**1. Introduction (2 pages)**

Introduce the student dropout problem in higher education and motivate the use of machine learning for early prediction.

**Content Suggestions:**

* **Context and Motivation:**
  + Prevalence and cost of student dropout
  + Importance of early detection and intervention
* **Problem Statement:**
  + Binary classification: dropout vs. continuation
  + Focus on early and midpoint prediction
* **Data Sources:**
  + Demographics, VLE activity, assessments, etc.
* **Aims and Objectives:**
  + Build predictive models using ML
  + Prioritise recall on dropout class
  + Understand key predictive features
* **Structure Overview:**
  + Briefly explain what each section will cover

**Section 2: Background (5 pages)**

Provide context on student dropout research and relevant machine learning literature.

**Suggested Subsections:**

1. **Student Dropout in Higher Education**
   * Statistics on dropout rates globally or in the UK
   * Reasons for dropout: academic, personal, engagement-related
2. **Predictive Analytics in Education**
   * Role of data-driven decision-making in student retention
   * Use cases of ML in academic settings
3. **Related Work**
   * Review studies using ML for dropout prediction
   * Strengths/weaknesses of prior models
4. **Rationale for Model Selection**
   * Why LR, SVM, MLP, and RF are suitable:
     + LR: Interpretable baseline
     + SVM: Handles high-dimensional spaces well
     + MLP: Captures nonlinear relationships
     + RF: Robust, handles mixed data types, feature importance
5. **Problem Statement and Objectives**
   * Define the classification task
   * Emphasise the need for early and accurate dropout prediction

**Section 3: Methods (10 pages)**

**Suggested Subsections:**

1. **Data Description**
   * Source of data, number of students, number of features
   * Overview of labels (dropout/continue)
2. **Feature Engineering**
   * Demographic, academic, and VLE features
   * Construction of early/mid/late/full datasets
3. **Preprocessing**
   * Missing value handling
   * Outlier detection and handling
   * Encoding and scaling
4. **Phase Division Logic**
   * How early, mid, late phases were defined based on time window
   * Justification of cutoffs (e.g., 0–90, 91–180, etc.)
5. **Model Selection and Justification**
   * Brief algorithm overview
   * Why each model is expected to help identify dropout patterns
6. **Training and Evaluation Protocol**
   * Train-test split
   * GridSearchCV
   * 5-fold cross-validation
   * Evaluation metrics: Accuracy, Macro F1, Precision, Recall (especially dropout class recall)
7. **Implementation**
   * Libraries and tools used (e.g., Scikit-learn, Pandas, NumPy)
   * Code structure (optional diagram or pseudocode)

**Section 4: Results and Analysis (10 pages)**

**Suggested Subsections:**

1. **Model Performance Overview**
   * Table summarising accuracy and macro F1 for each model across all dataset phases
2. **Detailed Model Comparisons**
   * **Logistic Regression**
     + Strength: Interpretable coefficients
     + Weakness: May underperform on nonlinear relationships
   * **SVM**
     + Performs well with a balanced margin
     + Limitations with scale
   * **MLP**
     + Stronger at capturing complex patterns
     + Sensitive to parameter tuning and training time
   * **Random Forest**
     + Best dropout recall
     + Visualisation of feature importances
3. **Dropout Class Analysis**
   * Precision and recall for dropout class (in-depth)
   * Compare false negatives and false positives for each model
   * Which model best balances sensitivity and specificity
4. **Impact of Time Phase**
   * How prediction accuracy improves from early → midpoint → full
   * Tradeoff between timeliness and confidence
   * Why early and midpoint are more practical for intervention
5. **Feature Importance Analysis**
   * For RF: Feature importance plot
   * Discussion of which features matter most: VLE activity, assessment scores, age band, etc.
6. **Error Analysis**
   * Where models struggle: borderline students, low activity students
   * Suggestions on improving prediction: more behavioural features, fine-tuned temporal segmentation

**Section 5: Conclusion (3 pages)**

**Suggested Content:**

1. **Summary of Findings**
   * Best performing models
   * Importance of early identification
   * Key impactful features
2. **Contributions**
   * Built a full ML pipeline for dropout prediction
   * Compared 4 major classifiers with time-based datasets
   * Showed practical trade-offs in early vs. late prediction
3. **Limitations**
   * Data granularity, generalisability, class imbalance, lack of external factors (e.g., mental health, finances)
4. **Future Work**
   * Add more real-time data (e.g., forum posts, login frequency)
   * Test deep learning architectures (e.g., LSTM for sequences)
   * Deploy as part of a student dashboard for support staff