🎯 Phase 1: Problem Definition & Planning

This document serves as the foundational agreement for the **Student Dropout Risk Prediction** machine learning project. It clearly defines the business objective, the technical ML approach, and the criteria for success.

1. Business Understanding

The core objective of this project is to create an early warning system that allows academic advisors and student support staff to proactively engage with students at high risk of dropping out.

Element	Description
Primary Business Goal	**Reduce the student dropout rate by 20%** within the next academic year.
Rationale	Dropout leads to significant revenue loss, administrative overhead, and reduced institutional reputation. Early, targeted intervention is the most cost-effective solution.
Intervention Trigger	A student flagged with a high risk score (above a dynamic threshold, TBD during evaluation) will trigger a personalized outreach campaign.

2. Machine Learning Problem Definition

The problem is framed as a supervised learning task designed to predict a binary outcome based on a student's profile and historical engagement data.

Element	Specification
Problem Type	**Binary Classification**

Target Variable (Y)	`Dropout` (1) or `Continue` (0).
Input Features (X)	A mixture of categorical and continuous variables including Demographics (Age, Income), Academics (Grades, Attendance), and Engagement metrics.
Key Output	A **Probability Score** (ranging from 0.0 to 1.0) representing the likelihood of dropout, which will be converted to a binary risk flag.
Model Choice (Initial)	LightGBM Classifier (selected for its speed, handling of diverse feature types, and built-in feature importance for interpretability).

3. Success Metrics and Thresholds

Given the need to identify the *minority* (at-risk students) and the high cost of a false negative (failing to identify an at-risk student), we prioritize metrics that handle class imbalance effectively.

Metric	Rationale	Target Threshold
Primary Metric: ROC-AUC	Measures the model's ability to discriminate between positive (dropout) and negative (continue) classes across all thresholds.	**> 0.85**
Secondary Metric: F1-Score	Represents the harmonic mean of precision and recall. Essential for balancing false positives (unnecessary intervention) and false negatives (missed risk).	**Optimized based on business need** (e.g., maximizing F1 for a specific recall level).
Cost Matrix Consideration	The cost of a **False Negative** (missing an at-risk student) is considered significantly	The threshold will be tuned to achieve high Recall, subject to

	higher than a **False Positive** (intervening unnecessarily).	acceptable Precision limits.

4. Constraints and Requirements

Constraint Category	Detail	Tracking & Documentation
Interpretability	**High Requirement:** Predictions must be traceable back to specific features (e.g., low attendance, poor grade in X course) to enable meaningful intervention strategies.	Documented in `docs/constraints.md`. Will use SHAP/LIME for explanation.
Data Privacy	All personally identifiable information (PII) must be anonymized or pseudonymized during preprocessing and storage. Adherence to institutional data governance policies is mandatory.	Legal review documented in a project log.
Model Latency	The prediction API must provide a response within **100ms** to support real-time application usage (e.g., student support dashboard lookups).	Monitored post-deployment via **AWS CloudWatch**.
Data Availability	Initial model training relies on data derived from the UCI Student Performance dataset. Any shift to internal institutional data will require a new data collection audit.	Data source and schema tracked in `data/schema.json`.

Tracking Tools Summary for Phase 1:

- **Project Goals:** Confluence/Notion (Business tracking).
- **ML Specs:** GitHub Wiki (Technical tracking).
- **Configuration:** `configs/training_config.yaml` (Stores target metrics and initial hyperparameter ranges).

- **Documentation:** `docs/constraints.md` (Formalizes non-functional requirements).