🎓 Final Project Report: Student Engagement Prediction

Course: AI for Software Engineering

Author: Bikila Keneni

🧩 1. Problem Definition

Universities face challenges in identifying disengaged students early enough to intervene effectively.
By leveraging behavioral and academic data, an AI model can help predict student engagement levels, enabling instructors and advisors to take timely action.

Objectives

1. Predict student engagement levels (low vs. high).
2. Identify the most influential factors driving engagement.
3. Support instructors in improving course delivery and student participation.

Stakeholders

- Academic Affairs and Course Instructors
- Students

Key Performance Indicator (KPI): *Recall* — percentage of disengaged students correctly identified.

📊 2. Dataset Overview

Source: Synthetic university dataset generated from Learning Management System (LMS) logs and feedback data.

Total Students: 300
Engagement Levels Distribution:

- Low Engagement (0): 271 students (90.3%)
- High Engagement (1): 29 students (9.7%)

Key Features:

- `attendance_rate`
- `average_study_time`
- `assignment_completion_rate`
- `discussion_participation`
- `instructor_rating`

Observation: The dataset is **imbalanced**, with far more disengaged students than engaged ones.

---

3. Data Preprocessing

Steps performed:

1. Removed duplicates and null values.
2. Normalized numerical features (e.g., study time, activity scores).
3. Encoded categorical variables (e.g., course session times).
4. Scaled data using `StandardScaler`.

Output Files:

- `cleaned_engagement.csv` — processed dataset
- `scaler.pkl` — saved feature scaler

---

4. Model Development

Model Used: Random Forest Classifier
Reasoning: Robustness, interpretability, and strong performance on tabular data.

Data Split:

- Training: 70%
- Validation: 15%
- Testing: 15%

Hyperparameters Tuned:

- `n_estimators`: 100 → 200
- `max_depth`: None → 10
- `min_samples_split`: 2 → 5

Saved Model: `engagement_model.pkl`

---

5. Evaluation Results

Confusion Matrix (Full Dataset)

	Predicted Low	**Predicted High**
Actual Low	268	3
Actual High	18	11

Classification Report

Metric	**Low (0)**	**High (1)**	**Weighted Avg**
Precision	0.94	0.79	0.92
Recall	0.99	0.38	0.93

Metric	Low (0)	High (1)	Weighted Avg
F1-Score	0.96	0.51	0.92
Accuracy	**0.93**		

Interpretation:
The model performs very well on predicting disengaged students (Recall = 0.99) but struggles with high-engagement cases due to dataset imbalance.

---

6. Insights & Discussion

- **Engagement is heavily skewed** toward low values, suggesting a need for more motivation or better learning design.
- **Instructor rating** and **study time** were among the strongest predictors.
- **Recall for high-engagement students (0.38)** could be improved using resampling or additional features.

---

7. Deployment Plan

1. Package the trained model into a Flask API.
2. Integrate predictions into the university LMS dashboard.
3. Create visualization panels to display real-time engagement insights.
4. Automate model retraining every semester to handle **concept drift**.

Compliance: Ensure student data is anonymized and follows institutional privacy guidelines (e.g., FERPA).

---

8. Ethical Considerations

- **Bias Risk:** Underrepresentation of certain programs may skew predictions.
- **Mitigation:** Apply oversampling (SMOTE) and regular fairness audits.
- **Transparency:** Ensure model explanations are accessible to educators.

---

9. Reflection

Challenges:

- Handling imbalanced data and defining engagement quantitatively.

Future Improvements:

- Include Natural Language Processing (NLP) on student feedback.
- Use explainable AI (e.g., SHAP) for transparent feature interpretation.
- Develop an interactive analytics dashboard.

Final Summary

Category	Description
Model	Random Forest Classifier
Accuracy	93%
Recall (Low Engagement)	0.99
Recall (High Engagement)	0.38
Dataset Size	300 students
Main Challenge	Class imbalance
Outcome	Successful AI workflow pipeline from data preprocessing to evaluation

End of Report

File generated as part of the AI Workflow Assignment.