UN3106 Applied Machine Learning – Final Project

Welcome to the final project of our Applied Machine Learning class! Throughout this course, you have been exploring various algorithms, techniques, and applications. Now, it's time to put your knowledge and skills to the test with a hands-on project that integrates all aspects of the machine learning pipeline.

This final project is designed to provide you with a comprehensive opportunity to showcase your understanding of the machine learning process, from data pre-processing to model building and evaluation, and to demonstrate your ability to apply machine learning concepts in a practical setting. This project is not just a culmination of everything you've learned; it's a chance to showcase your skills, creativity, and problem-solving abilities in a real-world context.

By the end of this project, you will have not only developed technical proficiency in machine learning and web development but also gained valuable experience in problem-solving, critical thinking, and communication – essential skills for any data scientist or machine learning practitioner.

Project Tasks:

1. Data Acquisition and Pre-processing:

- **Data Selection:** Choose a dataset(s) that aligns with your interests and goals for the project. Consider datasets from reputable sources such as Kaggle , UCI Machine Learning Repository, or governmental data portals (e.g., data.gov).
- **Data Exploration:** Perform exploratory data analysis (EDA) to gain insights into the dataset's characteristics, including data types, distributions, and potential relationships between variables. Utilize visualizations (e.g., histograms, scatter plots etc.) to uncover patterns and anomalies.
- **Data Pre-processing:** Cleanse the data by handling missing values, outliers, and inconsistencies. Apply appropriate techniques for data transformation (e.g., normalization, standardization etc.) to ensure consistency and improve model performance.
- **Feature Engineering:** Extract relevant features from the dataset and create new features that may enhance predictive power. This could involve techniques such as one-hot or dummy encoding for categorical variables, deriving new features from existing ones, removing near-zero variance variables etc.

2. **Model Building:**
    - **Algorithm Selection:** Choose at least three machine learning algorithms that are suitable for the chosen dataset(s) and problem. Consider a diverse range of algorithms, including but not limited to decision trees, random forests, support vector machines, neural networks, clustering, stacked models etc.
    - **Model Training:** Split the dataset into training and testing sets using an appropriate sampling method and ratio (e.g., 70:30, 80:20 etc.). Train each selected model on the training data using the chosen algorithms, tuning hyperparameters as necessary to optimize model performance. For every model, identify a selection of predictors that wield significant influence in constructing the model.
    - **Model Evaluation:** Evaluate the trained models using relevant performance metrics (e.g., accuracy, precision, recall, AUC, MSE etc.). Compare the performance of different models and identify the most effective one for your problem domain.

3. **Web Application Development with R-Shiny:**
    - **UI Design:** Design a user-friendly and visually appealing interface for your web application using the R Shiny package. Consider the target audience and aim to create an intuitive layout with clear navigation and interactive elements.
    - **Integration of Models:** Integrate the predictive models into the web application, allowing users to input data, perform data-splitting and data pre-processing, build models, and receive predictions in real-time. Ensure seamless communication between the front-end interface and the back-end model functionality.
    - **Visualization and Interpretation:** Provide visualizations and explanations that enhance the user's understanding of the model predictions. Use plots, tables, or interactive graphics to convey insights derived from the data and model outputs.

Deliverables [due on May 5th at 11:59PM]:

- R Markdown documents containing your code, including data pre-processing, model training & evaluation, and web application.
- A deployed R Shiny web application accessible via a URL.
- A written report summarizing each project task.
- Original dataset(s)

Evaluation Rubrics:

1. Data Acquisition and Pre-processing [0 – 32pt]:

- **Data Selection [0 – 8pt]:**
  - **Basic [2pt]:** Chooses a dataset without considering its complexity or data quality. Dataset has minimal/no data quality issues, making it relatively straightforward to work with.
  - **Intermediate [5pt]:** Selects a dataset with moderate complexity and some data quality issues. Demonstrates awareness of basic dataset.
  - **Advanced [8pt]:** Selects a dataset with high complexity and significant data quality challenges.
- **Data Exploration [0 – 8pt]:**
  - **Basic [2pt]:** Performs basic data exploration with limited visualizations and insights.
  - **Intermediate [5pt]:** Conducts thorough data exploration with appropriate visualizations, identifying key patterns and trends.
  - **Advanced [8pt]:** Conducts extensive data exploration, utilizing advanced visualizations and statistical techniques to uncover nuanced insights and relationships.
- **Data Pre-processing [0 – 8pt]:**
  - **Basic [2pt]:** Implements basic data cleaning techniques but overlooks some issues.
  - **Intermediate [5pt]:** Applies standard data preprocessing techniques effectively, handling most issues and optimizing data for modeling.
  - **Advanced [8pt]:** Implements advanced data preprocessing techniques with meticulous attention to detail, effectively addressing all data quality issues and optimizing data for optimal model performance.
- **Feature Engineering [0 – 8pt]:**
  - **Basic [2pt]:** Implements basic/no feature engineering techniques without considering the full potential of feature manipulation.
  - **Intermediate [5pt]:** Applies standard feature engineering techniques effectively, including normalization and standardization, and handling missing values appropriately.
  - **Advanced [8pt]:** Implements advanced feature engineering techniques with a high level of creativity and sophistication. Demonstrates a deep understanding of feature engineering principles and applies them effectively to enhance the predictive power of the dataset.

2. **Model Building [0 – 30pt]:**
   - **Algorithm Selection [0 – 10pt]:**
     - **Basic [3pt]:** Selects common algorithms without clear justification.
     - **Intermediate [6pt]:** Selects appropriate algorithms based on problem requirements and dataset characteristics, providing basic justification.
     - **Advanced [10pt]:** Selects algorithms strategically, considering a range of options and justifying selections based on thorough analysis and domain knowledge.
   - **Model Training [0 – 10pt]:**
     - **Basic [3pt]:** Splits data into training and testing sets but may not optimize hyperparameters or train/evaluate models thoroughly.
     - **Intermediate [6pt]:** Splits data appropriately, trains models effectively, but does not optimize hyperparameters to improve performance.
     - **Advanced [10pt]:** Implements advanced techniques for model training, performs hyperparameter tuning, optimizes models for maximum performance, and identifies influential predictors for every model.
   - **Model Evaluation [0 – 10pt]:**
     - **Basic [3pt]:** Evaluates models using basic metrics without in-depth analysis or comparison.
     - **Intermediate [6pt]:** Evaluates models using relevant metrics and compares performance on the testing set, providing some interpretation.
     - **Advanced [10pt]:** Evaluates models comprehensively, considering a range of metrics and providing insightful interpretation of results, including strengths and weaknesses of each model and metric used.

3. **Web Application Development with R-Shiny [0 – 30pt]:**
   - **UI Design [0 – 10pt]:**
     - **Basic [3pt]:** Creates a basic UI with limited functionality and aesthetics.
     - **Intermediate [6pt]:** Designs an intuitive UI with clear layout and interactivity, meeting basic user needs.
     - **Advanced [10pt]:** Designs a visually appealing and user-friendly UI with advanced features and interactive elements, enhancing user experience.
   - **Integration of Models [0 – 10pt]:**
     - **Basic [3pt]:** Integrates models into the web application but lacks seamless communication or real-time functionality. Data input, data-splitting, and data pre-processing options are not available.
     - **Intermediate [6pt]:** Successfully integrates models, allowing users to input data and receive predictions with minimal latency.

    - **Advanced [10pt]:** Implements advanced features for model integration, ensuring seamless communication and real-time predictions, enhancing overall application performance. Data input, data-splitting, and data pre-processing options are available.
- **Visualization and Interpretation [0 – 10pt]:**
    - **Basic [3pt]:** Presents model predictions without clear visualizations or explanations.
    - **Intermediate [6pt]:** Provides visualizations to aid interpretation of model outputs, offering basic insights into predictions.
    - **Advanced [10pt]:** Creates compelling visualizations and explanations that enhance user understanding of model predictions, facilitating informed decision-making.

4. Written Report [0 – 15pt]:

- **Basic [5pt]:** Provides a simple written report that lacks detail and coherence; lacks clarity in conveying key points; fails to provide sufficient explanation or analysis of the project tasks, methodologies used, and findings.
- **Intermediate [10pt]:** Prepares a clear and organized written report that effectively summarizes the project's approach, methodologies, and findings. Presents information in a logical manner, provides a moderate level of detail and analysis, offering some insight into project tasks and outcomes.
- **Advanced [15pt]:** Produces a comprehensive and well-articulated written report that demonstrates a deep understanding of the project tasks and methodologies. Presents detailed explanations and insightful analysis of each project task, including challenges faced and solutions implemented. Communicates findings effectively, providing clear conclusions.