Biniyam kefelegn ID 010/2022

1. **Handling Missing Values and Outliers**
   - **When to Fill Missing Values**
     - ✓ **Before Data Analysis:** Filling missing values is crucial before conducting data analysis to maintain the integrity and accuracy of the results.
     - ✓ **Before Machine Learning:** Machine learning algorithms can't process missing values, so it's essential to address them beforehand.
   - **Strategies for Filling Missing Values**
     - ✓ **Mean/Median Imputation:** Replace missing values with the column's mean or median. This method is suitable for numerical data.
     - ✓ **Mode Imputation:** Fill missing values with the mode of the column, ideal for categorical data.
     - ✓ **Forward/Backward Fill:** Use the previous or next value to fill missing values in time-series data.
     - ✓ **Interpolation:** Estimate missing values based on surrounding data points using methods like linear, polynomial, or spline interpolation.
     - ✓ **K-Nearest Neighbors (KNN):** Predict missing values based on the values of the nearest neighbors.
     - ✓ **Advanced Imputation Techniques:** Utilize models like regression or machine learning algorithms to predict and fill missing values.
   - **Handling Outliers**
     - ✓ **Identification:** Use box plots, z-scores, or the IQR method to identify outliers.
     - ✓ **Removal:** Remove outliers if they result from data entry errors or are irrelevant to the analysis.
     - ✓ **Transformation:** Apply transformations like log, square root, or Box-Cox to reduce the impact of outliers.
     - ✓ **Capping:** Limit outliers to a certain threshold to minimize their effect.
     - ✓ **Robust Statistical Methods:** Use statistical methods that are less sensitive to outliers, such as robust regression.
2. **Data Analytic Life Cycles**
   1. **Differences, Pros, and Cons of Methodologies**
   - **Data Mining Processes (KDD, CRISP-DM, SEMA):**

- ✓ **Differences:**
    - ❖ **KDD:** Focuses on the overall process of discovering knowledge from data.
    - ❖ **CRISP-DM:** Provides a structured approach specifically for data mining.
    - ❖ **SEMA:** Emphasizes software engineering for machine learning applications.
- ✓ **KDD (Knowledge Discovery in Databases):**
    - ❖ **Pros:** Systematic approach, comprehensive stages from selection to knowledge extraction.
    - ❖ **Cons:** Can be complex and time-consuming, requiring significant domain expertise.
- ✓ **CRISP-DM (Cross-Industry Standard Process for Data Mining):**
    - ❖ **Pros:** Widely adopted, flexible, iterative, focuses on business understanding and deployment.
    - ❖ **Cons:** May lack detailed guidance for specific stages, potentially overwhelming for smaller projects.
- ✓ **SEMA (Sample, Explore, Modify, Model, Assess):**
    - ❖ **Pros:** Simple and straightforward, emphasizes iterative and interactive steps.
    - ❖ **Cons:** Less structured, possibly unsuitable for complex projects requiring rigorous processes.

- ➢ **Data Science Life Cycle:**
    - ❖ **Differences:** Encompasses data collection, preparation, analysis, and deployment.
    - ❖ **Pros:** Holistic approach, integrates data engineering, analysis, and machine learning; iterative and adaptable.
    - ❖ **Cons:** Resource-intensive, may require multidisciplinary expertise.
- ➢ **Business Intelligence Life Cycle:**
    - ❖ **Differences:** Focuses on data analysis and reporting for business decision-making.
    - ❖ **Pros:** Leverages data for strategic decision-making, integrates well with existing business processes.
    - ❖ **Cons:** Often relies on historical data, may lack predictive analytics capabilities.

2. Proposed Data Analytic Process

I propose using **CRISP-DM** for its flexibility, business relevance, and structured approach, ensuring alignment with strategic goals and delivering relevant insights.

3. **Most Important Stage of Data Analytics Lifecycle**

- **Data Cleaning and Preparation:** Ensuring data quality and relevance is crucial, as it impacts the accuracy and reliability of the entire analysis process. Without clean data, even sophisticated models will produce poor results.

3. **What's the business problem given the above data sources?**

Business Problem: Determine the impact of income on patients' diabetes health outcomes.

4. **Hands-On Exercise: Customer Churn Prediction**

   1. **Business Problem:** Predict which customers are likely to churn.
   2. **Stakeholders:** Marketing team, customer service team, management.
   3. **Sources of Data:** Secondary data from CRM systems and customer feedback.
   4. **Variables:**
      - ✓ **Dependent Variable:** Churn (Yes/No).
      - ✓ **Independent Variables:** Customer interactions, transaction history, service usage.
   5. **Preprocessing Techniques:** Apply imputation techniques for missing values.
   6. **Variable with Highest Effect:** Number of customer service calls.
   7. **Type of Analytics:** Predictive analytics to forecast customer churn.
   8. **Type of Machine Learning:** Supervised learning.
      - ✓ **Best Model:** Random Forest (handles feature interactions well, robust).
      - ✓ **Least Effective Model:** Logistic Regression (limited by linear assumptions).
   9. **Performance Improvement Techniques:**
      - ✓ **Techniques:** Hyper parameter tuning, cross-validation.
      - ✓ **Validation Metrics:** Confusion matrix, accuracy, precision, recall, F1 score.
      - ✓ **Deployment:** Model integrated into the customer management system for real-time churn prediction.