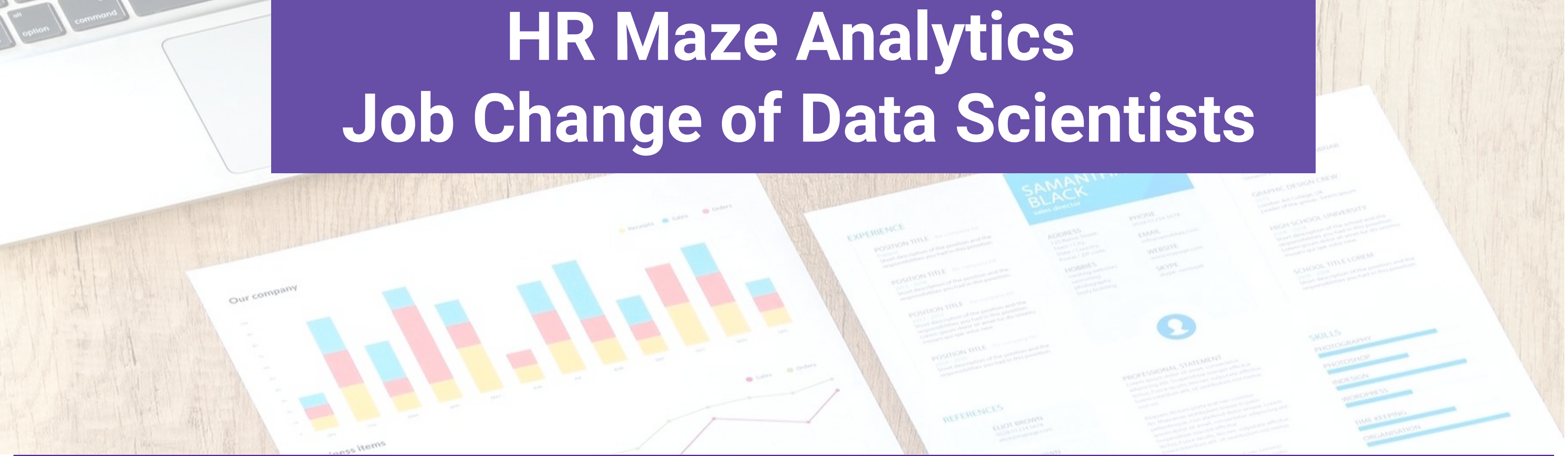


HR Maze Analytics
Job Change of Data Scientists

Data Science and Big Data Course

Spring 2024

Professor : Antonino Nocera

Team Work : Kamila Aburwais , Sepideh Hayati, Sahar Taheri

Imagine you're navigating through a maze of career opportunities....!!!

Content Table

- Project Overview
- Methodology and Architecture
- Analysis & Testing
- Limitation and Challenge
- Conclusion
- References

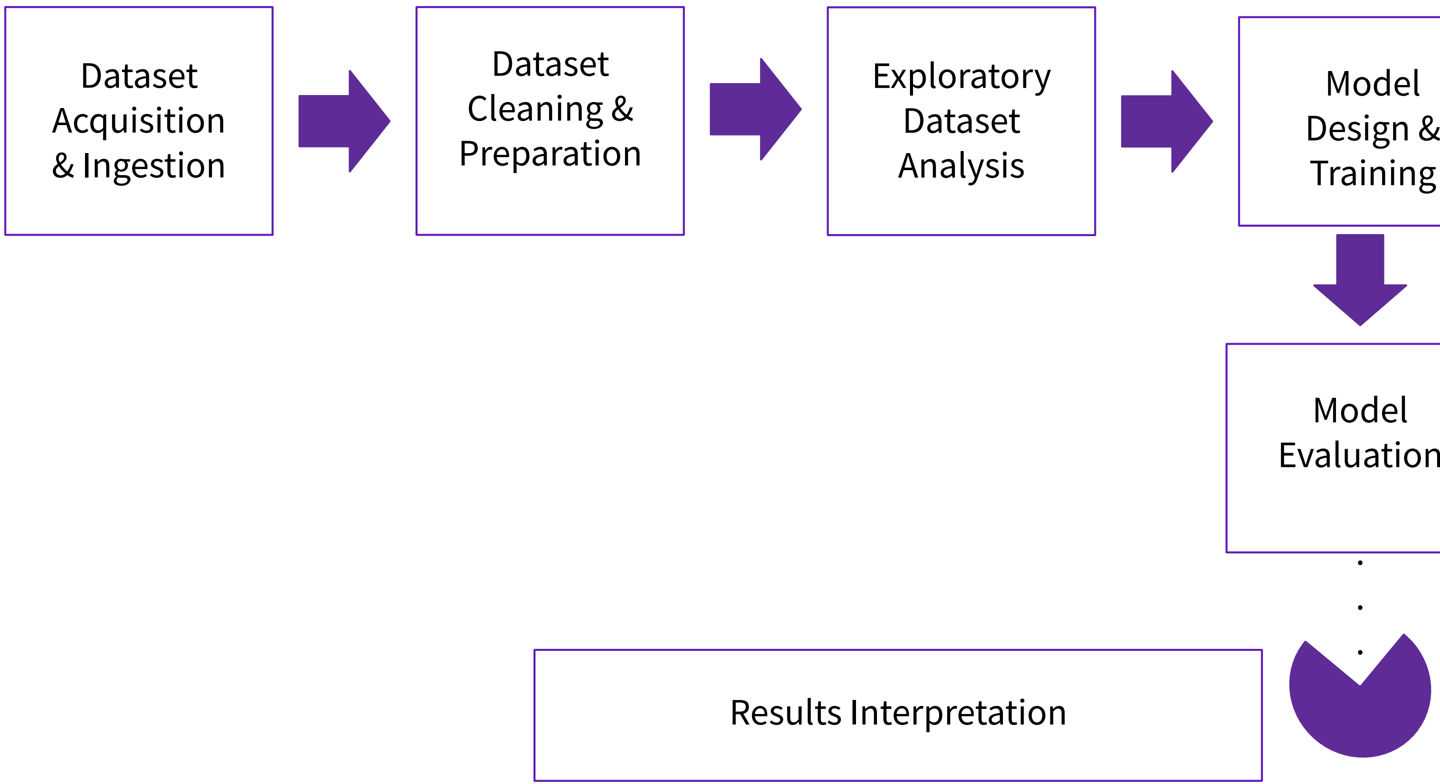
Introduction

This project rapidly evolving landscape of Big Data and Data Science, titled **"HR Maze Analytics"** delves into predicting job change intentions among candidates who have completed specialized training in data science. By leveraging machine learning algorithms and advanced analytics, the aim is to unravel the intricate interplay between educational background, professional experience, and career aspirations within the context of data science careers.

Project Objectives

1. Optimize the Influence of Education/Experience on Job Change Decisions,
2. Leverage Big Data Tools for Scalable Data Processing,
3. Investigate Key Hypotheses on Job-Seeking Behavior,
4. Build and Evaluate Predictive Models.

Project Workflow



Methodology & Architecture

Dataset Overview

HR Analytics: Job Change of Data Scientists

- Rows 2124
- Columns 13
- Usability 10
- Target predicting job changing

Data collection purpose:
A Big Data and Data Science company wants to identify which candidates from their training courses intend to work for the company versus those seeking new employment.

- To optimize training costs and quality.
- The dataset includes demographics, education, experience information, and …

Goal
To use this data to predict candidates' job-seeking intentions and understand the factors influencing these decisions

Tool and Techniques

- Virtual MachineBox Enviroment
- Hadoop (Mapper & Reducer)
- Spark
- Python Language
  - Pandas, Seaborn Libraries Python Language to Visualization
- Mongodb/PyMongoDb
- Machine Learning Models

Initial Hypotheses

#H1:

Candidates with longer experience are less likely to seek new job opportunities.

#H2:

Candidates enrolled in university courses are more likely to seek new job opportunities.

Reading Data

import findspark

findspark.init()

import pyspark

from pyspark.sql import SparkSession

spark = SparkSession.builder \
.master("local[1]") \ .appName("PySpark Read CSV and Convert to JSON") \
.getOrCreate()

csv_file = '/home/ubuntu/aug_test.csv'

df = spark.read.csv(csv_file, inferSchema=True, header=False)

df.show()

Results

```
+-----+--------+------------------+------+--------------------+---------------+--------+-----+---+---------+--------------+----+-----+
|  _c0|     _c1|               _c2|   _c3|                 _c4|            _c5|     _c6|  _c7|_c8|      _c9|          _c10|_c11| _c12|
+-----+--------+------------------+------+--------------------+---------------+--------+-----+---+---------+--------------+----+-----+
|10021| city_16|              0.91|Female|Has relevent expe...|  no_enrollment|Graduate| STEM|  5|   10000+|       Pvt Ltd|   1| 22\t|
|10049|city_160|              0.92|  Male|Has relevent expe...|  no_enrollment|Graduate| STEM|>20|  100-500|       Pvt Ltd|   1| 20\t|
|10050|city_103|              0.92|  Male|Has relevent expe...|  no_enrollment| Masters| STEM| 10|   10000+|       Pvt Ltd|  >4| 20\t|
|10162|city_160|              0.92|Female|Has relevent expe...|  no_enrollment|Graduate| STEM| 15|  500-999|       Pvt Ltd|   3| 29\t|
|10167|city_103|              0.92|Female|No relevent exper...|Part time course|Graduate| STEM|  2|    50-99|Funded Startup|   1| 35\t|
|10171| city_61|0.9129999999999999|  Male|Has relevent expe...|  no_enrollment| Masters| STEM|>20|    10/49|       Pvt Ltd|  >4| 40\t|
|10198|city_105|             0.794|  Male|Has relevent expe...|Full time course|Graduate| STEM| 11|  100-500|       Pvt Ltd|  >4|  3\t|
|10217|city_159|             0.843|  Male|Has relevent expe...|  no_enrollment|Graduate| STEM|  8|    50-99|       Pvt Ltd|   1| 33\t|
|10230|city_103|              0.92|  Male|Has relevent expe...|  no_enrollment| Masters| STEM|>20|  500-999|Funded Startup|   1| 60\t|
|10246|city_160|              0.92|  Male|Has relevent expe...|  no_enrollment| Masters| STEM| 17|     null|          null|   1|  9\t|
| 1026|city_162|             0.767|  null|Has relevent expe...|  no_enrollment| Masters| STEM| 11|1000-4999|       Pvt Ltd|   3| 62\t|
|10260| city_21|             0.624|  null|No relevent exper...|Part time course| Masters| STEM|  2|     null|          null|   2| 32\t|
|10279| city_71|             0.884|  null|No relevent exper...|Full time course|Graduate| STEM|  6|     null|          null|never| 45\t|
|10287|city_160|              0.92|  Male|Has relevent expe...|  no_enrollment|Graduate| STEM| 13|  500-999|       Pvt Ltd|   1| 20\t|
|10304|city_103|              0.92|  Male|Has relevent expe...|  no_enrollment| Masters| STEM| 10|     null| Public Sector|   1| 31\t|
|10308|city_160|              0.92|Female|No relevent exper...|  no_enrollment|Graduate| STEM|  1|  100-500|           NGO|   1|  8\t|
|10311| city_67|             0.855|  Male|Has relevent expe...|Full time course|Graduate| STEM| 11|    10/49|          null|   1|106\t|
|10324| city_16|              0.91|  Male|Has relevent expe...|Part time course|Graduate| STEM|>20|   10000+|       Pvt Ltd|   2| 12\t|
|10348|city_103|              0.92|  Male|No relevent exper...|  no_enrollment|Graduate| STEM| 18|    50-99|       Pvt Ltd|  >4| 28\t|
|10394|city_103|              0.92|  Male|Has relevent expe...|  no_enrollment|Graduate|Other| 15|    50-99|Funded Startup|   1| 10\t|
+-----+--------+------------------+------+--------------------+---------------+--------+-----+---+---------+--------------+----+-----+
only showing top 20 rows
```

Cleaning Data

from pyspark.sql.functions import round,when,regexp_replace, col

df_cleaned = df1.fillna("Others")

df_cleaned = df_cleaned.withColumn("experience", when(df_cleaned["experience"] == ">20", "21") .when(df_cleaned["experience"] == "<1","0").otherwise(regexp_replace(df_cleaned["experience"], "[^0-9]", "")))

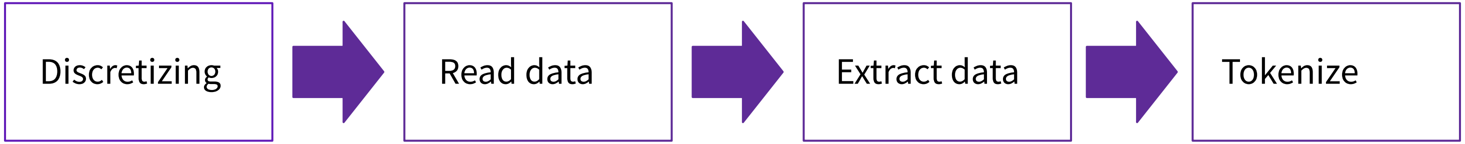
df_cleaned = df_cleaned.withColumn("last_new_job", when(col("last_new_job") == ">4", "5") .otherwise(col("last_new_job")))

Results

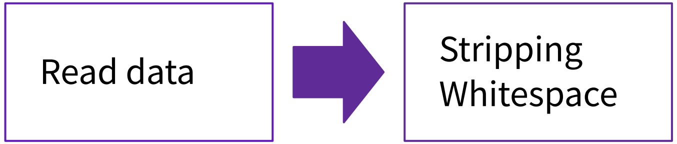
												raining_hours
10021	city_16	0.91	Female	yes	NO	Graduate	STEM	5	10000+	Pvt Ltd	1	22
10049	city_160	0.92	Male	yes	NO	Graduate	STEM	21	100-500	Pvt Ltd	1	20
10050	city_103	0.92	Male	yes	NO	Masters	STEM	10	10000+	Pvt Ltd	5	20
10162	city_160	0.92	Female	yes	NO	Graduate	STEM	15	500-999	Pvt Ltd	3	29
10167	city_103	0.92	Female	No	Parttime	Graduate	STEM	2	50-99	Funded Startup	1	35
10171	city_61	0.91	Male	yes	NO	Masters	STEM	21	10-49	Pvt Ltd	5	40
10198	city_105	0.79	Male	yes	Fulltime	Graduate	STEM	11	100-500	Pvt Ltd	5	3
10217	city_159	0.84	Male	yes	NO	Graduate	STEM	8	50-99	Pvt Ltd	1	33
10230	city_103	0.92	Male	yes	NO	Masters	STEM	21	500-999	Funded Startup	1	60
10246	city_160	0.92	Male	yes	NO	Masters	STEM	17	others	Others	1	9
1026	city_162	0.77	Others	yes	NO	Masters	STEM	11	1000-4999	Pvt Ltd	3	62
10260	city_21	0.62	Others	No	Parttime	Masters	STEM	2	others	Others	2	32
10279	city_71	0.88	Others	No	Fulltime	Graduate	STEM	6	others	Others	never	45
10287	city_160	0.92	Male	yes	NO	Graduate	STEM	13	500-999	Pvt Ltd	1	20
10304	city_103	0.92	Male	yes	NO	Masters	STEM	10	others	Public Sector	1	31
10308	city_160	0.92	Female	No	NO	Graduate	STEM	1	100-500	NGO	1	8
10311	city_67	0.86	Male	yes	Fulltime	Graduate	STEM	11	10-49	Others	1	106
10324	city_16	0.91	Male	yes	Parttime	Graduate	STEM	21	10000+	Pvt Ltd	2	12
10348	city_103	0.92	Male	No	NO	Graduate	STEM	18	50-99	Pvt Ltd	5	28
10394	city_103	0.92	Male	yes	NO	Graduate	Other	15	50-99	Funded Startup	1	10

only showing top 20 rows

Mapper



Reducer



Read data

```
for line in sys.stdin:
```

Stripping Whitespace

```
line.strip()
```

Mapper

Discretizing

```
Def
discretize_experience(
experience):
   try:
      if experience ==
'>20':
         return 21
      elif experience ==
'<1':
         return 0
      else:
         return
int(experience)
   except ValueError:
      return experience
```

Read and extract data

```
header = True

for line in sys.stdin:
   if header:


      print(line.strip())
      header = False
      continue
```

Tokenize

```
  columns =
line.strip().split(',')



discrete_experience =
discretize_experience(
columns[0])
```

output

```
     columns[0] =
str(discrete_experienc
e)

print(",".join(columns))
```

HDFS

```
start-dfs.sh
start-yarn.sh

hdfs dfs -put
/path/to/local/cleandata.csv
/input_path/cleandata.csv

hdfs dfs -rm -r /new_output_path2
hadoop jar
/usr/local/hadoop/share/hadoop/to
ols/lib/hadoop-streaming-*.jar \
  -input /input_path/cleandata.csv \
  -output /new_output_path2 \
  -mapper
/home/ubuntu/mapper2.py \
  -reducer
/home/ubuntu/reducer2.py

hdfs dfs -cat
/new_output_path2/part-00000
```

Result of mapp and reducer on HDFS

```
9106,city_104,0.92,Male,yes,NO,Graduate,STEM,21,50-99,Pvt Ltd,5,95
9134,city_21,0.62,Others,yes,NO,Graduate,STEM,7,50-99,Pvt Ltd,2,85
9149,city_23,0.9,Others,yes,NO,Masters,STEM,18,50-99,Pvt Ltd,3,60
915,city_103,0.92,Others,yes,NO,Masters,Humanities,16,1000-4999,Pvt Ltd,1,11
9163,city_21,0.62,Male,yes,NO,Graduate,Other,4,100-500,Pvt Ltd,1,24
9184,city_136,0.9,Male,yes,Parttime,Masters,STEM,3,10000+,Pvt Ltd,1,34
9195,city_21,0.62,Male,yes,NO,Graduate,STEM,10,100-500,Pvt Ltd,2,129
9205,city_143,0.74,Others,yes,NO,Graduate,STEM,13,others,Others,5,37
9207,city_114,0.93,Male,yes,NO,High School,Others,8,9,Pvt Ltd,never,39
9208,city_21,0.62,Others,No,Parttime,Graduate,STEM,2,10-49,Pvt Ltd,1,58
9209,city_121,0.78,Other,yes,NO,Graduate,STEM,0,50-99,Funded Startup,1,62
923,city_103,0.92,Male,yes,NO,Graduate,Humanities,21,50-99,Funded Startup,1,102
9234,city_61,0.91,Others,yes,NO,Graduate,STEM,21,9,Pvt Ltd,3,220
9237,city_65,0.8,Others,yes,NO,Phd,STEM,17,50-99,Pvt Ltd,5,41
9268,city_160,0.92,Male,yes,NO,Others,Others,21,9,Pvt Ltd,2,26
9270,city_162,0.77,Male,yes,Parttime,Graduate,STEM,6,10-49,NGO,4,31
9272,city_90,0.7,Male,yes,NO,Graduate,STEM,20,10-49,Pvt Ltd,2,51
9275,city_158,0.77,Male,yes,Parttime,Graduate,STEM,13,others,Others,5,324
9291,city_16,0.91,Male,yes,NO,Masters,STEM,15,10-49,Pvt Ltd,2,15
9302,city_149,0.69,Others,yes,NO,Graduate,STEM,9,others,Others,2,152
9335,city_103,0.92,Male,No,NO,Phd,STEM,18,100-500,NGO,1,25
9345,city_16,0.91,Other,yes,NO,High School,Others,8,50-99,Pvt Ltd,1,18
9462,city_21,0.62,Others,No,Parttime,Others,Others,1,others,Others,never,204
9487,city_21,0.62,Male,No,Fulltime,Graduate,STEM,5,others,Others,never,141
9501,city_71,0.88,Male,yes,NO,Graduate,STEM,5,100-500,Others,1,20
9514,city_160,0.92,Male,yes,Fulltime,High School,Others,12,100-500,Pvt Ltd,1,9
952,city_103,0.92,Male,yes,NO,Graduate,STEM,7,10000+,Pvt Ltd,1,96
9544,city_21,0.62,Male,No,NO,Graduate,STEM,4,others,Others,never,190
9548,city_114,0.93,Male,yes,NO,Masters,STEM,21,50-99,Pvt Ltd,5,65
9556,city_40,0.78,Others,yes,NO,Graduate,STEM,20,others,Others,1,23
9561,city_28,0.94,Others,yes,NO,Masters,STEM,19,others,Others,2,72
9562,city_73,0.75,Male,yes,NO,Graduate,STEM,21,others,Others,1,167
9564,city_103,0.92,Others,yes,NO,Graduate,STEM,9,1000-4999,Pvt Ltd,3,72
9586,city_103,0.92,Female,yes,NO,Masters,STEM,4,5000-9999,Public Sector,1,20
9618,city_103,0.92,Female,yes,NO,Graduate,Humanities,21,others,Others,5,22
9630,city_103,0.92,Male,yes,NO,Graduate,STEM,21,1000-4999,Pvt Ltd,5,43
9649,city_103,0.92,Male,yes,NO,Graduate,Business Degree,16,50-99,Pvt Ltd,1,33
9664,city_41,0.83,Male,yes,Parttime,Graduate,STEM,8,9,Pvt Ltd,2,86
9700,city_71,0.88,Female,No,NO,Graduate,Humanities,4,10000+,Pvt Ltd,2,34
9706,city_136,0.9,Male,No,NO,Masters,STEM,10,100-500,NGO,5,26
9707,city_103,0.92,Male,No,Fulltime,Graduate,STEM,8,others,Others,2,96
9726,city_160,0.92,Male,yes,NO,Graduate,STEM,16,others,Others,5,42
9740,city_103,0.92,Male,No,NO,Graduate,STEM,21,10000+,Pvt Ltd,5,21
9752,city_21,0.62,Others,yes,Fulltime,Graduate,STEM,6,others,Others,4,32
9753,city_37,0.79,Female,No,Fulltime,Graduate,STEM,4,others,Others,never,86
976,city_67,0.86,Male,yes,NO,Graduate,STEM,7,100-500,Pvt Ltd,2,57
9766,city_83,0.92,Male,yes,NO,Graduate,STEM,5,1000-4999,Pvt Ltd,2,14
9772,city_114,0.93,Male,No,NO,High School,Others,5,others,Others,never,32
9789,city_160,0.92,Male,yes,NO,Graduate,STEM,21,10000+,Pvt Ltd,5,4
9800,city_103,0.92,Others,No,NO,Masters,STEM,21,10000+,Pvt Ltd,3,59
9806,city_65,0.8,Male,yes,NO,Masters,STEM,15,10000+,Pvt Ltd,5,27
9827,city_138,0.84,Male,No,Fulltime,High School,Others,2,others,Others,never,112
9837,city_61,0.91,Male,yes,NO,Graduate,STEM,21,others,Others,1,42
9840,city_114,0.93,Male,No,Fulltime,High School,Others,8,others,Public Sector,1,81
9852,city_103,0.92,Male,yes,NO,Graduate,STEM,21,others,Others,3,23
```

Converting to Json

Listing Files in Directory

for filename in os.listdir(input_directory):

if filename.startswith('part-'):

file_path = os.path.join(input_directory, filename) data = []

Read csv file

with open(file_path, mode='r', encoding='utf-8') as file: csv_reader = csv.DictReader(file,fieldnames=fieldnames)

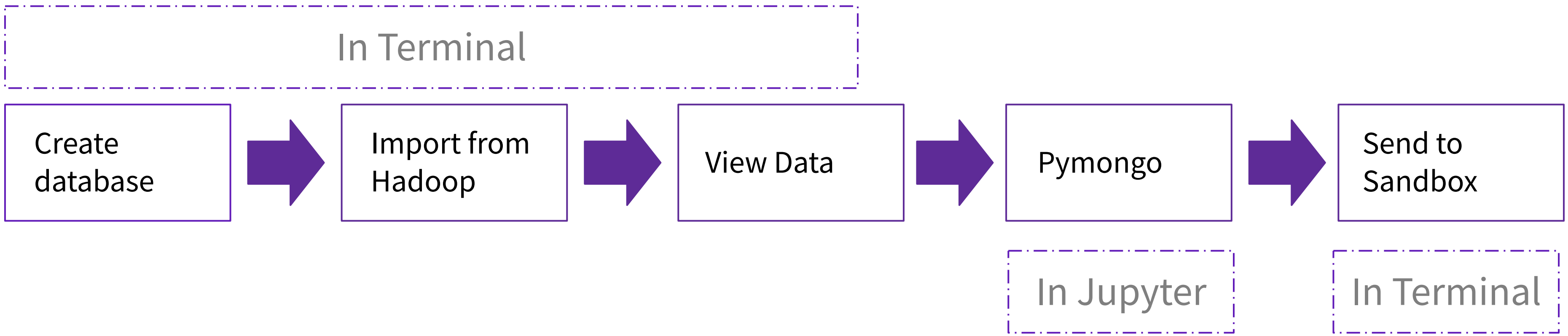
for row in csv_reader: data.append(row)

Writing Data to JSON Files

json_filename = os.path.join(output_directory, f"{filename}.json") with open(json_filename, mode='w', encoding='utf-8') as json_file:

json.dump(data, json_file, indent=4, ensure_ascii=False) print(f"Converted {file_path} to {json_filename}")

MongoDB



In Terminal
Create database
Import from Hadoop
View Data
Pymongo
Send to Sandbox
In Jupyter
In Terminal

Mongodb

Create database

```
use your_database
```

Importing from Hadoop

```
for file in *.json; do
  mongoimport --uri mongodb://localhost:27017/your_database \
        --collection your_collection \
        --file "$file" \
        --jsonArray
done
```

View Data

```
mongo --host localhost --port 27017

use your_database

db.your_collection.find().pretty()
```

All in Terminal

Results


```
,
{
        "_id" : ObjectId("66782a3171c97f29dcf53a7a"),
        "enrollee_id" : 10348,
        "city" : "city_103",
        "city_development_index" : 0.92,
        "gender" : "Male",
        "relevent_experience" : "No",
        "enrolled_university" : 1,
        "education_level" : "Graduate",
        "major_discipline" : "STEM",
        "experience" : 18,
        "company_size" : "50-99",
        "company_type" : "Pvt Ltd",
        "last_new_job" : "5",
        "training_hours" : "28",
        "target" : 0
}
{
        "_id" : ObjectId("66782a3171c97f29dcf53a7b"),
        "enrollee_id" : 10394,
        "city" : "city_103",
        "city_development_index" : 0.92,
        "gender" : "Male",
        "relevent_experience" : "yes",
        "enrolled_university" : 1,
        "education_level" : "Graduate",
        "major_discipline" : "Other",
        "experience" : 15,
        "company_size" : "50-99",
        "company_type" : "Funded Startup",
        "last_new_job" : "1",
        "training_hours" : "10",
        "target" : 1
}
Type "it" for more
>
```

Work with data in PyMongo

Connect to mongodb

```
from pymongo import MongoClient
import pandas as pd

client = MongoClient('mongodb://localhost:27017/')
db = client.your_database
collection = db.your_collection
```

Fetch data from MongoDB

```
data = list(collection.find())
```

Convert to Pandas DataFrame

```
df = pd.DataFrame(data)
```

Send to sandbox

```
scp /path/to/sandbox/data.parquet user@your-sandbox-ip:/desired/path
```

Results

	_id	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_ty
0	66782a3171c97f29dcf53a68	10021.0	city_16	0.91	Female	yes	1	Graduate	STEM	5.0	10000+	Pvt
1	66782a3171c97f29dcf53a69	10049.0	city_160	0.92	Male	yes	1	Graduate	STEM	21.0	100-500	Pvt
2	66782a3171c97f29dcf53a6a	10050.0	city_103	0.92	Male	yes	1	Masters	STEM	10.0	10000+	Pvt
3	66782a3171c97f29dcf53a6b	10162.0	city_160	0.92	Female	yes	1	Graduate	STEM	15.0	500-999	Pvt
4	66782a3171c97f29dcf53a6c	10167.0	city_103	0.92	Female	No	2	Graduate	STEM	2.0	50-99	Funded Star
5	66782a3171c97f29dcf53a6d	10171.0	city_61	0.91	Male	yes	1	Masters	STEM	21.0	10-49	Pvt
6	66782a3171c97f29dcf53a6e	10198.0	city_105	0.79	Male	yes	0	Graduate	STEM	11.0	100-500	Pvt
7	66782a3171c97f29dcf53a6f	10217.0	city_159	0.84	Male	yes	1	Graduate	STEM	8.0	50-99	Pvt
8	66782a3171c97f29dcf53a70	10230.0	city_103	0.92	Male	yes	1	Masters	STEM	21.0	500-999	Funded Star
9	66782a3171c97f29dcf53a71	10246.0	city_160	0.92	Male	yes	1	Masters	STEM	17.0	others	Oth

Analysis & Testing

Steps

1. Loading Data
2. Data Preprocessing
3. Selecting Features and Target
4. Visualization the distribution of categorical feature (Bar and Histogram plot)
5. Train-Test Split
6. Classification ML Model for prediction
7. Evaluate ML models
8. Analysis

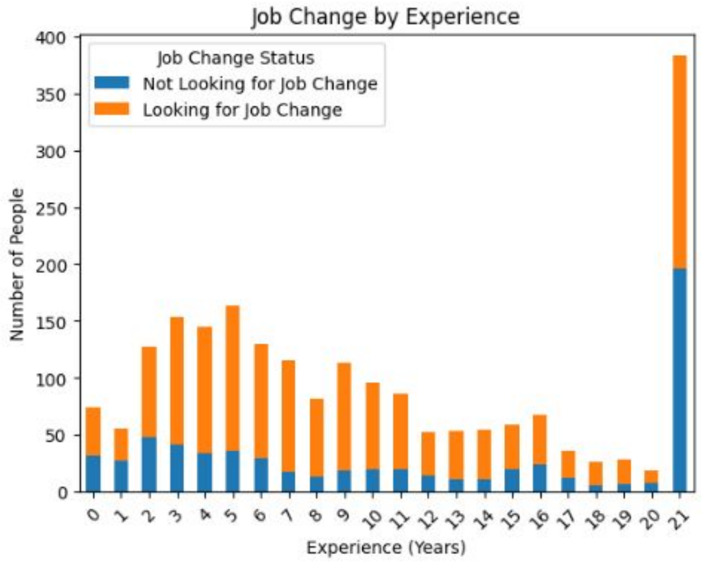
H1: **Candidates with <u>longer experience</u> are less likely to seek new job opportunities.**

Data Preprocessing

- The `experience` column is converted to numerical values using the `convert_experience` function.

Features and Target (showing in bar plot)

- The feature (`X`) includes **only** the `experience_category`,
- The target (`y`) includes the `changing_job`



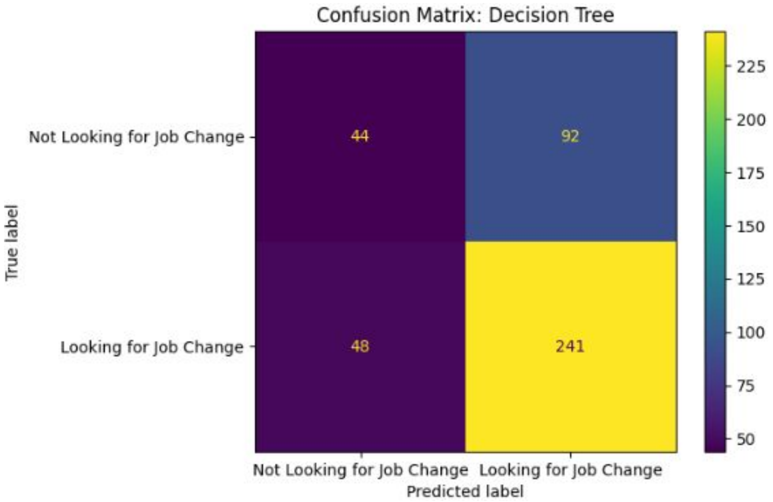
H1: Candidates with longer experience are less likely to seek new job opportunities.

- **Decision tree for Model Prediction**

  - **Precision**: when the model predicts a candidate is looking for a job change, it is correct **72%** of the time.
  - **Recall**: The model correctly identifies **83%** of the candidates who are looking for a job change.
  - The **F1-score** is the harmonic mean of precision and recall.

Decision Tree Model Performance:

		precision	recall	f1-score	support
NL	0.0	0.48	0.32	0.39	136
L	1.0	0.72	0.83	0.77	289
accuracy				0.67	425
macro avg		0.60	0.58	0.58	425
weighted avg		0.65	0.67	0.65	425



- **Confusion Matrix (Evaluation)**

0	(TN): 44	(FP): 92
1	(FN): 48	(TP): 241

  - Positive/ Negative: Nl/L for job
  - True/ False: Identified correctly/ incorrectly by model

H1: Candidates with <u>longer experience</u> are less likely to seek new job opportunities.

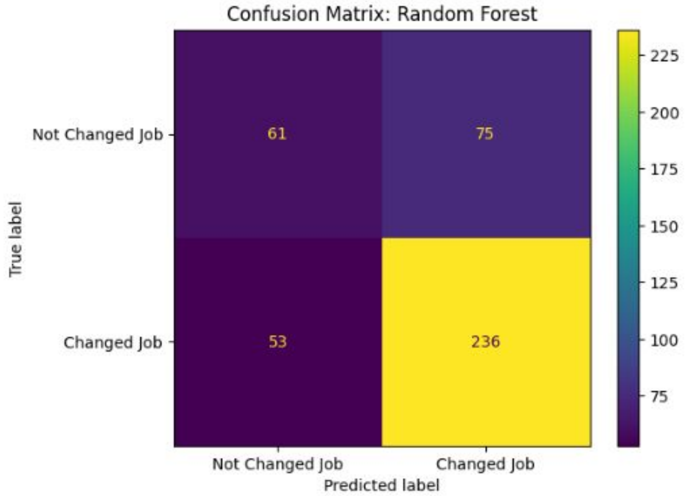
- **Random forest for Model Prediction**

Changing the algorithm and considering more variables and re-examining their effect on the target.

- Education_level, relevent_experience
- Overall accuracy: 0.70

Random Forest Model Performance:

		precision	recall	f1-score	support
NL	0.0	0.54	0.45	0.49	136
L	1.0	0.76	0.82	0.79	289
	accuracy			0.70	425
	macro avg	0.65	0.63	0.64	425
	weighted avg	0.69	0.70	0.69	425

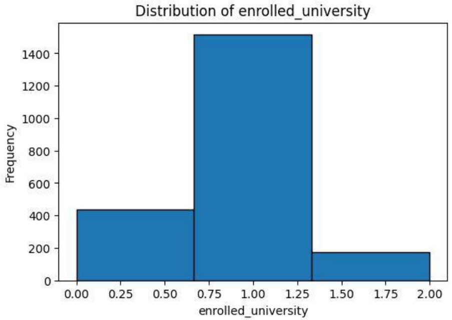


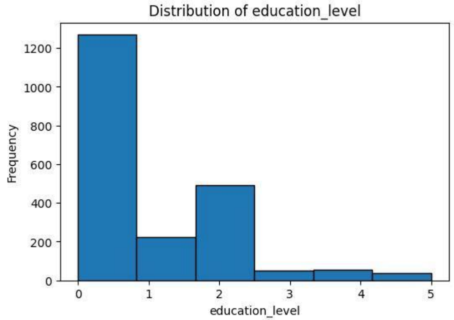
Given that the "experience" feature is the most important feature in the model:

- Individuals with less experience (class 0) are more likely to change jobs.
- Individuals with more experience (class 1) are less likely to change jobs.

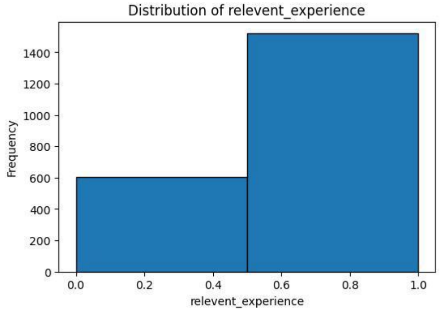
H2 Candidates enrolled in university courses are more likely to look for new job opportunities.

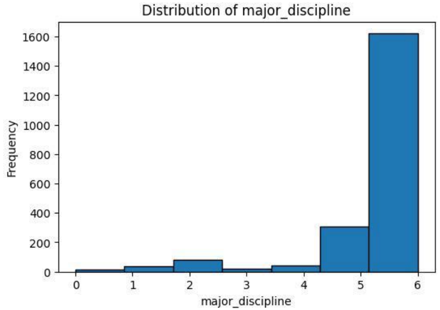
Firstly, we made histogram plots in our sample code visualize the distribution of categorical features (`enrolled_university`, `education_level`, `relevent_experience`, `major_discipline`).





Here's what each part does:





H2: Candidates enrolled in university courses are more likely to look for new job opportunities.

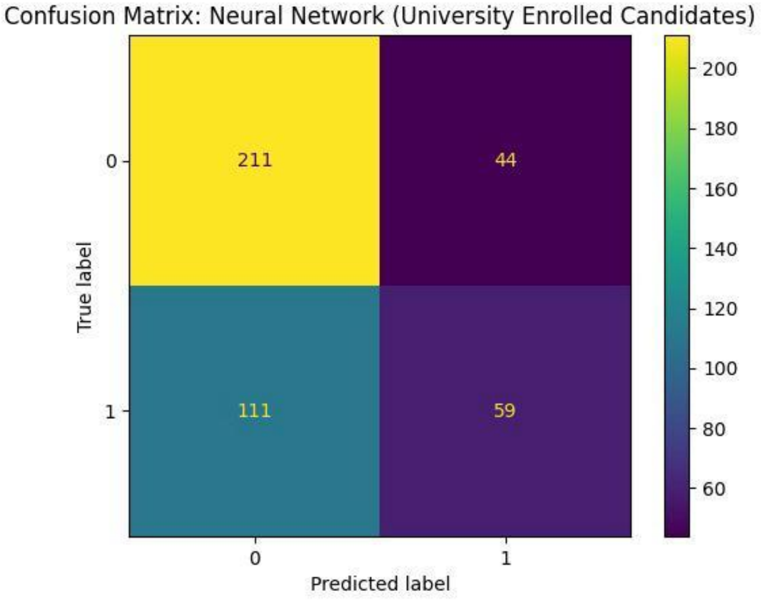
Neural Network Model based on the figure results, the neural network (NN) model achieves an accuracy of **0.61,** with a **precision** of **0.57,** recall of **0.35,** and **F1-score of 0.43** for candidates likely to seek new job opportunities **(class 1).**

These metrics indicate that NN effectively identifies a substantial number of candidates who are actually looking for new job opportunities while maintaining a balanced performance in terms of precision and recall.

Therefore, NN is recommended for predicting candidates' likelihood to seek new job opportunities based on their experience.

Neural Network Model Performance (University Enrolled Candidates):

	precision	recall	f1-score	support
0	0.66	0.83	0.73	255
1	0.57	0.35	0.43	170
accuracy			0.64	425
macro avg	0.61	0.59	0.58	425
weighted avg	0.62	0.64	0.61	425

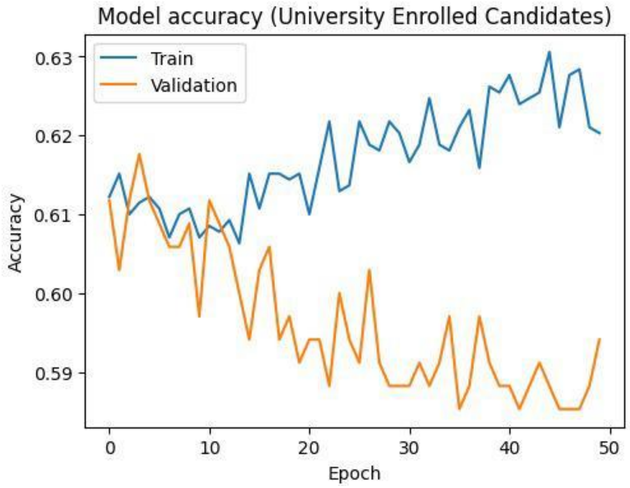


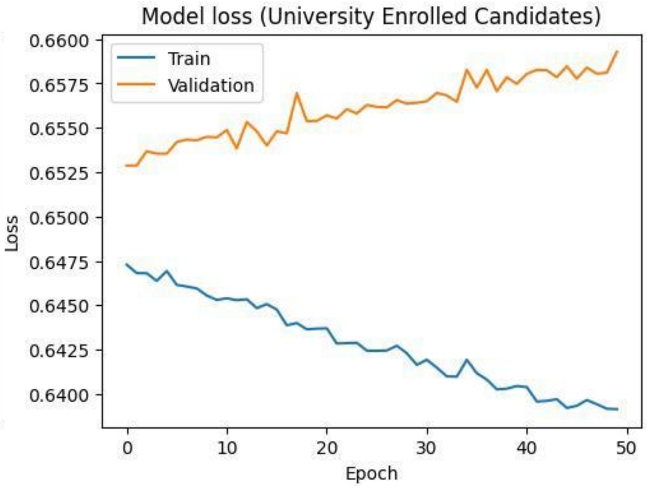
H2 Candidates enrolled in university courses are more likely to look for new job opportunities.

```
# Evaluate the model on university enrolled candidates
y_pred_prob_univ = model.predict(X_test_univ)
y_pred_univ = (y_pred_prob_univ > 0.5).astype(int)
```

Training & validation accuracy values for university enrolled candidates

Training & validation loss values for university enrolled candidates





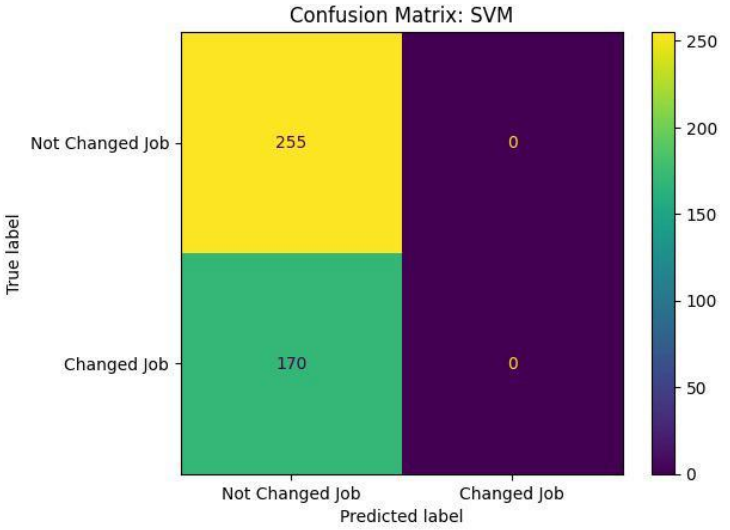
H2 Candidates enrolled in university courses are more likely to look for new job opportunities.

SVM Support Vector Machine show a precision, recall, and F1-score of 0.00 for candidates seeking new job opportunities (class 1), this means the model fails to correctly identify any of these candidates.

As a result, SVM does not effectively capture the relationship between candidates' experience and their likelihood to seek new job opportunities.

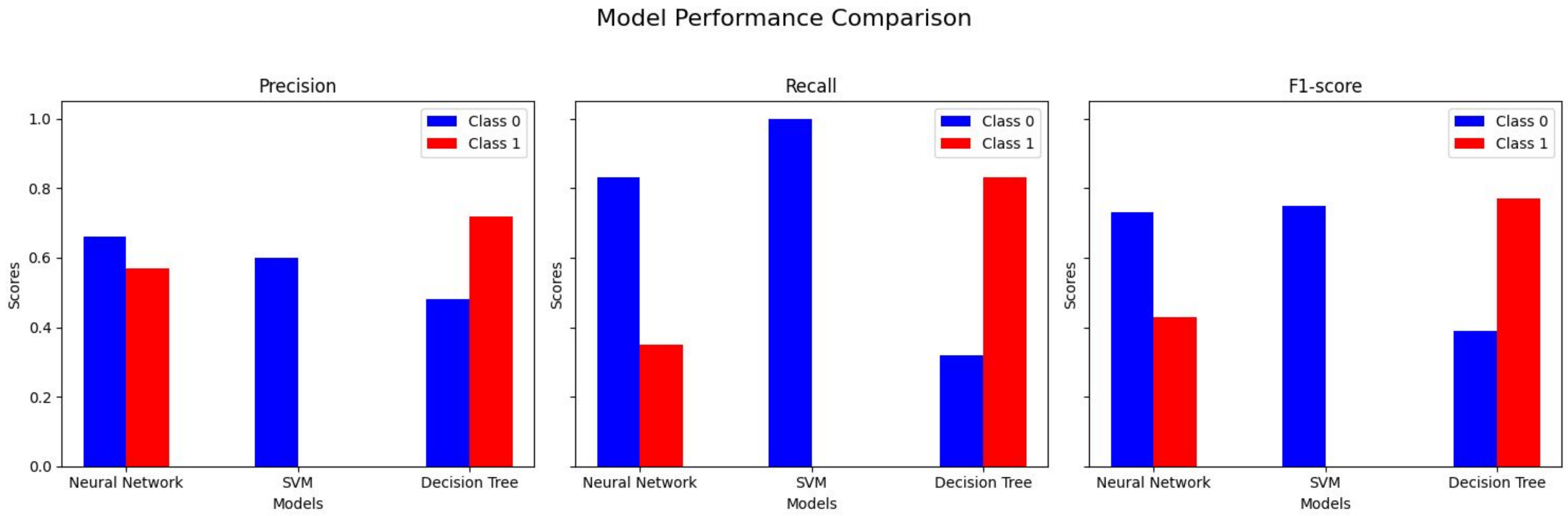
Therefore, SVM's performance in this regard is inadequate for drawing conclusions about this hypothesis

```
Support Vector Machine (SVM) Model Performance:
              precision    recall  f1-score   support

           0       0.60      1.00      0.75       255
           1       0.00      0.00      0.00       170

    accuracy                           0.60       425
   macro avg       0.30      0.50      0.37       425
weighted avg       0.36      0.60      0.45       425
```



Conclusion

This comparison helps evaluate the models' capabilities in validating hypotheses related to candidate job change behavior, such as the influence of university enrollment and work experience length.



Challenge & Limitation

- HDFs restrict connectivity in our devices, causing problems with VMs connecting effectively to intensive processes.
- The sandbox environment struggles to establish a reliable connection between MongoDB data storage and machine learning code files, leading to operational issues.
- In hypothesis 1, there are limitations in working with only one values, so we improved the problem in the second hypothesis.
- Attempts to use machine learning models for prediction were unsuccessful, even after switching models, as results remained inaccurate. Further efforts to categorize data into two columns did not yield useful outcomes.

References

DataSet Source :

https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists

DS&BD Course Notes:

https://elearning.unipv.it/course/view.php?id=6951

VM Environment Configuration :

https://drive.google.com/file/d/1KxCrpDNotz2kuo-G8O5VvRp_YFNRNDdX/view

Any Questions?

Thank you

شكرًا - Moteshakeram

Contact

Sepideh Hayati

Master Student Computer Engineering - Computer Science Track

sepideh.hayati01@universitadipavia.it

SAHAR TAHERI

Master Student Computer Engineering - Computer Science Track

sahar.taherimoghadar01@universitadipavia.it

Kamila Aburwais

Master Student Computer Engineering - Data Science Track

kamilamustafaa.aburwais01@universitadipavia.it