

STUDENT DROPOUT PREDICTION: A DATA- DRIVEN APPROACH TO IDENTIFYING AT-RISK HIGH SCHOOL STUDENTS



Team Members



Phork Soksan



Pen Sithol



Mat Ehak



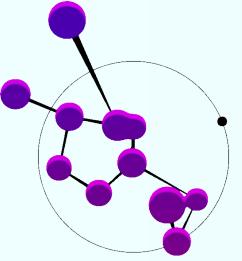
Leang Honghav



Lorn Thornpunleu



OUR CONTENT



Research Background

Problem Statement

Research Objective

Methodology

Results

Findings & Feature
Importance

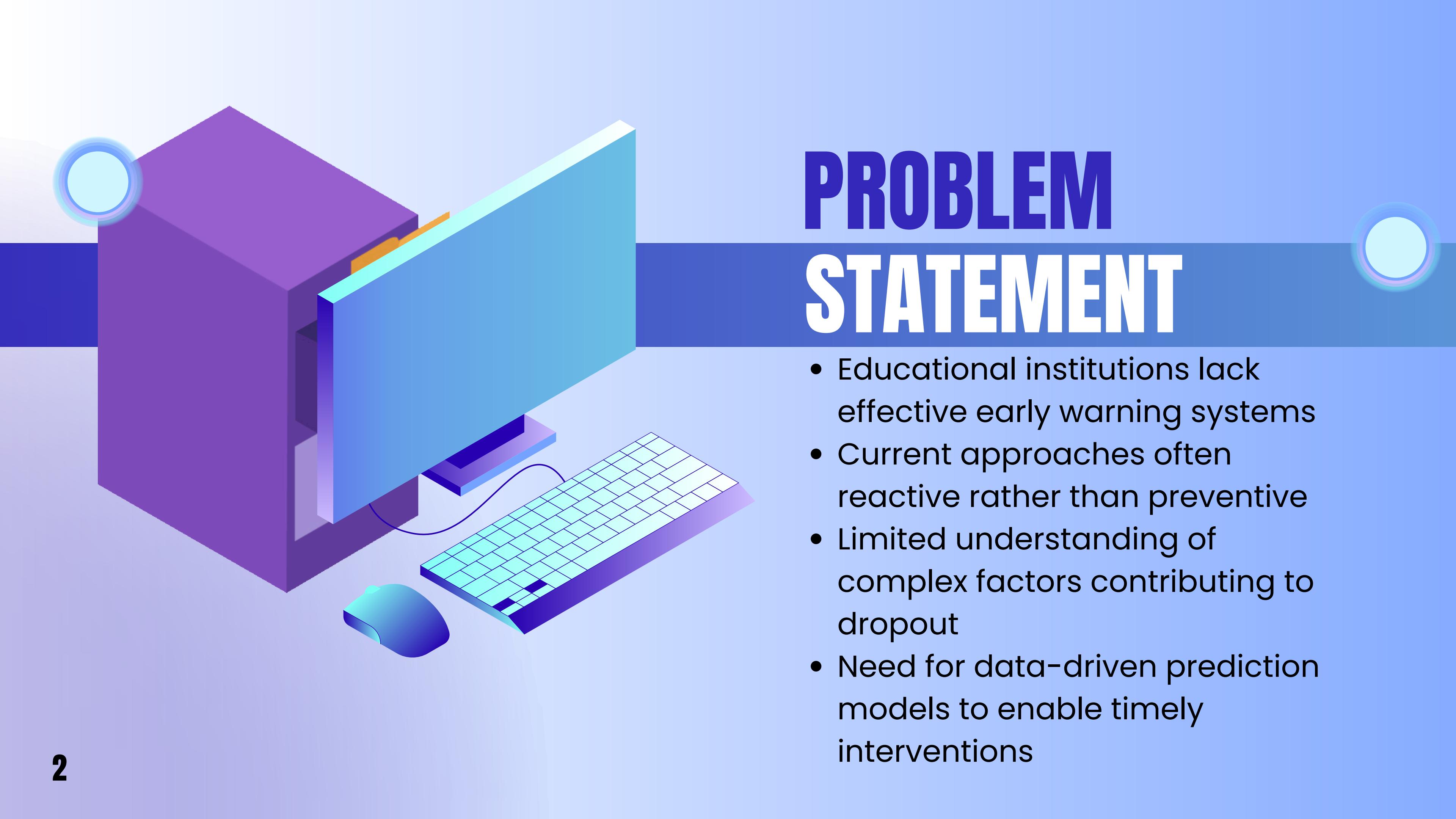
Recommendations

Conclusion



RESEARCH BACKGROUND

- High school dropout remains a critical educational issue.
- Approximately 5.1% of high school students fail to complete education.
- Significant consequences:
 - Lower lifetime earnings
 - Higher unemployment rates
 - Increased incarceration likelihood
 - Substantial economic cost to society

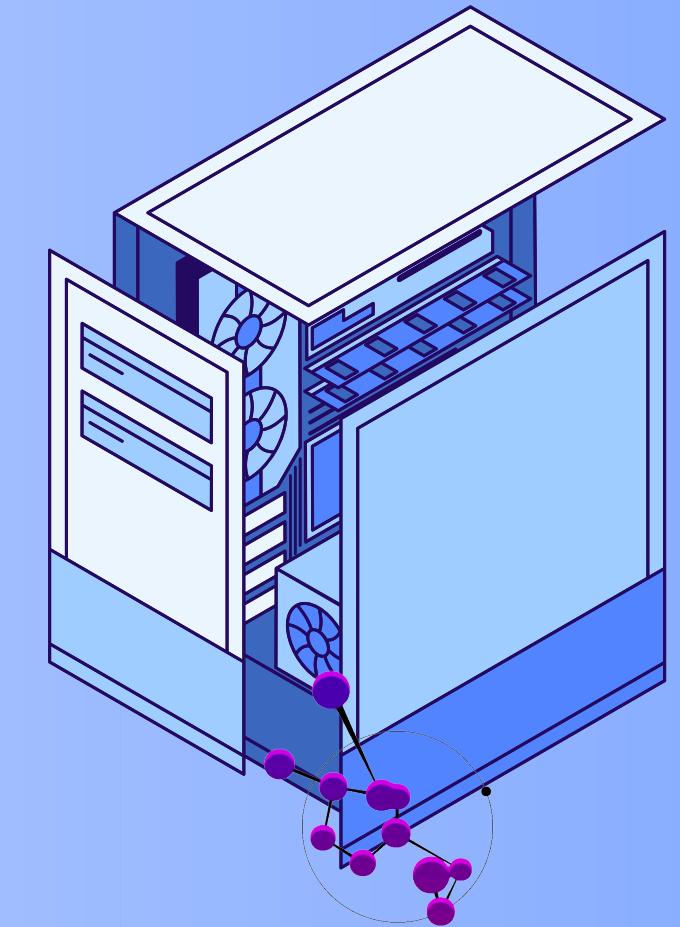
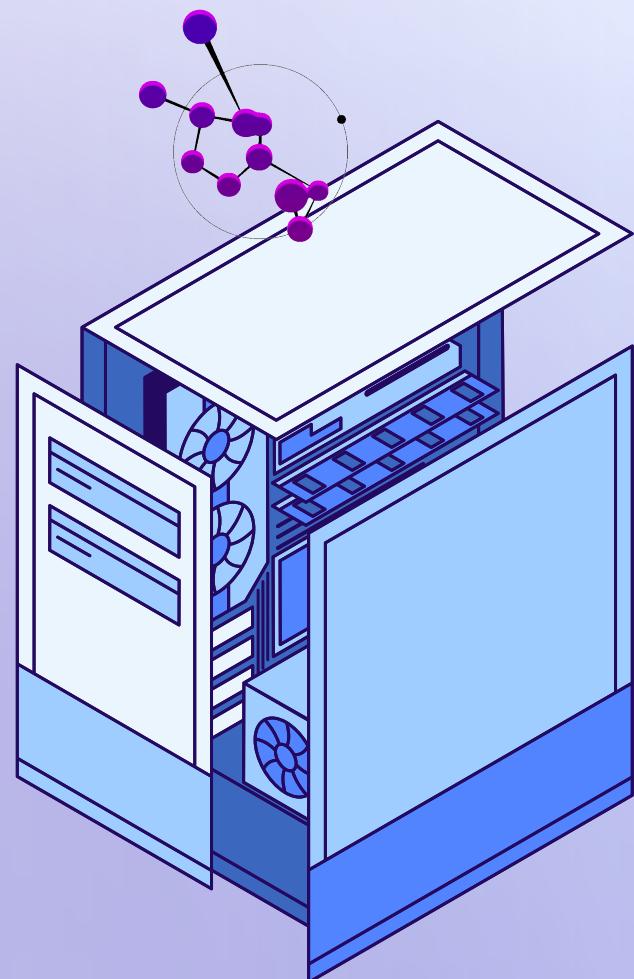


PROBLEM STATEMENT

- Educational institutions lack effective early warning systems
- Current approaches often reactive rather than preventive
- Limited understanding of complex factors contributing to dropout
- Need for data-driven prediction models to enable timely interventions

RESEARCH OBJECTIVE

- 🎯 Develop machine learning models to predict dropout risk.
- 🎯 Identify key factors associated with dropout
- 🎯 Provide recommendations for early warning systems.
- 🎯 Contribute to understanding of modern dropout patterns.



DATASET OVERVIEW

- 5,400 US high school student records
- Multiple states represented (CA, NV, UT, OR, AZ)
- Features include:
 - Demographics (age, gender, location)
 - Academic performance (GPA, test scores)
 - Family background (parent education, occupation)
 - Behavioral indicators (attendance, suspensions)
 - Support systems (counseling, teacher support)

METHODOLOGY - DATA PREPROCESSING

- Removed irrelevant columns (Name, Address)
- Handled missing values using median imputation
- Standardized categorical variables
- Created binary dropout target variable based on:
 - GPA < 2.0
 - Attendance rate < 60%
 - Expulsions > 0
- Encoded categorical features

METHODOLOGY - MODELS

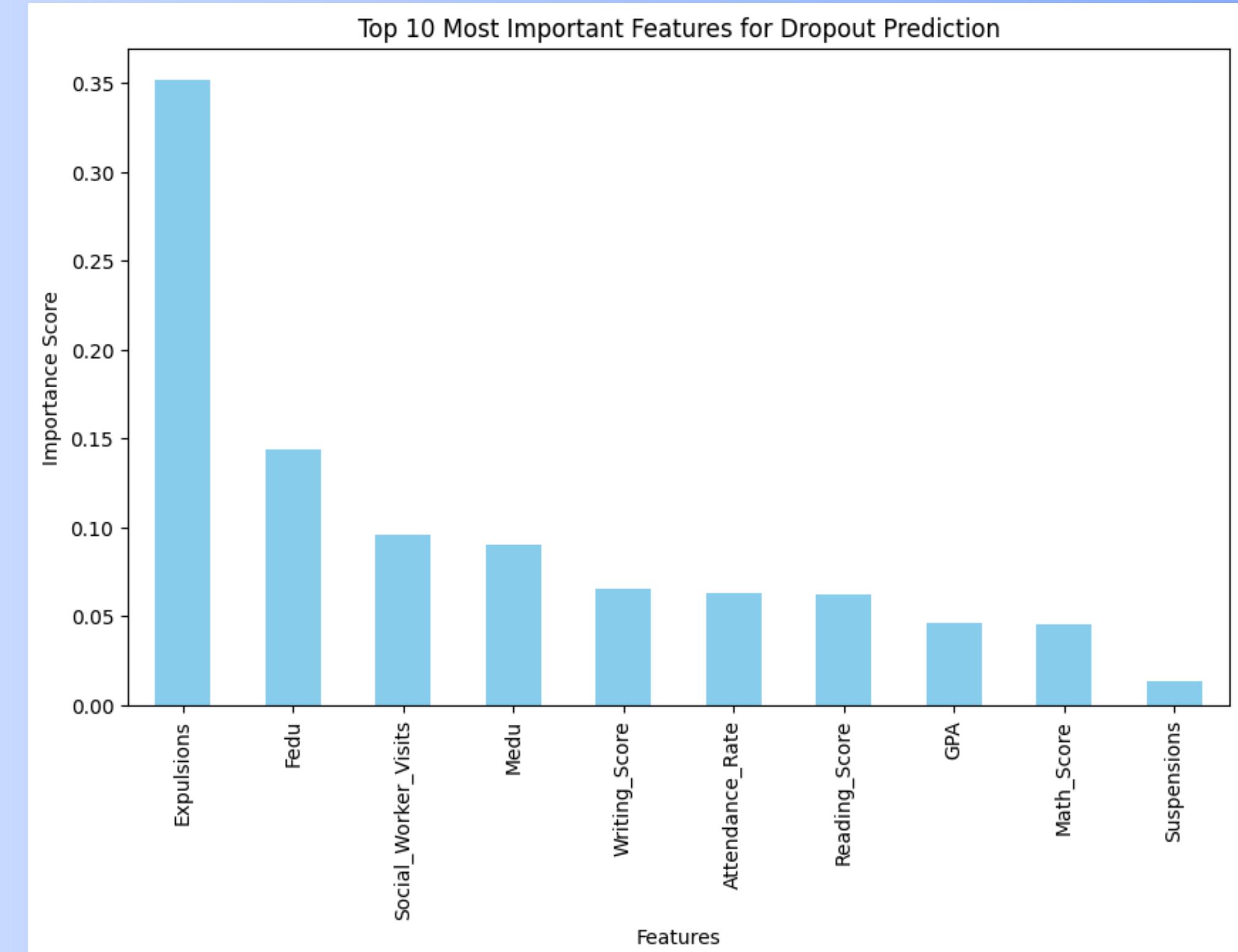
- Random Forest
 - 100 estimators
 - 80/20 train-test split
 - Evaluated with accuracy, precision, recall
- XGBoost
 - Feature selection based on importance
 - Learning rate: 0.1, Max depth: 5
 - Particularly effective for imbalanced data

EVALUATE RESULTS MODEL PERFORMANCE

- Random Forest:
 - Overall accuracy: ~94%
 - Dropout class: Precision 0.85,
Recall 0.65
 - F1-score (dropout): 0.73
- XGBoost:
 - Comparable accuracy to
Random Forest
 - Improved recall for dropout
class (0.68)
 - F1-score (dropout): 0.75

FEATURE IMPORTANCE

- Top Predictors:
 - Attendance Rate
 - GPA
 - Parental Education
 - Test Scores (Math, Reading, Writing)
 - Parental Involvement

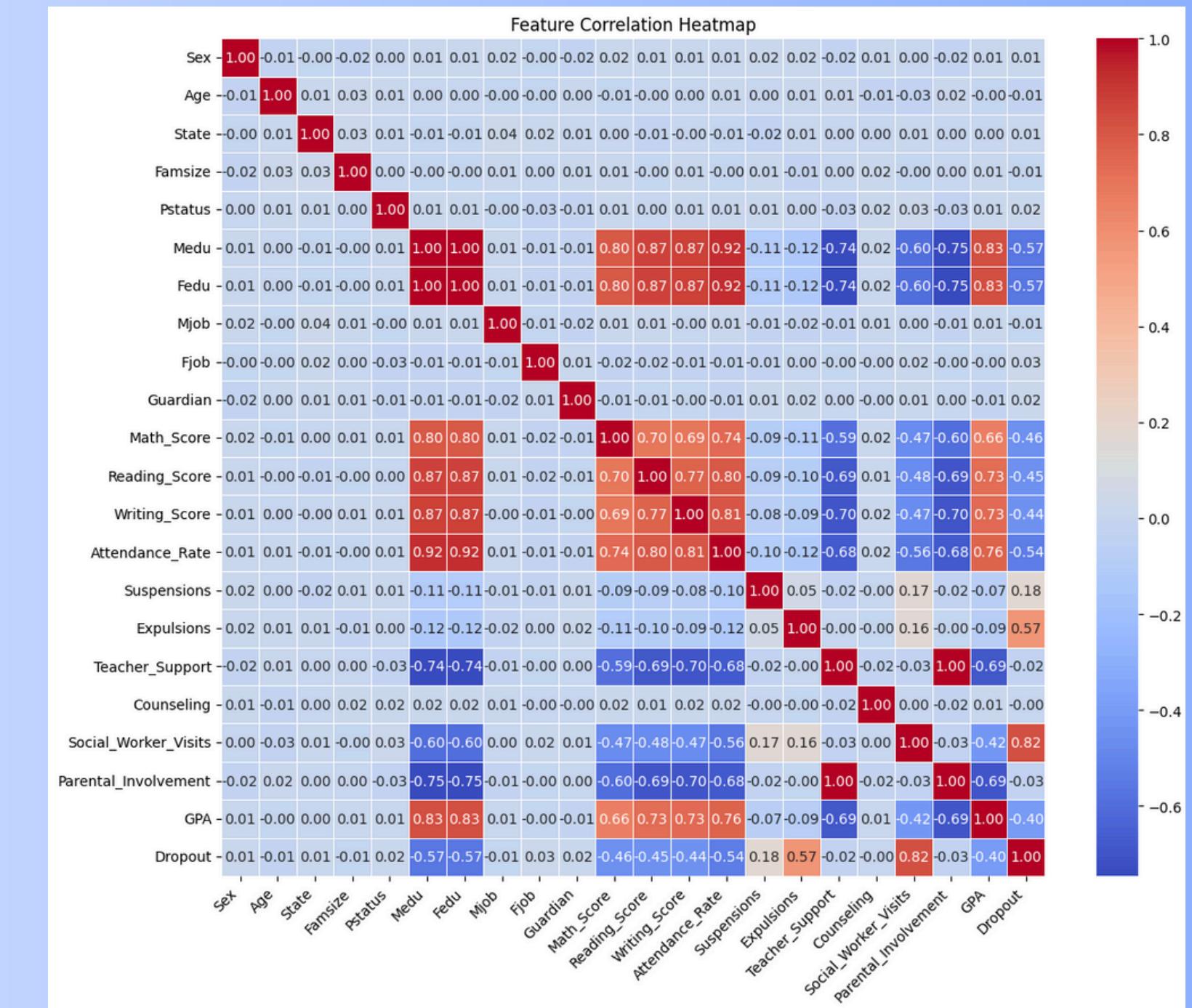


KEY FINDINGS

- Attendance is the strongest predictor of dropout risk
- Academic performance (GPA, test scores) highly influential
- Family background factors (parental education) significantly impact risk
- Parental involvement plays a crucial role
- Demographic factors (gender, location) less important than expected

CORRELATION ANALYSIS

- Strong negative correlation between:
 - Attendance and dropout risk (-0.538)
 - Parental education and dropout risk (-0.567)
 - Academic scores and dropout risk (-0.45)



RECOMMENDATIONS FOR SCHOOLS

- Implement automated early warning systems tracking attendance
- Develop targeted intervention strategies for academic support
- Create specific programs for first-generation students
- Strengthen family engagement initiatives
- Use data-driven approaches to evaluate intervention effectiveness

RECOMMENDATIONS FOR POLICY

- Allocate resources based on evidence-driven risk factors
- Support educator professional development
- Develop policies to increase parental involvement
- Fund targeted research on effective interventions
- Focus on holistic approaches addressing multiple risk factors

FUTURE RESEARCH DIRECTIONS

- Investigate effectiveness of targeted interventions
- Explore interactions between different risk factors
- Conduct longitudinal studies tracking causal relationships
- Include additional data on social-emotional factors
- Develop more sophisticated predictive models

THANK YOU!

Thank you for your attention!
Questions?