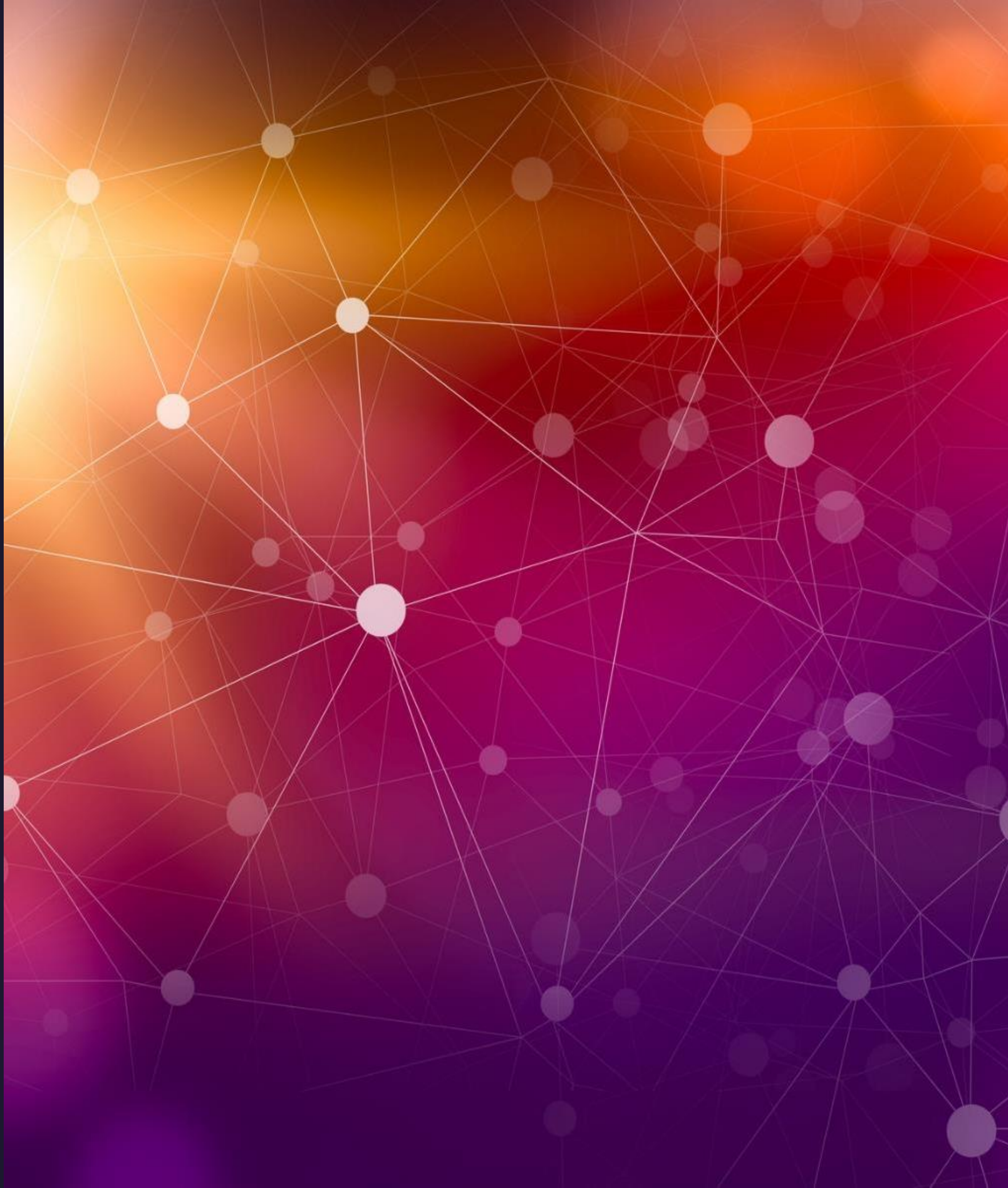
Student Performance Prediction & Support System



By student :
Ahmed Diefallah

Agenda

Dataset

Challenges

Data preprocessing

Machin learning model

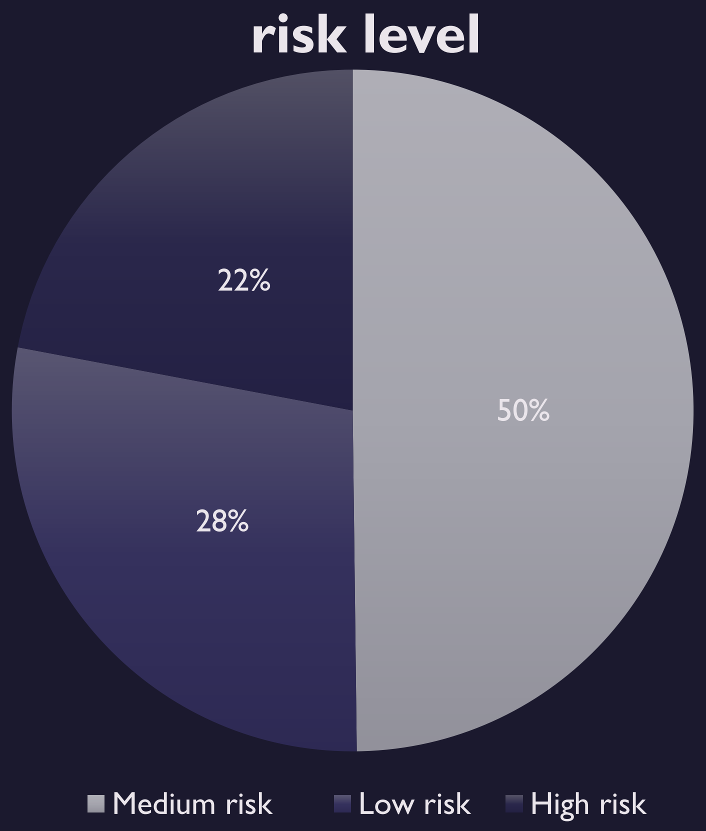
Key Insights

Conclusion

Dataset

Source: Student performance dataset (Math & Portuguese).
Size: ~650 records for Portuguese and ~350 records for Math .
Features (33 features) :

- Demographics: (age, gender, address, family background).
- Academic: (study time, failures, absences, grades).
- Social: (activities, internet, romantic relationship).
- Target variable: Student performance category (Low / Medium / High Risk).



Data Quality

No Missing Values.

Balanced Classes: Well-distributed risk categories.

Rich Features: 33 comprehensive attributes

Challenges

No unique student IDs → Difficult to merge Math and Portuguese datasets without duplication Class

To solve this make new column and combine two datasets

Feature selection → Need to identify which features (e.g., grades, study time, absences) are most relevant.

Categorical variables → Many features are non-numeric (e.g., school, address, family status), requiring encoding.

Apply binary and one-hot encoder to make it numeric

Overfitting risk → Simple models (Decision Tree) may perform well on training but poorly on unseen data.

Tuning model to handle and not get overfitting

Data preprocessing

Data Cleaning

Removed irrelevant features (e.g., guardian, nursery, romantic).

Feature Engineering

Created risk_level column from Final Grade (G3):High Risk: G3 < 10Medium Risk: 10 ≤ G3 ≤ 13Low Risk: G3 ≥ 14

Computed Average Grade = (G1 + G2 + G3)/3.

And attendance ratio column by 1 - absence/100.

Encoding Categorical Features Converted binary (e.g., yes/no → 0/1).One-Hot Encoding for nominal features (e.g., job, reason).

Normalization & Scaling Standardized numerical features (age, absences, grades) using MinMaxScale. Ensured fair contribution across features.

Machin learning model

- Logistic Regression 📈

  Baseline linear model.

  Simple and interpretable.

- Decision Tree 🌳

  Splits data based on feature conditions.

  Easy to visualize but can overfit.

- Random Forest 🌲🌲

  Ensemble of decision trees.

  Reduces overfitting, improves accuracy.

- Support Vector Machine (SVM) ⚡

  Finds the best boundary between classes.

  Works well with high-dimensional data.

- Neural Network (TensorFlow/Keras) 🧠

  Multi-layer perceptron for classification.

  Handles complex non-linear patterns.

- NLP Sentiment Analysis 💬

  TF-IDF + Logistic Regression.

  Used on student feedback text to detect

sentiment → correlated with performance.

Result

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	**93.3%**	**93.3%**	**93.3%**	**93.2%**
Random Forest	88.5%	88.6%	88.5%	88.5%
Logistic Regression	84.2%	85.6%	84.2%	84.0%
SVM	83.7%	85.3%	83.7%	83.5%
Neural Network (TF)	~84%	~86%	~84%	~84%
NLP Sentiment Model	93.5%	92.4%	93.5%	92.5%

Decision Tree highest accuracy

- 93.30% accuracy - highest performance
- Balanced across all metrics
- Clear interpretability for educators
- Handles feature interactions well

Key Insights

- Grades matter most 📘

G1, G2, and G3 (period grades) were the strongest predictors of student risk level.

- Decision Tree vs. Random Forest 🌳

Decision Tree had the highest accuracy but risk of overfitting.

Random Forest provided more balanced predictions and robustness.

- Neural Network 🧠

Achieved comparable accuracy to Random Forest.

Flexible and scalable for future, larger datasets.

- NLP Sentiment Analysis 💬

Student feedback sentiment correlated with performance risk.

Negative sentiment often linked to High Risk.

- Feature Relevance 🔑

Academic effort (study time, failures, absences) had stronger influence than social factors (activities, romantic, internet).

Ethics in AI

- Data Privacy
  1. No personal identifiers (names, IDs) stored
  2. Anonymized student records

- Bias Awareness
  1. Risk of bias from gender, socioeconomic status, school type
  2. Need fair training data and monitoring

- Responsible Use
  1. Predictions should support, not replace, teacher judgment
  2. System used as a decision aid, not a final verdict

- Transparency & Accountability
  1. Models and decisions should be explainable
  2. Schools should remain accountable for actions taken

System Overview

- Purpose: Predict student academic risk using ML + NLP
- Architecture:
    1. Presentation Layer → Flask web app (UI, visualization)
    2. Business Logic Layer → ML models & NLP pipeline
    3. Data Layer → Pre-trained models & datasets
- Models: Logistic Regression, Decision Tree, Random Forest, SVM, Neural Network
- NLP: Text cleaning, TF-IDF, Logistic Regression, sentiment & keyword analysis
- Risk Levels: ⚠️ High (<10) → urgent intervention 🔶 Medium (10–13) → needs support ✅ Low (14+) → safe
- Features: Confidence scoring, color-coded results, REST API support

Conclusion

✅ Built a Student Risk Prediction System combining ML & NLP

🎯 Accurately classifies students into High / Medium / Low risk

📊 Provides actionable insights for early intervention

🌐 Simple web interface + API for real-world usability

⚖️ Considered ethics & bias (gender, socioeconomic factors)

🔮 Future work:

- Improve NLP models with larger datasets
- Continuous retraining with new student data
- Integration with Learning Management Systems (LMS)

Thank you