

Id: 0358445

Name: Anuj Bhandari

Module-Code: ITS65704

Module-Name: Data Science Principles

Intake: August

Question 1

Ans:- the lifecycle of a datascience projects consists of following things:

1. Problem Definition: The problem definition Step is very important as it lays the groundwork for the whole project. A well-defined problem ensures the project align with the strategic goals of the business, which increases the likelihood of the project to deliver meaningful value. also, it helps to identify some potential risks early on.
2. Collection of data Identifying the data sources is a important step, we have to determine where to obtain the necessary data, which includes various external or internal APIs,Databases. we have to combine those data from multiple sources to create a dataset . the process like data cleaning, transforming and merging data to ensure the consistency and accuracy this process is also known as data integration
3. Data preprocessing Here, we handle missing values , remove duplicate values etc. also some data transforming techniques like normalizing are done. creating new features to better represent the underlying patterns.
4. Modeling here, we choose the appropriate machine learning algorithm that is suitable for the data we have. and train the model using the prepared data or preprocessed data. model should also be evaluated using relevant metrics and validation techniques.
5. model deployment integrating the model into the real world (production environment) for the existing systems.Also continuous monitor the model's performance and updates as needed to ensure the effectiveness overtime. 5 Decision Making Using the various insights derived for the data decisions

should be made. this is the lifecycle of the datascience project contribution of a structured framework to data-driven Decisionmaking

Clarity and Focus: A well-defined structure guarantees that the problem and its objectives are apparent right away. This emphasis aids in coordinating scientific endeavors with commercial objectives. **Efficiency and Coherence:** Using a standardize process reduces the amount of time needed to figure out the next step, which improves efficiency. Additionally, it guarantees methodology consistency, which is essential for reproducibility and dependability of outcomes. **Data Quality:** The organization's collection and preparation procedures guarantee that the data used for analysis is clear, pertinent, and of excellent quality. Good data quality is essential for developing accurate models and obtaining trustworthy insights.

Better Models: A structured approach to data analysis and modeling helps in selecting the right algorithms, tuning parameters, and evaluating models effectively. This leads to better-performing models that can provide more accurate predictions.

Transparency and Accountability: Documenting each step of the process and communicating results transparently builds trust with stakeholders. It also ensures accountability as all decisions and actions are traceable.

Scalability and Maintenance: A structured framework makes it easier to scale data science efforts and maintain models in production. It facilitates continuous monitoring and updating of models, ensuring they remain relevant over time.

Question 2

```
In [2]: # create a pandas data frame from a dictionary
import pandas as pd
# create a simple database of people
data = {'Name': ['Anuj', 'Sushil', 'Prabin'],
        'Location': ['Dhapasi', 'Balkumari', 'Kirtipur'],
        'Age': [20, 21, 22]}

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

	Name	Location	Age
0	Anuj	Dhapasi	20
1	Sushil	Balkumari	21
2	Prabin	Kirtipur	22

Question 3

```
In [15]: #importing the libraries for visualization

import matplotlib.pyplot as plt
```

```
In [5]: #Data
Hours_Studied = [10, 9, 3, 2, 12, 7, 5, 11, 8, 6, 4, 15, 10, 13, 2, 14, 6, 9, 7, 5]
Exam_Score = [95, 90, 75, 60, 98, 85, 70, 96, 88, 80, 65, 100, 95, 98, 55, 99, 80, 90,
```

```
In [17]: #Scatter plot

plt.figure(figsize=(10, 6))
plt.scatter(Hours_Studied, Exam_Score, color='Grey')

#Labels
plt.title('Relationship')
plt.xlabel('Hours Studied')
plt.ylabel('Exam Score')

plt.show()
```

