Project Proposal

Project Title: Scikit-Learn Recreation and Implementation

Prepared By: Group-5

Salman Shafi – 2211386042

Jarinah Tasnim-2121026642

Jesmin Akter-2111382642

Sazid Rahman Bhuiyan-2131267642

Date: June 4, 2025

Situation/Problem/Opportunity:

Machine learning has become a cornerstone of modern data science and artificial intelligence applications. Scikit-learn stands as one of the most widely used and comprehensive machine learning libraries in Python, providing robust implementations of numerous algorithms for classification, regression, clustering, and data preprocessing. However, understanding the underlying mechanisms and mathematical foundations of these algorithms often requires hands-on implementation experience. Currently, students and practitioners rely heavily on pre-built libraries without fully comprehending the internal workings of machine learning algorithms.

Purpose Statement (Goals):

Working in coordination with faculty requirements and academic objectives, this project will engage in a comprehensive recreation of the scikit-learn library functionality. The project scope includes:

- **Phase 1:** Implementation of core scikit-learn examples across all major categories
- **Phase 2:** Complete recreation of scikit-learn library from scratch
- **Command-line interface:** All functionality accessible through terminal/command prompt
- **Educational focus:** Deep understanding of machine learning algorithm implementations

A key deliverable of this work is a comprehensive machine learning library that can be executed entirely through command-line interface, designed to integrate with local development environments and provide educational insight into algorithm mechanics.

This project will establish and publish a complete machine learning framework that ensures consistency across all implemented algorithms. The scope includes mathematical foundations, algorithm implementations, data preprocessing capabilities, and compliance with standard machine learning practices.

Objectives/Deliverables:

Deliverables for this project include:

• **Phase 1 Deliverables:**

- Complete implementation of scikit-learn classification examples (SVM, Random Forest, Logistic Regression, etc.)
- Regression algorithm examples (Linear Regression, Ridge, Lasso, etc.)
- Clustering algorithm implementations (K-Means, DBSCAN, Hierarchical Clustering)
- Dimensionality reduction techniques (PCA, t-SNE, LDA)
- Model selection and validation methods (Cross-validation, Grid Search)
- Data preprocessing utilities (Scaling, Encoding, Feature Selection)

• **Phase 2 Deliverables:**

- Complete recreation of scikit-learn library architecture
- Custom implementation of all major machine learning algorithms
- Comprehensive command-line interface for all functionalities
- Documentation and user guides for the recreated library
- Performance benchmarking against original scikit-learn
- Implementation plan for adoption of the custom library

Methods/Approach:

The following strategies will be implemented to ensure the success of this project:

• **Collaborative Planning:** Coordinate with faculty members, academic advisors, and project stakeholders to establish clear objectives and make key decisions

• **Systematic Implementation:** Work systematically through scikit-learn documentation and examples, starting with foundational algorithms and progressing to advanced techniques

• **Algorithm Study:** Partner with mathematical foundations to understand core algorithms, conduct detailed analysis of existing implementations, and develop optimized versions

• **Command-Line Development:** Design and implement a comprehensive command-line interface that provides easy access to all machine learning functionalities

• **Testing and Validation:** Implement rigorous testing procedures to ensure algorithm correctness, performance optimization, and compatibility with standard datasets

• **Documentation Focus:** Maintain detailed documentation throughout development, focusing on educational value and practical implementation guides

---

Success Criteria:

This project will be successful if the following conditions are created:

• **Phase 1 Completion:** All major scikit-learn examples are successfully implemented and can be executed via command prompt with identical or superior results

• **Phase 2 Achievement:** Complete recreation of scikit-learn functionality with custom implementations that demonstrate understanding of underlying algorithms

• **Performance Standards:** Custom implementations achieve comparable accuracy and performance metrics to original scikit-learn library

• **Usability Requirements:** Command-line interface provides intuitive access to all functionalities with comprehensive help documentation

• **Educational Value:** Final deliverable serves as an effective learning tool for understanding machine learning algorithm implementations

• **Code Quality:** All implementations follow best practices for code organization, documentation, and maintainability