Task 4

**4. Data Science Lifecycle Example:**

**Example:- Diabetes Prediction in Healthcare**:

**1. Problem Definition:**

* **Objective:** Predict if a person will get diabetes based on their health data.

**2. Data Collection:**

* Collect patient data from electronic health records (EHRs), medical databases, and health surveys.
* Age, gender, blood sugar levels, BMI, family history

**3. Data Cleaning:**

Remove duplicates and irrelevant entries.

Handle missing values by imputing or discarding incomplete records.

Replacing null values using mean,mode,median.

**4. Exploratory Data Analysis (EDA):**

Understand the data distribution and identify patterns.

Visualize data using histograms, scatter plots, and box plots.

Identify correlations between features and diabetes occurrences.

**5. Feature Engineering:**

Create new features based on domain knowledge,Encode categorical variables (e.g., one-hot encoding for gender).

Normalize numerical features to ensure consistent scaling.

**6. Model Building:**

* **Algorithm Selection:** Choose suitable machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting, or Neural Networks.
* **Training:** Split the data into training and testing sets. Train the model using the training set.
* **Hyperparameter Tuning:** Optimize model parameters using techniques like Grid Search or Random Search.

**7. Model Evaluation:**

* **Metrics:** Evaluate the model's performance using metrics like accuracy, precision, recall, F1-score.
* **Validation:** Perform cross-validation to ensure the model's robustness and avoid overfitting.
* Refine the model based on evaluation results and re-train as needed.

**8. Deployment:**

* **Integration:** Integrate the trained model into the healthcare system's EHR or patient management system.