Project: Predictive Modeling on a Real-World Dataset

Description:

For your final project in the Machine Learning course, you will have the opportunity to apply your knowledge and skills in machine learning to a real-world dataset of your choice. This project will allow you to demonstrate your understanding of the entire machine learning workflow, from data preprocessing to model selection and evaluation.

Project Tasks:

1. Dataset Selection:

- Select a real-world dataset that aligns with your interests or domain of study. It could be related to healthcare, finance, social sciences, or any other field that you find intriguing.
- Ensure that the dataset is sufficiently large and diverse to allow for meaningful analysis and modeling.

2. Data Preprocessing:

- Perform comprehensive data preprocessing and cleaning on the selected dataset.
- Handle missing values, encode categorical variables, and scale/normalize numerical features as necessary.
- Utilize exploratory data analysis (EDA) techniques to gain insights into the dataset.

3. Feature Engineering and Selection:

- Explore feature engineering techniques to create new relevant features from the existing dataset.
- Apply feature selection methods such as correlation analysis, statistical tests, or dimensionality reduction techniques to select the most informative features.

4. Model Building and Evaluation:

- Implement and evaluate different machine learning algorithms for your chosen problem.
- Use appropriate evaluation metrics based on the problem type (classification, regression, etc.).
- Utilize techniques such as cross-validation and hyperparameter tuning to optimize your models.

5. Model Comparison and Selection:

- Compare the performance of different models and select the best-performing model based on evaluation metrics.
- Provide a clear rationale for your selection, considering factors such as interpretability, complexity, and generalization ability.

6. Model Presentation:

- Prepare a presentation summarizing your project, including the problem statement, dataset description, data preprocessing steps, model selection, evaluation results, and insights gained from the project.
- Highlight any challenges you faced during the project and discuss potential areas for further improvement.
- Show Andreas the finished notebook on the 9th of June

This project is your opportunity to showcase your understanding of the entire machine learning process. It encourages creativity, critical thinking, and problem-solving skills while working on a real-world problem of your choice. Feel free to explore different techniques, algorithms, and strategies to achieve the best possible model performance. Your presentation will allow you to demonstrate your project outcomes and share your insights.

Please note that the project will be evaluated based on the clarity and organization of your presentation, the quality of your code, the effectiveness of your modeling techniques, and your ability to explain the rationale behind your decisions.

Good luck, and have fun exploring and building predictive models on your chosen real-world dataset!

Hand-in:

One well-commented notebook (.ipynb-file) with all the relevant steps included. It should include markdown, describing the steps and choices, all the code, as well as the names of any team member(s). 1-2 team members per notebook!