FDS Experiment no. :- 01

Name :- Om Sandip Jadhao

UID:- 2024310006

Question 1: - **List 10 Uses of Data Science for Industry**

1. **Fraud Detection**
   - Anomaly detection algorithms in transaction data.
   - Pattern recognition to predict fraudulent activities.

2. **Supply Chain Optimization**
   - Predictive models for demand forecasting.
   - Optimization algorithms for efficient resource allocation.

3. **Healthcare Diagnostics**
   - Analysis of patient data for diagnosis prediction.
   - Machine learning models for personalized treatment plans.

4. **Product Development**
   - Sentiment analysis from customer feedback.
   - Market trend analysis using machine learning.

5. **Risk Management**
   - Historical data analysis for risk assessment.
   - Predictive models for forecasting financial risks.

6. **Optimizing Marketing Campaigns**
   - Audience segmentation using clustering algorithms.
   - Marketing effectiveness analysis with A/B testing and predictive analytics.

7. **Churn Prediction**
   - Behavioural data analysis to predict churn.
   - Machine learning models to identify at-risk customers.

8. **Customer Personalization**
   - Data collection and analysis of customer behaviour.
   - Predictive analytics for personalized recommendations.

9. **Predictive Maintenance**
   - Sensor data and machine learning for equipment failure prediction.
   - Time-series analysis for maintenance scheduling.

10. **Automating Decision-Making**
    - Real-time data processing for automated decision models.
    - Machine learning algorithms for operational decision-making.

***Question 2*:- Identify Ten Dataset from Kaggle and which application are possible from them.**

1. Titanic - Machine Learning from Disaster
    1. Dataset: Titanic Machine Learning from Disaster
    2. Applications:
        - Predicting survival rates using classification models.
        - Feature engineering and missing data imputation.

2. Iris Flower Dataset
    1. Dataset: Iris Species
    2. Applications:
        - Classification of flower species using basic machine learning algorithms.
        - Exploratory data analysis and visualization.

3. Pima Indians Diabetes Database
    1. Dataset: Pima Indians Diabetes
    2. Applications:
        - Predicting the onset of diabetes using classification models.
        - Feature importance analysis and health risk assessment.

4. Wine Quality Dataset
    1. Dataset: Wine Quality
    2. Applications:
        - Predicting wine quality scores based on physicochemical properties.
        - Regression analysis and model evaluation.

5. NYC Taxi Trip Duration
    1. Dataset: NYC Taxi Trip Duration
    2. Applications:
        - Predicting taxi trip duration using regression models.
        - Time-series analysis and route optimization.

6. Student Performance Data
    1. Dataset: Student Performance Data
    2. Applications:
        - Predicting student performance based on socio-economic factors.
        - Correlation analysis and educational insights.

7. House Prices - Advanced Regression Techniques
    1. Dataset: House Prices
    2. Applications:
        - Predicting house prices using regression models.
        - Feature engineering and outlier detection.

8. Fake News Detection
    1. Dataset: Fake News Detection
    2. Applications:
        - Classifying news articles as fake or real using NLP techniques.
        - Text preprocessing and sentiment analysis.

9. COVID-19 World Vaccination Progress
    1. Dataset: COVID-19 World Vaccination Progress
    2. Applications:
        - Analysing global vaccination trends and predicting future vaccinations.
        - Time-series forecasting and public health insights.

10. Global Terrorism Database
    1. Dataset: Global Terrorism Database
    2. Applications:
        - Analysing global terrorism patterns and predicting future attacks.
        - Geographic data visualization and clustering.

Question 3:- **Identify a research paper from year2024 on data science and Ai write important from the paper**

1. **Objective**:
   - Develop a chatbot to simplify user interaction with complex datasets.
   - Democratize data analysis by making it accessible to users with varying levels of data science expertise.

2. Key Technologies:
   - Natural Language Processing (NLP): Uses tokenization, part-of-speech tagging, and syntactic analysis to understand and process user queries.
   - Natural Language Toolkit (NLTK): Essential for the chatbot's language processing abilities and enables continuous improvement through ongoing model training.
   - Machine Learning (ML): Integrated to provide data-driven insights, enhance recommendations, and improve over time.

3. System Architecture:
   - Questioner Module: Processes and interprets user input using NLP techniques.
   - Answer Module: Combines ML and data analysis to generate relevant responses based on user queries.
   - Chatbot Integration: Acts as the hub, combining both modules to facilitate seamless user interaction with data.

4. Main Features:
   - Natural language query handling.
   - Data-driven insights and recommendations.
   - Contextual and accurate responses.
   - Continuous improvement using ML models and NLTK.

5. Performance Results:
   - Accuracy of 88% with NLTK + NLP compared to 75% with NLP alone.
   - NLTK's linguistic analysis capabilities contribute to the improved accuracy in language processing.
   - User Experience:
   - The system is designed to be user-friendly, providing simple interfaces for users to engage with data through natural language, making data science more accessible.

6. Future Work:
   - Improve NLP algorithms to handle more complex queries.
   - Integrate deep learning models and multimodal interfaces (e.g., speech recognition).
   - Enhance data-driven recommendations and expand data sources