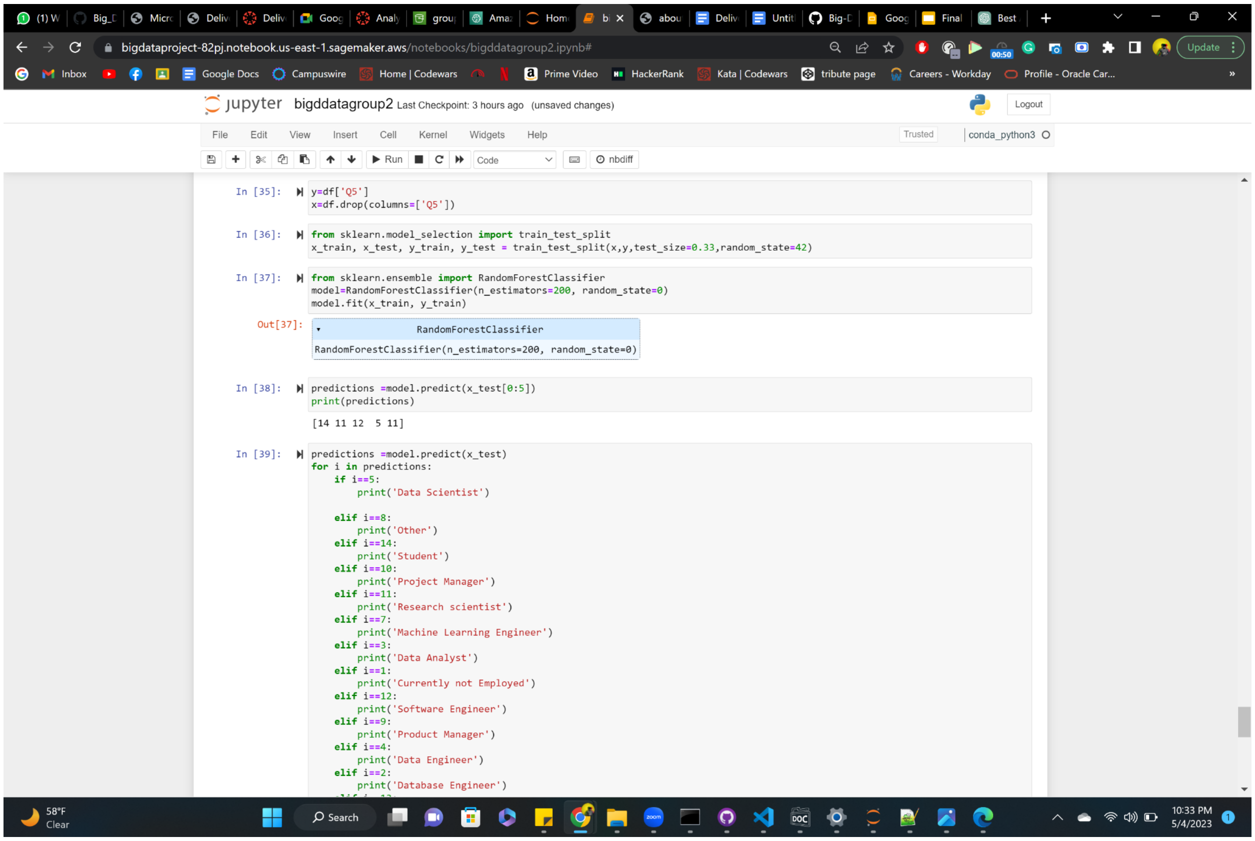
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
AWS Marketplace for S3
Close
Files and folders
Configuration
Find by name
voclabs/user2484879=vdarimid @ 2290-4847-0624
Global
meet.google.com is sharing your screen.
Stop sharing
Hide
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Feedback
Language
7:29 PM
5/4/2023


bigdataproject-82pj.notebook.us-east-1.sagemaker.aws/notebooks/bigddatagroup2.ipynb
jupyter bigddatagroup2 Last Checkpoint: 27 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
error Trusted conda_python3
Run Code nbdiff

```
In [1]:  import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns

In [2]:  bucket='group2kaggledatasurvey/kaggledatasetsurvey.csv'
         data_loc='s3://{}/'.format(bucket)
         df=pd.read_csv(data_loc)
         df

/tmp/ipykernel_27015/2054090626.py:3: DtypeWarning: Columns (1) have mixed types. Specify dtype option on import or set low_
memory=False.
  df=pd.read_csv(data_loc)

Out[2]:
```

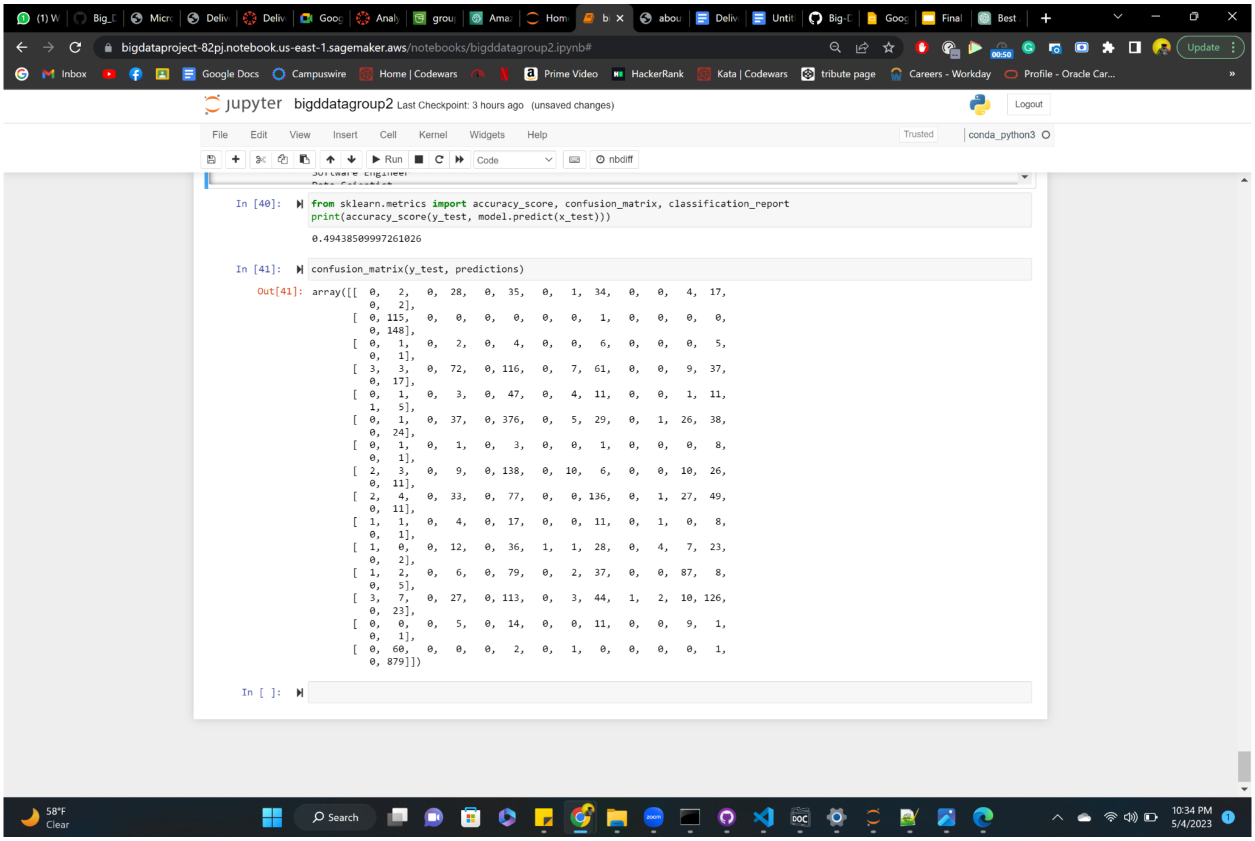
| | Unnamed: 0 | Time from Start to Finish (seconds) | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7_Part_1 | Q7_Part_2 | ... | Q36_B_Part_3 | Q36_B_Part_5 | Q36_B_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | Duration (in seconds) | What is your age (# years)? | What is your gender? - Selected Choice | In which country do you currently reside? | What is the highest level of formal education ... | Select the title most similar to your current ... | For how many years have you been writing code ... | What programming languages do you use on a reg... | What programming languages do you use on a reg... | ... | Which categories of automated machine learning... | Which categories of automated machine learning... | catego auto m lea |
| **1** | 1 | 484 | 30-34 | Man | India | Bachelor's degree | Data Scientist | 5-10 years | Python | R | ... | Automated model selection (e.g. auto-sklearn, ... | Automated hyperparameter tuning (e.g. hyperopt... | Automa pipeline Goo |
| **2** | 2 | 655 | 30-34 | Man | Turkey | I prefer not to answer | Other | 1-3 years | Python | R | ... | Automated model selection (e.g. auto-sklearn, ... | Automated hyperparameter tuning (e.g. hyperopt... | Automa pipeline Goo |
| **3** | 3 | 1777 | 40-44 | Man | Australia | Doctoral degree | Other | 1-3 years | Python | R | ... | Automated model selection (e.g. auto-sklearn | Automated hyperparameter tuning (e.g. | Automa pipeline |

We have used Random Forest Classifier model:


bigdataproject-82pj.notebook.us-east-1.sagemaker.aws/notebooks/bigddatagroup2.ipynb#
jupyter bigddatagroup2 Last Checkpoint: 3 hours ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted conda_python3
Run Code nbdiff

```
In [35]:  y=df['Q5']
          x=df.drop(columns=['Q5'])

In [36]:  from sklearn.model_selection import train_test_split
          x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.33,random_state=42)

In [37]:  from sklearn.ensemble import RandomForestClassifier
          model=RandomForestClassifier(n_estimators=200, random_state=0)
          model.fit(x_train, y_train)

Out[37]:  ▾            RandomForestClassifier
          RandomForestClassifier(n_estimators=200, random_state=0)

In [38]:  predictions =model.predict(x_test[0:5])
          print(predictions)

          [14 11 12  5 11]

In [39]:  predictions =model.predict(x_test)
          for i in predictions:
              if i==5:
                  print('Data Scientist')

              elif i==8:
                  print('Other')
              elif i==14:
                  print('Student')
              elif i==10:
                  print('Project Manager')
              elif i==11:
                  print('Research scientist')
              elif i==7:
                  print('Machine Learning Engineer')
              elif i==3:
                  print('Data Analyst')
              elif i==1:
                  print('Currently not Employed')
              elif i==12:
                  print('Software Engineer')
              elif i==9:
                  print('Product Manager')
              elif i==4:
                  print('Data Engineer')
              elif i==2:
                  print('Database Engineer')
```

We have calculated accuracy and Confusion matrix for the above model.



```
In [40]: from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
         print(accuracy_score(y_test, model.predict(x_test)))

         0.49438509997261026
```

```
In [41]: confusion_matrix(y_test, predictions)

Out[41]: array([[  0,   2,   0,  28,   0,  35,   0,   1,  34,   0,   0,   4,  17,
                   0,   2],
                [  0, 115,   0,   0,   0,   0,   0,   0,   1,   0,   0,   0,   0,
                   0, 148],
                [  0,   1,   0,   2,   0,   4,   0,   0,   6,   0,   0,   0,   5,
                   0,   1],
                [  3,   3,   0,  72,   0, 116,   0,   7,  61,   0,   0,   9,  37,
                   0,  17],
                [  0,   1,   0,   3,   0,  47,   0,   4,  11,   0,   0,   1,  11,
                   1,   5],
                [  0,   1,   0,  37,   0, 376,   0,   5,  29,   0,   1,  26,  38,
                   0,  24],
                [  0,   1,   0,   1,   0,   3,   0,   0,   1,   0,   0,   0,   8,
                   0,   1],
                [  2,   3,   0,   9,   0, 138,   0,  10,   6,   0,   0,  10,  26,
                   0,  11],
                [  2,   4,   0,  33,   0,  77,   0,   0, 136,   0,   1,  27,  49,
                   0,  11],
                [  1,   1,   0,   4,   0,  17,   0,   0,  11,   0,   1,   0,   8,
                   0,   1],
                [  1,   0,   0,  12,   0,  36,   1,   1,  28,   0,   4,   7,  23,
                   0,   2],
                [  1,   2,   0,   6,   0,  79,   0,   2,  37,   0,   0,  87,   8,
                   0,   5],
                [  3,   7,   0,  27,   0, 113,   0,   3,  44,   1,   2,  10, 126,
                   0,  23],
                [  0,   0,   0,   5,   0,  14,   0,   0,  11,   0,   0,   9,   1,
                   0,   1],
                [  0,  60,   0,   0,   0,   2,   0,   1,   0,   0,   0,   0,   1,
                   0, 879]])

In [ ]:
```

## Evaluation and Optimization:

- Accuracy scores and confusion matrix were used to evaluate the performance of the models. The accuracy metric assesses the proportion of correct predictions, while the confusion matrix presents a summary of the model's prediction results.

- For 40% of the testing data, the model's accuracy score of 0.4943 indicates that it is not accurately identifying the cover type.

## Results:

Accuracy:0.4943 and Confusion Matrix.

## Future work, Comments:

### 1) What was unique about the data? Did you have to deal with imbalance? What data cleaning did you do? Outlier treatment? Imputation?

The dataset that we are dealing with is conducted an industry-wide survey that presents a truly comprehensive view of the state of data science and machine learning. What we did as a research team was utilize these data in order to do both descriptive and predictive analysis. As we already know, descriptive analytics tells what happened in your business in the past week, month, or year. On the other hand, Predictive analytics determines the potential outcomes of present and past actions and trends. Some of the features in your dataset, such as age, gender, country, education level, and current role, are descriptive because they reflect on things past consumer behavior. Some of the features, such as programming languages and years of coding experience, are predictive because they can help assess the likelihood of a certain event happening in the future, such as getting a job role as a data scientist.

The unique thing about the dataset is that it is relevant to a wide variety of industries, including genetics, health sciences, economics, chemometrics, sociological surveys, environmental sciences, and finance. Furthermore, The features are flexible for both descriptive and predictive analysis depending on our goal. For example, you can use descriptive analytics to compare the demographics and skills of different job roles in your dataset. You can also use predictive analytics to forecast the demand and supply of different job roles based on the trends and patterns in your dataset. Here we are trying to predict the job role based on the remaining factors like languages they know and degree they achieved and age and gender.

After cleaning the data and preparing it for preprocessing, one of the things that we considered was whether the data is imbalanced or not. Imbalanced data refers to a situation, where one target class represents a significant portion of observations. Also, There are different ways to deal with imbalanced data, depending on the context such as Down sampling, Up sampling, and Using different metrics. For our dataset, we were dealing with imbalanced data which crate problems when we try to predict any field in the data frame and specifically the columns related to program languages that our users know had some issues with imbalanced data. Balanced data will help the model to predict accurately. Therefore, we used a resampling method called Synthetic Minority Oversampling Technique to address this issue. Our dataset mostly consisted of object data types. But one column contains integer data but it shows the object type in its property. So processed the particular field by changing the type of the column. Null values are replaced by the mode of the column. The rows which consist of maximum null values are deleted.

**2). Did you create any new additional features / variables?**

One of the ways to deal with imbalanced data was to create additional variables and columns but it would make multicollinearity or redundancy in your data or an overfitting problem with the final prediction. In the same way, we did not apply any outlier treatment because our target column is a categorized column so outlier treatment is not needed. Overall, it is important to carefully consider data preprocessing and handling techniques when dealing with imbalanced datasets in order to ensure accurate model predictions.

**3) What was the process you used for evaluation? What was the best result?**

For the project, the process we used for evaluation using was the random forest classifier which starts by splitting the dataset into three subsets: train, validation, and test. For the ratio, we used an 80/20/10 split ratio for the train/validation/test sets. The reason that we choose a random forest classifier is that it uses averaging to improve the predictive accuracy and control over-fitting. It can handle both categorical and numerical features, and it is robust to imbalanced data and outliers. It performs well with imbalanced data and can handle non-linear relationships between the features. In the end, after creating and setting the model we evaluated the model performance on the test set using various metrics. I used accuracy, precision, recall, F1-score, ROC curve, and PR curve to measure how well the model can predict the job role of new data points.

**4) What were the problems you faced? How did you solve them?**

During the project we all faced a few challenges in the preprocessing phase. Deciding the threshold value to delete rows and fitting data with the right estimates is a challenging task. Furthermore, handling imbalanced data can be difficult as it requires careful consideration of various techniques such as resampling, boosting, and bagging.

**5) What future work would you like to do?**

for future work, we could explore other machine learning algorithms and techniques specifically designed to handle imbalanced data or use different features to predict other variables related to job roles or entrepreneurial competency

**6) Instructions for individuals that may want to use your work**

In order to obtain comparable outcomes, individuals interested in utilizing our work should consider the following guidelines:

Install the essential Python libraries, including Numpy, Pandas, Matplotlib, Seaborn, and Scikit-learn, that are widely used in this project

Ensure that the dataset is thoroughly cleaned, transformed, and prepared to eliminate the majority of the outliers. This will result in the development of accurate machine learning models and visualizations