CSE4288 Introduction to Machine Learning – Team Project – Fall 2024

Project Duration: 7 weeks

Project Proposal: end of Week 7

Progress Presentation: Week 10
Presentation: Week 13
Team Size: 5 students per team

Objective

The objective of this project is to provide hands-on experience in applying machine learning techniques to solve real-world problems. By working in teams, you will:

- Understand the end-to-end process of a machine learning project.
- Enhance your skills in data preprocessing, model development, and evaluation.
- Learn to collaborate effectively within a team.
- Present your findings professionally.

Project Overview

Each team will:

1. **Select a Problem and Dataset:**
    - Choose a real-world problem that can be addressed using machine learning.
    - Find relevant and publicly available datasets suitable for the problem.
    - Define clear objectives and scope for your project.
2. **Data Preprocessing and Exploration:**
    - Clean and preprocess the data (handle missing values, encode categorical variables, normalize features, etc.).
    - Perform exploratory data analysis (EDA) to understand data distributions, relationships, and patterns.
    - Visualize data using appropriate charts and graphs.

3. **Model Development:**
    - Choose suitable machine learning algorithms for your problem (e.g., regression, classification, clustering, etc.).
    - Implement the models using Python and appropriate libraries (e.g., scikit-learn, TensorFlow, Keras).
    - Split your data into training and testing sets (consider using cross-validation).
4. **Model Evaluation and Optimization:**
    - Evaluate your models using appropriate metrics (e.g., accuracy, precision, recall, F-Score, RMSE, MAPE, etc.).
    - Compare different models and select the best-performing one.
    - Perform hyperparameter tuning to optimize your model.
    - Analyze errors and discuss potential improvements.
5. **Documentation and Presentation:**
    - Document all stages of your project in a comprehensive report.
    - Prepare a presentation summarizing your project.

Project Timeline

Week 1: Project Initiation

- **Tasks:**
    - Form teams of five students.
    - Brainstorm and select a project topic and dataset.
    - Open a GitHub repo named CSE4288F24_GrpX, share with team members and your Professor (mcganiz) and your TA
    - Submit a project proposal for approval.
- **Deliverables:**
    - **Project Proposal Document** including:
        - Team members' names and contact information.
        - Project title and description.
        - Problem statement and objectives.
        - Brief overview of the dataset.
        - Proposed methodologies and timeline.
- **Due Date:** End of Week 7.

Week 2: Data Preprocessing and EDA

- **Tasks:**
    - Acquire and understand your dataset.
    - Perform data cleaning and preprocessing.
    - Conduct exploratory data analysis.
    - Identify key features and consider feature engineering.
- **Deliverables:**
    - **Data Preprocessing Report** including:

        - Description of data cleaning steps.
        - EDA findings with visualizations.
        - Rationale for feature selection or engineering.
- **Due Date:** End of Week 9.

Week 3: Model Development

- **Tasks:**
    - Select appropriate machine learning algorithms.
    - Implement initial models.
    - Train models using the training data.
- **Deliverables:**
    - **Initial Model Implementation Code** with comments.
    - **Progress Report** summarizing work done and challenges faced.
- **Due Date:** End of Week 10.

Week 4: Model Evaluation and Optimization

- **Tasks:**
    - Evaluate model performance using test data.
    - Experiment with different algorithms and hyperparameters.
    - Optimize your model based on evaluation results.
- **Deliverables:**
    - **Model Evaluation Report** including:
        - Performance metrics.
        - Comparison of different models.
        - Description of optimization techniques used.
- **Due Date:** End of Week 11.

Week 5: Finalization and Presentation Preparation

- **Tasks:**
    - Finalize your model and results.
    - Compile all work into a comprehensive project report.
    - Prepare presentation slides.
    - Rehearse the presentation as a team.
- **Deliverables:**
    - **Final Project Report**.
    - **Presentation Slides**.
- **Due Date:** End of Week 12.

Project Deliverables

1. Project Proposal

- **Content:**
  - Clear problem statement and objectives.
  - Dataset description and source.
  - Planned methodology and tools.
  - Team roles and responsibilities.
  - Timeline and milestones.
- **Format:** PDF document named `CSE4288_F24_GrpX_Project_Proposal.pdf`.

2. Data Preprocessing Report (Week 2)

- **Content:**
  - Detailed data cleaning and preprocessing steps.
  - EDA results with charts and graphs.
  - Feature selection and engineering decisions.
- **Format:** PDF document named `CSE4288_F24_GrpX_Data_Preprocessing_Report.pdf`.

3. Progress Report and Code

- **Content:**
  - Summary of modeling work completed.
  - Challenges encountered and solutions.
  - Code files with proper documentation.
- **Format:** PDF report and code files in GitHub repo named CSE4288F24_GrpX `CSE4288_F24_GrpX_Model_Development.pdf`

4. Model Evaluation Report

- **Content:**
  - Evaluation metrics and results.
  - Comparison between models.
  - Optimization steps taken.
- **Format:** PDF document named `CSE4288_F24_GrpX_Model_Evaluation_Report.pdf`.

5. Final Project Report

- **Content:**
  - **Abstract:** Brief summary of the project.
  - **Introduction:** Background and significance of the problem.
  - **Methodology:** Detailed explanation of data preprocessing, modeling, and evaluation methods.
  - **Results:** Presentation of findings with tables and figures.
  - **Discussion:** Interpretation of results, challenges faced, and how they were addressed.

    - **Conclusion:** Summary of work and potential future improvements.
    - **References:** List of sources cited.
    - **Appendices:** Additional material (e.g., code snippets, detailed tables).
- **Format:** PDF document named `CSE4288_F24_GrpX_Final_Project_Report.pdf`.

6. Presentation Slides

- **Content:**
    - Key points from each section of the project.
    - Visual aids (graphs, charts, images).
    - Speaker notes or prompts.
- **Format:** PowerPoint or PDF named `CSE4288_F24_GrpX_Presentation_Slides.pdf`

Presentation Details

- **Duration:** 15 minutes presentation + 5 minutes Q&A.
- **All team members** must participate in the presentation.
- **Content to Cover:**
    - Introduction to the problem and its relevance.
    - Overview of the dataset and key findings from EDA.
    - Description of the models used and why they were chosen.
    - Summary of results and evaluation metrics.
    - Challenges faced and lessons learned.
    - Conclusion and potential future work.

Assessment Criteria

Your project will be evaluated based on the following criteria:

1. Understanding of the Problem (5%)

- Clarity and relevance of the problem statement.
- Appropriateness of the chosen dataset.

2. Data Preprocessing and EDA (20%)

- Effectiveness of data cleaning and handling of missing values.
- Depth of exploratory data analysis.
- Justification for feature selection and engineering.

3. Model Development (25%)

- Correct implementation of machine learning algorithms.
- Appropriateness of chosen models for the problem.
- Proper use of training and validation techniques.

4. Model Evaluation and Optimization (30%)

- Use of appropriate performance metrics.
- Depth of model comparison and analysis.
- Effectiveness of optimization strategies.

5. Documentation and Reporting (10%)

- Quality and clarity of the final report.
- Logical structure and flow of information.
- Proper citation of sources and adherence to academic standards.

6. Presentation (10%)

- Clarity and professionalism of the presentation.
- Engagement with the audience during Q&A.
- Team coordination and time management.

Suggested Project Ideas

While you are free to choose any relevant machine learning problem, here are some suggestions:

1. Predictive Analytics

- **House Price Prediction:** Using datasets like the Boston Housing dataset to predict house prices based on various features.
- **Stock Market Analysis:** Predicting stock prices or market trends using historical data.

2. Classification Tasks

- **Spam Detection:** Classify emails as spam or not spam.
- **Image Classification:** Recognize handwritten digits using the MNIST dataset.

3. Clustering and Segmentation

- **Customer Segmentation:** Group customers based on purchasing behavior.
- **Anomaly Detection:** Identify fraudulent transactions in financial datasets.

4. Natural Language Processing

- **Sentiment Analysis:** Determine the sentiment of movie reviews or social media posts.
- **Topic Modeling:** Identify topics in a collection of documents.
- **Fine**-tuning LLM models for specific tasks

5. Time Series Analysis

- **Weather Forecasting:** Predict future weather conditions based on historical data.
- **Energy Consumption:** Forecast energy usage in buildings or cities.

Guidelines and Expectations

- **Team Collaboration:**
  - Allocate tasks according to each member's strengths.
  - Maintain regular communication and document meetings.
- **Use of Tools and Libraries:**
  - Python is recommended for implementation.
  - Libraries such as NumPy, pandas, scikit-learn, TensorFlow, and Keras are allowed.
  - Ensure all code is original and not plagiarized.
- **Code Documentation:**
  - Comment your code thoroughly.
  - Use meaningful variable names and follow coding standards.
- **Data Ethics:**
  - Ensure you have the right to use the dataset.
  - Avoid using sensitive or confidential data without proper authorization.
- **Academic Integrity:**
  - All work must be your own.
  - Plagiarism or academic dishonesty will result in severe penalties.

Submission Instructions

- **Deadline:** All deliverables are due by 11:59 PM on the specified due dates assigned in Google classroom.
- **Format:**
  - Submit all documents in PDF format.
  - Code should be submitted as .py files or Jupyter notebooks (.ipynb) into project group’s GitHub repo.
  - Include a README file explaining how to run your code.
- **Submission Platform:**
  - Upload your files to the course's learning management system under the appropriate assignment links.

- **File Naming Convention:**
  - Use the provided naming conventions for all files.
  - Example: `CSE4288_F24_GrpX_DocumentName.pdf` or `CSE4288_F24_GrpX_CodeFile.py`.

Support and Resources

- **Instructor and TA:**
  - Available during office hours for guidance and support.
- **Discussion Forum:**
  - Use Google classroom for all project related discussion.
- **Recommended Resources:**
  - **Books:**
    - Introduction to Machine Learning by Ethem Alpaydin.
    - Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron.
    - Machine Learning by Tom Mitchell (https://www.cs.cmu.edu/~tom/mlbook.html)
    - "Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville" (https://www.deeplearningbook.org/)
    - "Speech and Language Processing (3rd ed. draft) by Dan Jurafsky and James H. Martin" (https://web.stanford.edu/~jurafsky/slp3/ed3bookaug20_2024.pdf)
    -
  - **Online Courses:**
    - Coursera, edX, and other platforms offer courses that might be helpful.
  - **Datasets:**
    - UCI Machine Learning Repository.
    - Kaggle Datasets.
    - HuggingFace
  - Possible Journals/Conferences/Resources that might be helpful for topic selection
  - You can Access the following journal through your Marmara University library account and the web page of electronic databases (https://katalog.marmara.edu.tr/vetisbt/)
    - 1. IEEE Transactions on Pattern Analysis and Machine Intelligence
    - 2. IEEE Transactions on Image Processing
    - 3. Pattern Recognition
    - 4. Pattern Recognition Letters
    - 5. International Journal of Pattern Recognition and Artificial Intelligence
    - 6. International Conference on Pattern Recognition (ICPR)
    - 7. Neural Information Processing Systems (NEURIPS)
    - 8. International Conference on Learning Representations (ICLR)
  - Kaggle challenges: A machine learning competition website: https://www.kaggle.com/competitions Note: You can use the datasets on Kaggle

but you can not use the code available at Kaggle!! You need to write your own code.
- https://ibug.doc.ic.ac.uk/courses/machine-learning-course-395/
- CS229 Machine Learning, Stanford University: https://cs229.stanford.edu/
- Machine Learning Course, University of Washington https://courses.cs.washington.edu/courses/cse446/17wi/project.html

---

Frequently Asked Questions

Q: Can we use data from multiple sources?
A: Yes, as long as it is relevant and you properly preprocess and integrate the data.

Q: Is it mandatory to try multiple machine learning algorithms?
A: Yes, you must use multiple ML algoritms to train and compare different models to find the most suitable one.

Q: Can we use deep learning models?
A: Yes, if appropriate for your problem and you can justify their use.

Q: How much emphasis should be placed on model accuracy versus understanding?
A: Both are important. While achieving high accuracy is desirable, understanding and explaining your approach is crucial.

Good luck with your project! This is an excellent opportunity to apply what you've learned and gain practical experience in machine learning. Be creative, collaborative, and diligent.