

MIDOCEAN UNIVERSITY – UAE

College of Informatics Technology

Master of Science in Artificial Intelligence

**Systematic Algorithm Selection Framework for CO₂
Emissions Prediction in Saudi Arabia**

Task 5 – Detailed Action Plan for the Project

Prepared by:

Mohammad Wajeih Bader

Student ID: 251000462

Supervised by:

Dr. Amjad Qataish

July 2025

Contents

Executive Summary	3
Literature Review and Research Gap	3
Current State Analysis	3
Identified Research Gap	3
Research Framework	4
Problem Statement	4
Research Questions	4
Research Objectives	4
Research Problem-Questions-Objectives Integration	5
Methodology	5
Data Sources and Processing	5
Algorithm Selection Framework.....	6
Validation and Evaluation	7
Implementation Schedule.....	7
Project Gantt Chart.....	8
Key Milestones.....	8
Quality Assurance	8
Expected Outcomes and Risk Assessment	9
Primary Deliverables	9
Risk Assessment	9
Success Probability	9
Technical Requirements	10
Development Environment	10
Data Storage and Management	10
Version Control and Collaboration	10
Documentation Standards.....	10
Success Factors	10

Executive Summary

This 4-month research project develops a systematic algorithm selection framework for CO2 emissions prediction in Saudi Arabia, addressing inconsistent algorithm selection practices in regional emissions modeling. The framework provides evidence-based guidelines for selecting and optimizing machine learning algorithms, tailored to Saudi Arabia's unique environmental and economic context.

Single Research Contribution:

"Systematic Algorithm Selection and Optimization Framework for Regional CO2 Emissions Prediction" - A standardized methodology for algorithm evaluation, optimization, and selection, customized for Saudi Arabian conditions, integrating regional data characteristics and practical implementation guidelines.

Research Objective:

Develop and validate a systematic framework through comprehensive testing of four machine learning algorithms (XGBoost, LightGBM, Random Forest, CatBoost), aiming to approach or exceed the 93.13% baseline accuracy (Nassef et al., 2023) while establishing clear selection guidelines.

Expected Value:

The framework offers researchers a replicable methodology, reducing trial-and-error in algorithm selection and setting performance benchmarks for regional emissions prediction studies, with potential applications in environmental policy support.

Approach:

Systematic testing and optimization of four algorithms using standardized procedures, statistical comparison, sensitivity analysis, and comprehensive documentation, with a focus on Saudi-specific data characteristics.

Literature Review and Research Gap

Current State Analysis

Review of CO2 emissions prediction studies reveals inconsistent algorithm selection practices:

Saudi-Specific Studies:

- Nassef et al. (2023): 93.13% accuracy using ensemble approach (FFNN, ANFIS, LSTM)
- Rahman et al. (2023): 89% accuracy using Decision Tree ensemble
- Haq (2022): 98% accuracy using Random Forest for agricultural applications

International Benchmarks:

- Johnson et al. (2024): 94% accuracy using XGBoost (USA)
- Patel & Kumar (2023): 92% accuracy using LightGBM
- Liu et al. (2023): 94% accuracy using CNN-LSTM (China)

Identified Research Gap

Problem: Regional studies achieve varying performance (70-98%) with different algorithms but lack systematic comparison and selection methodology, particularly in Saudi Arabia's unique context (e.g., oil-driven economy, desert climate).

Evidence:

- Most studies test 1-2 algorithms without systematic comparison
- Limited justification for algorithm choice
- Inconsistent optimization approaches across studies
- No standardized framework for algorithm selection incorporating regional characteristics

Opportunity: Develop a systematic framework for algorithm selection and optimization, incorporating Saudi-specific data characteristics (e.g., oil production, industrial energy consumption), to improve consistency and performance in regional emissions prediction.

Research Framework

Problem Statement

Current CO₂ emissions prediction studies for Saudi Arabia use inconsistent algorithm selection approaches, limiting the ability to identify optimal methods and achieve consistent performance for regional environmental modeling in the context of Saudi Arabia's oil-dependent economy and environmental conditions.

Research Questions

Primary: What systematic framework can be developed for selecting and optimizing machine learning algorithms to achieve optimal CO₂ emissions prediction accuracy for Saudi Arabian conditions?

Supporting Questions:

1. Which algorithms (XGBoost, LightGBM, Random Forest, CatBoost) perform best for Saudi emissions data, considering regional factors like oil production and energy consumption?
2. What systematic optimization procedures yield optimal performance for each algorithm under computational constraints?
3. What evidence-based guidelines can be established for algorithm selection in regional studies?

Research Objectives

Primary Objective: Develop and validate a systematic algorithm selection framework that integrates testing procedures, optimization methodology, evidence-based selection guidelines, performance benchmarks, and practical implementation documentation, tailored to Saudi Arabian conditions.

Framework Components:

- Systematic algorithm testing procedure
- Standardized optimization methodology
- Evidence-based selection guidelines incorporating regional factors
- Practical implementation documentation and best practices
- Performance benchmarks with sensitivity analysis for robustness

Success Criteria:

- Targeting performance improvement beyond the 93.13% baseline where feasible, while focusing on methodological rigor
- Clear algorithm ranking with appropriate statistical comparison
- Practical implementation documentation enabling replication
- Evidence-based guidelines for future algorithm selection
- Sensitivity analysis demonstrating framework robustness under varying conditions

Research Problem-Questions-Objectives Integration

Research Problem	Research Questions	Research Objectives
Primary Problem: Current CO2 emissions prediction studies for Saudi Arabia use inconsistent algorithm selection approaches, limiting the ability to identify optimal methods and achieve consistent performance for regional environmental modeling	Primary Question: What systematic framework can be developed for selecting and optimizing machine learning algorithms to achieve optimal CO2 emissions prediction accuracy for Saudi Arabian conditions?	Primary Objective: Develop and validate a systematic algorithm selection framework that integrates testing procedures, optimization methodology, evidence-based selection guidelines, performance benchmarks, and practical implementation documentation into an adaptable methodology for regional CO2 emissions prediction
Sub-problems: • Performance variation (70-98%) without systematic comparison • Limited justification for algorithm choice • Inconsistent optimization approaches • No standardized framework for selection • Lack of regional context integration	Sub-questions: 1. Which algorithms perform best for Saudi emissions data, considering regional factors? 2. What systematic optimization procedures yield optimal performance? 3. What evidence-based guidelines can be established for algorithm selection?	Framework Elements: • Algorithm testing component • Optimization methodology component • Selection guidelines component • Documentation component • Performance evaluation and sensitivity analysis component <i>All elements serve the single framework contribution</i>

Logical Connection:

- **Problem** identifies inconsistency in algorithm selection and lack of regional context
- **Questions** explore systematic solutions and regional adaptation
- **Single Objective** defines a comprehensive framework addressing all issues

Methodology

Data Sources and Processing

Historical Data (1954-2020):

- **CO2 Emissions:** Our World in Data (global CO2 dataset, filtered for Saudi Arabia)
- **Economic Indicators:** World Bank (GDP, population, industrial production)
- **Energy Data:** BP Statistical Review (oil production, energy consumption)
- **Regional Context:** Saudi GASTAT (industrial activity, energy use patterns)
- **Backup Sources:** IEA (International Energy Agency), Saudi Vision 2030 reports for recent energy transition data

Data Processing:

- **Quality Control:** Outlier detection using IQR method, cross-verification across sources
- **Missing Value Imputation:** Time-series interpolation (linear or spline) for gaps, with documentation of imputation impact
- **Feature Engineering:**
 - Lag variables (e.g., previous year's emissions)

- Moving averages (e.g., 3-year energy consumption trends)
- Saudi-specific features: oil production volume, industrial energy intensity, renewable energy adoption
- **Exploratory Data Analysis (EDA):** Correlation analysis and PCA to identify key predictors
- **Standardization:** Train-test splits (80-20) with temporal validation to respect time-series nature

Data Challenges and Mitigation:

- **Inconsistent Formats:** Standardize units (e.g., CO2 in metric tons, energy in MWh) across sources
- **Missing Recent Data:** Use proxy variables (e.g., Vision 2030 energy trends) or contact GASTAT for updates
- **Regional Specificity:** Prioritize features reflecting Saudi Arabia's oil-driven economy and climate policies

Algorithm Selection Framework

Selected Algorithms with Rationale:

1. **XGBoost:** Proven temporal regression performance, handles missing data effectively
2. **LightGBM:** Computational efficiency, good performance on structured data
3. **Random Forest:** Proven success in Saudi studies (Haq 2022: 98%), robust baseline
4. **CatBoost:** Effective categorical data handling for economic indicators

Implementation Process:

Phase 1: Baseline Establishment

- Replicate Nassef et al. (2023) methodology for comparison baseline
- Validate baseline performance (target: $R^2 \geq 90\%$)
- Contact authors for clarification if replication issues arise

Phase 2: Systematic Algorithm Testing

This phase focuses on systematically testing the four selected algorithms (XGBoost, LightGBM, Random Forest, CatBoost) to evaluate their performance on Saudi Arabia's CO2 emissions data. The process includes the following steps:

- **Basic Implementation:** Implement each algorithm using default parameters to establish a performance baseline. This ensures a fair starting point for comparison.
- **Hyperparameter Optimization:** Optimize each algorithm's hyperparameters to maximize performance while respecting computational constraints:
 - Conduct **grid search** with focused parameter ranges to systematically explore key hyperparameters.
 - Perform **random search** for efficient exploration of broader parameter spaces.
 - Apply **early stopping** to enhance computational efficiency and prevent overfitting.
- **Cross-Validation:** Use an expanding window cross-validation approach tailored for time-series data to ensure robust performance evaluation across temporal splits.
- **Performance Evaluation and Documentation:** Measure and document performance using metrics such as R^2 , RMSE, MAE, and MAPE, ensuring comprehensive records for comparison.
- **Sensitivity Analysis:** Assess algorithm robustness by testing performance under varying conditions, such as different data periods or feature sets, to understand model stability.

- **Statistical Comparison:** Conduct statistical tests (e.g., t-tests) to identify significant performance differences, providing a rigorous basis for ranking.

Phase 3: Framework Development

This phase synthesizes the results from Phase 2 to create a comprehensive algorithm selection framework tailored for CO₂ emissions prediction in Saudi Arabia. The framework development includes the following steps:

- **Performance-Based Algorithm Ranking:** Rank the algorithms based on their validated performance metrics to identify the most effective model for Saudi-specific data.
- **Statistical Comparison and Validation:** Perform statistical tests (e.g., ANOVA, t-tests) to validate performance differences and ensure reliable comparisons.
- **Selection Criteria and Guidelines Development:** Develop evidence-based guidelines for selecting algorithms, incorporating Saudi-specific factors such as oil production and energy consumption patterns.
- **Implementation Best Practices Documentation:** Document step-by-step procedures and best practices for implementing the framework, ensuring reproducibility.
- **Sensitivity Analysis Results:** Integrate sensitivity analysis outcomes to demonstrate the framework's robustness under varying data conditions.
- **Practical Usage Recommendations:** Provide actionable recommendations for applying the framework in Saudi Arabia, including considerations for environmental policy support.

Validation and Evaluation

Performance Metrics:

- **Primary:** R², RMSE, MAE, MAPE for prediction accuracy
- **Secondary:** Computational efficiency (training time), implementation complexity
- **Statistical:** Confidence intervals, significance testing (p-values < 0.05)

Validation Approach:

- Time-series cross-validation with expanding windows
- Hold-out test set (last 10% of data) for final validation
- Sensitivity analysis: Test models under varying conditions (e.g., reduced data periods, missing features)
- Statistical significance testing for performance comparisons
- Robustness testing: Evaluate performance with noisy or incomplete data

Implementation Schedule

4-Month Timeline

Month 1: Foundation

- Week 1-2: Environment setup, data collection, EDA, literature completion
- Week 3-4: Data processing, baseline replication, validation framework, local environment setup testing

Month 2: Algorithm Implementation

- Week 5-6: Individual algorithm implementation and basic testing
- Week 7-8: Systematic optimization and parameter tuning for all algorithms

Month 3: Framework Development

- Week 9-10: Systematic comparison, statistical analysis, performance ranking, sensitivity analysis
- Week 11-12: Framework development, guidelines creation, validation

Month 4: Documentation

- Week 13-14: Results analysis, framework finalization, implementation guide
- Week 15-16: Thesis writing, documentation completion, defense preparation, draft conference paper

Project Gantt Chart

Task	Month 1 (W1-W4)	Month 2 (W5-W8)	Month 3 (W9-W12)	Month 4 (W13-W16)
Environment & Data Setup	W1-W2			
EDA & Data Processing	W1-W3			
Baseline Replication	W3-W4			
Algorithm Implementation		W5-W6		
Algorithm Optimization		W7-W8		
Performance Comparison			W9-W10	
Sensitivity Analysis			W9-W10	
Framework Development			W11-W12	
Results Analysis				W13-W14
Thesis & Conference Paper				W15-W16
Milestones				
M1: Data & EDA Complete	W2			
M2: Baseline Established	W4			
M3: Algorithms Optimized		W8		
M4: Framework Complete			W12	
M5: Documentation Ready				W16

Key Milestones

Week	Milestone	Deliverable
Week 2	Data & EDA Complete	Processed datasets, EDA report
Week 4	Baseline Established	Replicated Nassef et al. methodology
Week 8	Algorithms Optimized	All 4 algorithms implemented and tuned
Week 12	Framework Complete	Selection framework with guidelines, sensitivity analysis
Week 16	Documentation Ready	Thesis, implementation guide, draft conference paper

Quality Assurance

- Weekly progress reviews with supervisor
- Statistical validation of all performance comparisons
- Complete code documentation and version control (GitHub)
- Reproducibility testing in Colab and local environments
- Peer review of draft thesis and conference paper

Expected Outcomes and Risk Assessment

Primary Deliverables

Framework Components:

1. **Algorithm Evaluation Methodology:** Practical testing and optimization procedures adapted for academic constraints
2. **Selection Guidelines:** Evidence-based recommendations incorporating Saudi-specific factors
3. **Implementation Guide:** Step-by-step procedures with documented code (Jupyter notebooks, GitHub repository)
4. **Performance Benchmarks:** Validated results with sensitivity analysis for robustness
5. **Conference Paper Draft:** Summary of findings for potential publication

Academic Contribution:

- Master's thesis (80-100 pages) with complete framework documentation
- Systematic algorithm comparison for Saudi emissions prediction
- Replicable methodology for regional environmental modeling studies
- Sensitivity analysis demonstrating robustness under varying conditions
- Potential conference paper for academic dissemination

Risk Assessment

Moderate Risks (30-40% probability):

- **Algorithm Performance Variation:** Some algorithms may not exceed baseline performance
 - *Mitigation:* Focus on systematic comparison value, use sensitivity analysis to explore performance drivers
- **Optimization Resource Constraints:** Limited computational resources for extensive tuning
 - *Mitigation:* Use efficient methods (random search, early stopping), test local environment as backup
- **Data Quality Issues:** Inconsistent formats or missing recent data
 - *Mitigation:* Cross-verify sources, use proxy variables, contact GASTAT for clarification

Low Risks (15-20% probability):

- **Baseline Replication Issues:** Difficulty reproducing exact Nassef et al. results
 - *Mitigation:* Accept 90% baseline performance, document discrepancies clearly

Success Probability

High Confidence (85%): Complete framework development with systematic algorithm comparison and practical documentation

Moderate Confidence (70%): Measurable improvement over baseline performance (1-2% above 93.13%)

Guaranteed Academic Value: Even if performance improvements are modest, the systematic methodology and sensitivity analysis provide substantial value for regional emissions modeling research.

Value of Negative Results: Even if no algorithm significantly outperforms the baseline, the comparative methodology, reproducibility framework, sensitivity analysis, and systematic evaluation provide substantial academic and practical value for future research.

Technical Requirements

Development Environment

- **Primary Platform:** Google Colab (free tier with GPU access)
- **Backup Platform:** Local environment (Python 3.9+, same libraries) for handling Colab session limits
- **Python Libraries:** pandas, numpy, scikit-learn, xgboost, lightgbm, catboost
- **Optimization:** Random search and early stopping for efficient tuning within Colab limits
- **Documentation:** Colab notebooks with markdown and code comments

Data Storage and Management

- **Primary Storage:** Google Drive for datasets and results
- **Backup Strategy:** Weekly GitHub commits, local backups
- **Session Management:** Work in 8-10 hour chunks to accommodate Colab limits
- **Data Organization:** Folders for raw data, processed data, models, results, EDA reports

Version Control and Collaboration

- **Code Repository:** GitHub for version control and collaboration
- **Notebook Management:** Regular .ipynb exports to GitHub
- **Documentation:** LaTeX (Overleaf) for thesis, README for repository
- **Progress Tracking:** Weekly backups, issue tracking on GitHub

Documentation Standards

- **Development Documentation:** Well-commented Colab notebooks with clear markdown explanations
- **Code Repository:** GitHub repository with comprehensive README and environment specifications
- **Academic Documentation:** Methodology documentation suitable for thesis and replication
- **Reproducibility:** Clear instructions for running notebooks in Colab and local environments

Success Factors

- **Platform-adapted methodology** ensuring consistent evaluation within Colab constraints
- **Robust backup strategy** preventing data loss from session timeouts
- **Practical documentation** enabling replication in Colab and local environments
- **Regional focus** ensuring applicability to Saudi context and environmental policy support

Expected Completion: 16 weeks from project start

Primary Risk Mitigation: Multiple algorithms, sensitivity analysis, and systematic approach ensure success

Academic Value: Framework methodology valuable regardless of performance improvements

Practical Impact: Immediately usable for future regional emissions prediction studies and environmental policy support