Data Science Research Exercise for MSc Students

Objective

To explore a data science application area, identify relevant datasets, examine applicable machine learning algorithms, and investigate the tools and frameworks suited to these algorithms. This exercise will deepen students' understanding of the data science project lifecycle from data acquisition to model implementation and evaluation.

Outline

1. **Application Area Identification**

Choose a specific application area within data science that aligns with your academic or career interests. Potential application areas include:

- Healthcare (e.g., disease diagnosis, medical imaging)
- Finance (e.g., fraud detection, credit scoring)
- E-commerce (e.g., recommender systems, customer segmentation)
- Social Media (e.g., sentiment analysis, trend forecasting)
- Climate Science (e.g., weather forecasting, environmental monitoring)

2. **Dataset Selection**

i. Identify and obtain a dataset that is relevant to your chosen application area. Look for publicly available datasets on platforms like **Kaggle**, **UCI Machine Learning Repository**, **Google Dataset Search**, or **GitHub** repositories.

ii. Provide a summary of the dataset, including:

- Source and description of the dataset
- Number of records and features
- Key attributes relevant to the application area

- Any ethical considerations or data quality issues

3. **Applicable Machine Learning Algorithms**

Research and identify machine learning algorithms that are commonly used in this application area. Provide a brief description of each algorithm, including:

- Purpose and typical use cases of the algorithm
- Why this algorithm is suitable for your application area
- The algorithm's strengths and limitations

Examples:

- **Classification**: Logistic regression, decision trees, random forests, support vector machines (SVM), neural networks,deep learning (CNN,RNN,GAN,Transformers,LSTM,GRU etc)
- **Clustering**: K-means, hierarchical clustering, DBSCAN
- **Regression**: Linear regression, Ridge regression, Lasso regression
- **Time Series**: ARIMA, Prophet, Long Short-Term Memory (LSTM) networks

4. **Machine Learning Frameworks and Tools**

Investigate the frameworks and tools that support the implementation of the chosen algorithms. Briefly describe each framework or tool's purpose and role in the data science pipeline for your chosen application area.

These may include:

- **Libraries**: TensorFlow, Keras, PyTorch, Scikit-Learn, XGBoost
- **Data Processing**: Pandas, NumPy, Spark

- **Visualization**: Matplotlib, Seaborn, Plotly
- **Workflow Management**: Jupyter Notebook, Google Colab, Docker for environment management, MLflow for experiment tracking

5. **Application-Specific Challenges and Considerations**

Identify and discuss any unique challenges or considerations in your chosen application area, such as:

- Data quality and preprocessing needs
- Ethical concerns (e.g., bias, fairness)
- Interpretability and explainability of machine learning models
- Scalability and deployment in production environments

6. **Final Report and Presentation**

- Compile your findings into a structured report, with sections for each part of the exercise (Application Area, Dataset Summary, Algorithms, Frameworks and Tools, Challenges and Considerations).
- Prepare a 10-minute presentation summarizing your research. Focus on explaining why the application area, algorithms, and tools were chosen, and highlight any challenges encountered in data selection or algorithm suitability.

Final Document Submission:

- Revise your assignment based on the key topics outlined.
- Include any results from your analysis or experiments, explaining how the AI methods were applied and what the results indicate.

- Submit your assignment with proper references in a recognized citation style (e.g., APA, MLA, Chicago).