This is a team project that is designed to provide hands-on experience in building a complete machine learning pipeline, from selecting a real-world problem to deploy and evaluating a solution. You will collaborate with your team to apply the techniques you’ve learned through the course, including data processing, model development, and ethitical consideration.

This project will not only asses your technical skills but also your ability to work collaboratively and address the ethitical implications of using machine learning in real world application.

Task Description

1. Problem Section
2. Select a real-word problem relevant to the domains covered in the course, such as finance, healthcare, marketing, etc.
3. Justify your choice of problem and discuss its relevance to the industry and society
4. Data Collection and Preparation
5. Use existing data set relevant to the selected problem.
6. Apply techniques such as dimensionality reduction (PCA, t-SNE) if necessary for feature optimization.
7. Model Development.
8. Develop multiple machine learning models (both supervised and unsupervised, as is applicable to your problem.)
9. Techniques such as train-set splits, cross-validation, hyperparameter tuning, and assemble learning (eg Random forests. Gradient boosting.)
10. For unsupervised planning problems, implement clustering algorithm and evaluate their effectiveness.
11. Model evaluation and comparison.
12. Evaluate the performance of each model using appropriate metrics (eg K-means recall, F1-score or clustering evaluation metrics)
13. The models and select the best performing best models in the evaluation metrics.
14. Deployment considerations.
15. Propose a strategy for deploying the chosen model in a real-world scenario.
16. Discuss consideration for scalability and model maintenance, monitoring and production environment.
17. Ethical implications.
18. Analyse the ethical consideration of your machine learning solution, including issues related to data privacy, algorithmic bias, fairness, and societal impact of your model.

7. Final Reports.

1. Submit a comprehensive report (2000-2500 words) detailing the entire project lifecycle:
2. Problem selection and relevance.
3. Data collection and preparation.
4. Model development and evaluation.
5. Ethical implications.
6. A section describing the contribution of each team member.

Healthcare: Disease Prediction or Early Detection

Problem: Predicting heart disease risk or diabetes progression

Potential Dataset:

Kaggle's "Heart Disease Dataset"

"Diabetes 130-US hospitals for years 1999-2008" dataset

Relevance: Critical for preventive healthcare and personalized medicine