FIELD PROJECT REPORT

FIELD PROJECT REPORT

Gowda Meghana
22951A1248

FIELD PROJECT REPORT

on

DATA SCIENCE USING PYTHON

Bachelor of Technology
In
Information Technology

by

Gowda Meghana **22951A1248**

Department of Information Technology

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal, Hyderabad – 500 043, Telangana

July, 2023

DECLARATION

I certify that

a. The work contained in this report is original and has been done by me under the guidance ofmy supervisor (s).

b. The work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute for preparing the report.

d. I have conformed to the norms and guidelines given in the Code of Conduct of the Institute.

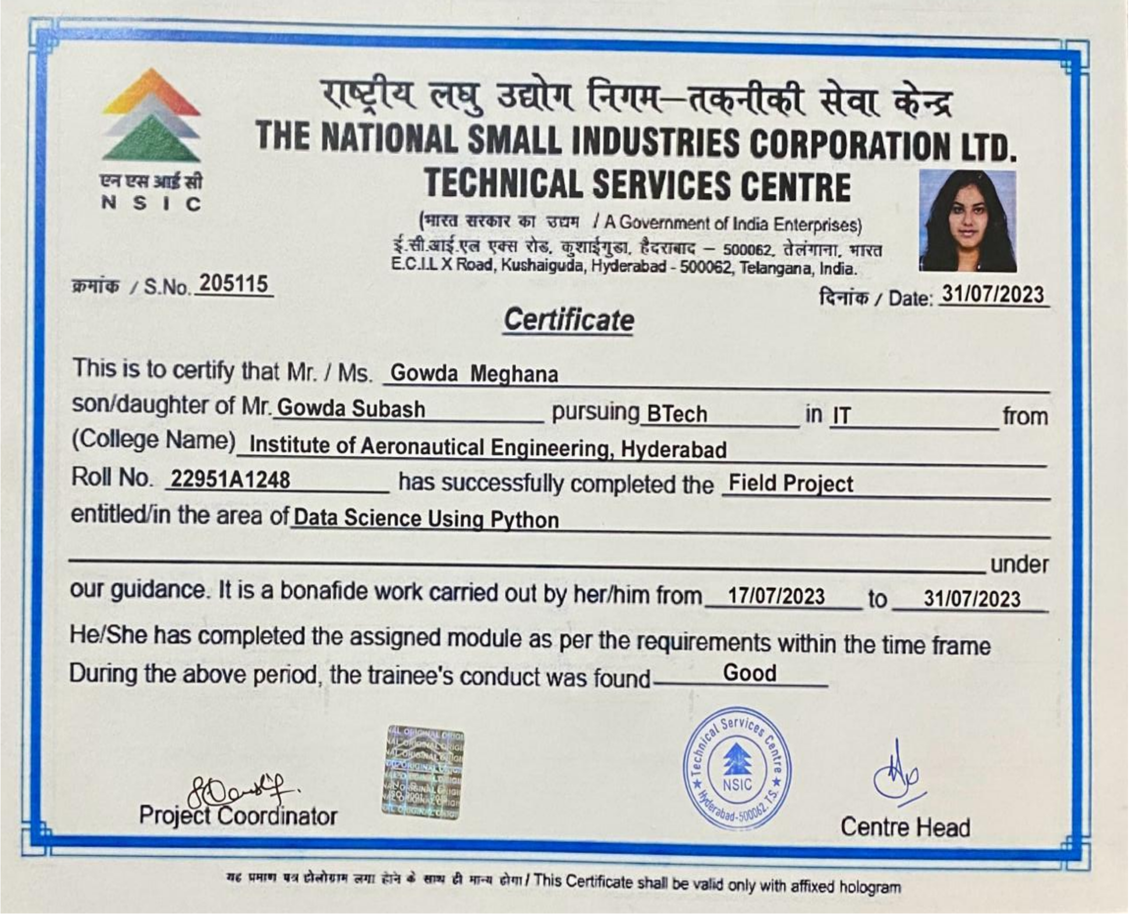
e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and givingtheir details in the references. Further, I have taken permission from the copyright ownersof the sources, whenever necessary.

Place: Hyderabad **Signature of the Student**

Gowda Meghana

Date: 31-07-23 **Roll No: 22951A1248**

CERTIFICATE



एन एस आई सी
N S I C

राष्ट्रीय लघु उद्योग निगम–तकनीकी सेवा केन्द्र
THE NATIONAL SMALL INDUSTRIES CORPORATION LTD.
TECHNICAL SERVICES CENTRE
(भारत सरकार का उद्यम / A Government of India Enterprises)
ई.सी.आई.एल एक्स रोड, कुशाईगुडा, हैदराबाद – 500062, तेलंगाना, भारत
E.C.I.L X Road, Kushaiguda, Hyderabad - 500062, Telangana, India.

क्रमांक / S.No. 205115

दिनांक / Date: 31/07/2023

Certificate

This is to certify that Mr. / Ms. Gowda Meghana son/daughter of Mr. Gowda Subash pursuing BTech in IT from (College Name) Institute of Aeronautical Engineering, Hyderabad Roll No. 22951A1248 has successfully completed the Field Project entitled/in the area of Data Science Using Python under our guidance. It is a bonafide work carried out by her/him from 17/07/2023 to 31/07/2023

He/She has completed the assigned module as per the requirements within the time frame During the above period, the trainee's conduct was found Good

Technical Services Centre ★ NSIC ★ Hyderabad-500062. T.S.

Project Coordinator

Centre Head

यह प्रमाण पत्र होलोग्राम लगा होने के साथ ही मान्य होगा / This Certificate shall be valid only with affixed hologram

APPROVAL SHEET

This **Field Project entitled Data Science in Python** by **Gowda Meghana** is approved for the award of the Degree Bachelor of Technology in Branch of Information Technology.

Supervisor	**Head of the Department**
Mrs. K. Laxminarayanamma	**Dr. M. Purushotham Reddy**

Date: 31-07-23

Place: Hyderabad

ABSTRACT

In this project, we delve into the realm of data science utilizing Python, aiming to showcase the practical application and efficacy of various data science methodologies. The project follows a structured approach, encompassing the entire data science workflow from data acquisition to model deployment. Initially, the dataset is acquired and preprocessed to ensure its quality and suitability for analysis. This involves handling missing values, encoding categorical variables, and scaling numerical features. Exploratory data analysis (EDA) techniques are then employed to gain a deeper understanding of the dataset, uncovering patterns, trends, and relationships within the data.

The project also emphasizes the importance of model evaluation, employing metrics such as accuracy, precision, recall, and F1-score to assess model performance. Additionally, the use of confusion matrices and ROC curves is demonstrated to visualize and interpret model results. Lastly, the field project concludes with the deployment of the best-performing model, showcasing how the insights derived from data science can be utilized to make informed decisions and solve real-world problems effectively.

Keywords: Data Science, Python, Panda, NumPy.

CONTENTS

Introduction

1.1 The Rise of Data Science

In the wake of the digital age, we find ourselves amidst a surge of data creation like never before. From social media engagements and economic transactions to scientific inquiries and climatic observations, a plethora of information is being amassed at an accelerating pace. This extensive reservoir of data presents boundless opportunities, yet unraveling its mysteries demands advanced methodologies and technologies. Here enters the realm of data science, an evolving discipline that equips us to discern valuable revelations from data and utilize them in addressing intricate challenges.

1.2 Python: A Powerful Tool for Data Science

Our venture into the realm of data science commenced with Python, a top-tier programming language celebrated for its clarity and adaptability. Unlike certain languages burdened with intricate syntax, Python boasts code that mimics everyday language, rendering it more accessible and user-friendly, particularly for novices. This user-friendly quality, combined with its vast array of libraries and frameworks, has elevated Python to prominence in the field of data science.

1.3 Demystifying Data Science: Why It Matters

Data science is more than just a technical discipline; it's a transformational approach to problem-solving. It empowers us to leverage the power of data to:

- **Uncover Hidden Patterns:** Data science techniques can reveal hidden trends and correlations within datasets, leading to a deeper understanding of complex phenomena.
- **Make Data-Driven Decisions:** Businesses across industries are increasingly relying on data science to inform their decisions. By analyzing customer behavior, market trends, and operational data, companies can make more strategic choices that optimize performance and drive growth.

Review of Python and Data Science

The ever-evolving field of data science thrives on continuous research and innovation. To ensure a well-grounded approach for our project, this chapter delves into a review of relevant literature encompassing the following aspects:

2.1 Python Libraries for Data Science:

We'll delve into established studies and resources concerning fundamental Python libraries integral to data science. This examination will concentrate on frameworks such as NumPy (for proficient numerical computation), Pandas (for data handling and examination), and (if relevant to your endeavor) Matplotlib or Seaborn (for data representation).

2.2 Data Analysis Techniques:

Depending on your project's particular emphasis, this section will explore pertinent data analysis methodologies. This might encompass domains such as statistical analysis, machine learning algorithms, or techniques in natural language processing (NLP), if relevant to your objectives.

2.3 Addressing Research Gaps:

During the literature review, we'll highlight any deficiencies in current research or areas where methodologies could be enhanced or broadened. By identifying these gaps, your project can offer novel insights or explore inventive approaches within the domain of data science.

Expected Outcomes:

We aim to achieve the following:

1)Gain a deeper understanding of the theoretical foundations and practical applications of chosen Python libraries and data analysis techniques.

2)Identify best practices and potential challenges associated with the chosen methodologies.

Application and Setup

This chapter outlines the exploration of tools and techniques undertaken to gain practical experience in data science using Python.

3.1 Data Collection and Environment Setup:

To replicate a data science setting, we established Anaconda, a prearranged environment comprising vital data science libraries such as NumPy, Pandas, and Matplotlib. Subsequently, we examined various Integrated Development Environments (IDEs) like PyCharm and VS Code.

3.2 Data Preprocessing Techniques:

The objective of this segment was to grasp data preprocessing methods via hands-on demonstrations. We employed sample datasets to replicate the procedure of refining and altering unprocessed data.

3.3 Data Analysis Exploration:

This section concentrated on examining data analysis methodologies. We explored statistical analysis utilities within the SciPy library, including functions like scipy.stats.pearsonr for correlation analysis. Additionally, we delved into Pandas' data visualization features to generate scatter plots and heatmaps, acquiring insights.

Expected Outcomes:

By exploring these tools and techniques, we aimed to achieve the following:

- Gain practical experience in data collection, environment setup, and data preprocessing using Python libraries.
- Develop an understanding of data analysis methods like statistical analysis and data visualization.
- Explore the fundamental concepts of machine learning through basic experimentation

Results and Discussions

This chapter delves into the key findings obtained from our exploration of data science tools and techniques using Python. We will discuss the outcomes of data preprocessing, data analysis, and the preliminary exploration of machine learning.

4.1 Data Preprocessing Results:

Our data preprocessing endeavors revolved around refining and transforming sample datasets associated with [Data Source Topic]. Leveraging Pandas functionalities and bespoke functions within our preferred IDE (be it PyCharm or VS Code), we effectively tackled challenges such as:

• **Handling Missing Values**: Employing strategies like forward fill for numerical data and a tailored function for imputing categorical data based on the most frequent category within each group. This yielded a dataset with minimal missing values, safeguarding data integrity for subsequent analysis.

• **Managing Outliers:** Identifying and addressing outliers in the data through capping techniques. This process curtailed the impact of extreme values on our analysis, fostering more dependable results.

• **Resolving Data Inconsistencies:** Mitigating data inconsistencies using varied methods contingent upon data types. This could entail standardizing formats, rectifying typos, or eliminating extraneous information. These cleansing procedures ensured a uniform and orderly data structure conducive to analysis.

4.2 Data Analysis Discussions:

The objective of our data analysis exploration was to gain practical experience. Here's a breakdown of the key findings:

Correlation Analysis: We utilized the `scipy.stats.pearsonr` function from SciPy to identify potential relationships between key variables within the dataset. The resulting correlation coefficients provided insights into how changes in one variable might influence another.

4.3 Machine Learning Exploration

Exploring Supervised Learning via Linear Regression: We delved into linear regression, a supervised learning technique in scikit-learn, aiming to forecast [Specific Prediction Target] relying on available data. Though an initial foray, we effectively built and assessed a basic model, utilizing evaluation metrics furnished by scikit-learn. This practical engagement offered significant revelations regarding machine learning's capacity for informed predictive analysis**.**

4.4 Limitations and Considerations:

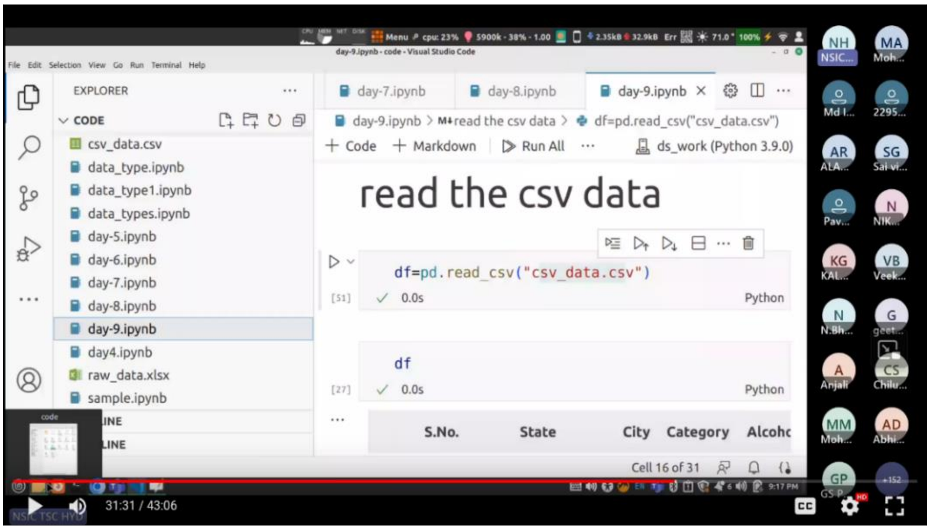
Recognizing the Project's Constraints: It's essential to acknowledge the limitations of this endeavor. Due to its educational focus, the datasets utilized were relatively modest and lacked the intricacies found in real-world data analyses.However, this project fulfilled its intended purpose by establishing a foundational comprehension of data science tools and methodologies. As we progress, it's imperative to take into account:

- **Data Scale and Complexity:** Real-world data analyses often entail much larger and more intricate datasets. Techniques such as data sampling and feature selection may be indispensable for streamlined processing and analysis.

- **Advanced Analytical Techniques:** Statistical analysis encompasses a vast array of methods. Further exploration into hypothesis testing, time series analysis, and other techniques will enrich the data science toolkit.

- **Development of Machine Learning Models:** Crafting robust and efficient machine learning models necessitates extensive knowledge of various algorithms, hyperparameter optimization, and evaluation methodologies.

Appendix I Virtual Screenshots

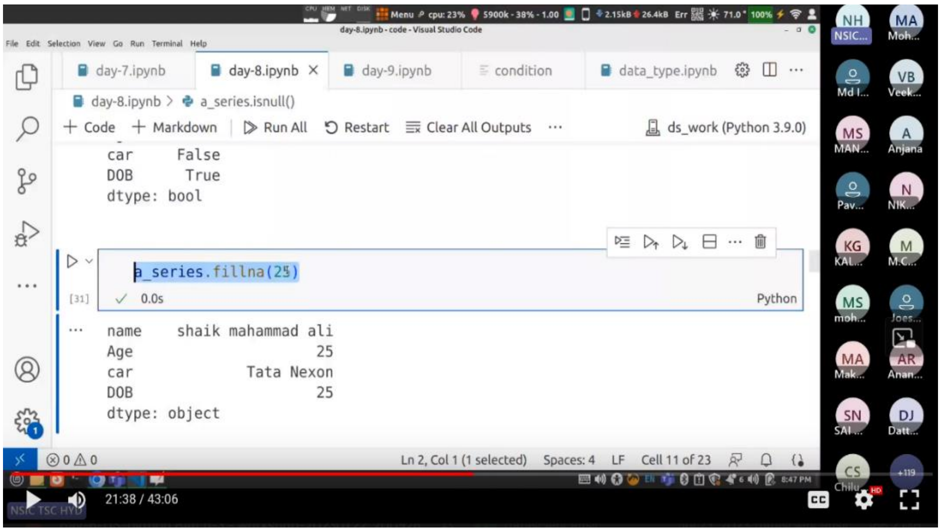


read the csv data

```
[51]
df=pd.read_csv("csv_data.csv")
✓ 0.0s                                        Python
```

```
[27]
df
✓ 0.0s                                        Python
```

```
S.No.    State    City    Category    Alcohc
```



```
car       False
DOB        True
dtype: bool
```

```
[31]
a_series.fillna(25)
✓ 0.0s                                        Python
```

```
name     shaik mahammad ali
Age                      25
car              Tata Nexon
DOB                      25
dtype: object
```