Data Science Capstone Project Plan

Duration: October 28 – November 21, 2025
Team Size: 3–4 students
Total Time: 2 weeks

Key Dates

- November 5 (Stand-up 1): Share progress, challenges, and questions with coach
- November 12 (Stand-up 2): Discuss model results, evaluation strategy, and visuals
- November 21 (Final Presentation):
    - 13–15 min project presentation
    - 5–7 min Q&A
    - Open to all attendees

Suggested Project Topics

Education-Focused Projects

Theme	**Project Idea**	**Dataset Suggestion**	**Techniques Covered**	**Expected Output**
Learning Analytics	Student Performance Prediction	UCI Student Performance Dataset, Kaggle Exam Scores	Decision Trees, KNN, Neural Network	Academic success predictor dashboard

Student Engagement	Classroom Engagement Recognition	DIPSER Dataset (2025)	Ensemble Models, Neural Network	Engagement classification dashboard
Online Learning	MOOC Dropout Prediction	EdNet (2020), MOOCCube (2020)	Random Forest, XGBoost	Dropout likelihood predictor
Teacher Analytics	Predicting Instructor Workload	Synthetic LMS Activity Data	Decision Tree, Regression KNN	Workload forecasting dashboard
Cognitive Load	Predict Cognitive Load from Student Data	Synthetic Educational Datasets (Python generator)	Neural Network, KNN	Visual load analysis
Computational Thinking	Analyze Coding Patterns in Programming Courses	IT Academy Inquiry Skills Dataset	Decision Tree, Ensemble	Behavioural metrics dashboard

Other Domains

Theme	**Project Idea**	**Dataset Suggestion**	**Techniques Covered**	**Possible Outputs**
Healthcare	Heart Disease Prediction	UCI Heart Disease Dataset, Framingham Study	RF, AdaBoost, Neural Network	Feature importance and comparison of ensemble vs NN
Business	Customer Churn Prediction	Kaggle “Telco Customer Churn”	Decision Trees, Boosting, ANN	Churn analytics and model comparison
Environment	Air Quality Index Forecasting	UCI Air Quality Dataset	Neural Networks, Bagging, KNN	AQI forecast charts
Finance	Credit Risk Prediction	German Credit Data (Kaggle)	XGBoost, Random Forest	ROC curve summary
Transportation	Taxi Fare Prediction	NYC Taxi Dataset	KNN, Decision Tree Regressor	Fare estimation and residual visualization
Social Media	Twitter Sentiment Analysis	Kaggle Twitter Sentiment Dataset	Simple NN, Voting Classifier	Sentiment visualization dashboard
Agriculture	Crop Yield or Disease Prediction	UCI Crop Dataset	Decision Tree, Neural Network	Predictive dashboard for yield

Weekly Work Plan

Week 1 (Oct 28 – Nov 7)

- Dataset exploration and EDA
- Identify research question, target variable, baseline metrics
- Begin preprocessing (missing data, outliers, encoding)
- Stand-up 1 (Nov 5):
  - Present progress on dataset and EDA
  - Ask technical and modeling questions

Week 2 (Nov 8 – Nov 20)

- Train baseline models (Decision Tree, KNN, Neural Network, Ensemble)
- Tune hyperparameters and compare models
- Generate interpretability plots (feature importance, SHAP)
- Prepare report slides
- Stand-up 2 (Nov 12):
  - Share validation results
  - Discuss challenges and next steps

November 21 – Final Presentation

- 13–15 min project presentation + 5–7 min Q&A
- All students and faculty invited

Final Deliverables

1. Short Written Report (max 8 pages): problem statement, dataset, methods, evaluation, and future work
2. Presentation Slides: clear storyline, visuals, and insights
3. Code Repository: clean, well-commented notebook (with reproducibility)

Capstone Rubric (Total 60 Marks)

Assessment Area	**Criteria**	**Description**	**Weight**
1. Problem Definition & Motivation	Relevance and clarity	Clear research problem, background, and goal setting	5 pts
2. Dataset Quality & Preparation	Data Source & EDA	Appropriate dataset selection, preprocessing, visualization clarity	5 pts
3. Modeling Approach	Algorithm Implementation and Justification	Appropriate choice of algorithms (Decision Tree, Ensemble, KNN, NN) and rationale	15 pts
4. Evaluation & Interpretation	Metrics and Explainability	Correct evaluation metrics (accuracy, F1, RMSE, AUC, etc.); interpretation using visuals	15 pts
5. Insight & Application	Value of Findings	Quality and depth of domain insights; reproducible implications	10 pts

6. Presentation Quality	Storyline & Communication, following the timelimit	Coherent narrative, visuals, time management, professional Q&A	5 pts
7. Team Collaboration	Coordination & Participation	Equal task sharing, engagement from all the team members in both stand-ups and presentation+Q&A	5 pts

Total: 60 points

Notes for Presentations

- Keep slides concise (max 20-25 slides)
- Begin with “why the problem matters” and finish with “what the results mean”
- Use color-coded charts for model comparison
  (AUC, Accuracy, Confusion Matrices)
- Avoid technical overload – focus on explanation quality