Program : Applied AI Solutions Development

Course : Applied Math. Concepts for Machine Learning

Term Project

Student Performance Classification

Instructor : Dr. Reza Moslemi

Group 8 :

- Ashwin
- Gil
- Saranya
- Lokesh
- Waliuddin
- Yesha
- Sergio
- Akash
- Abirudh
- Agha

Introduction

- **Data Source :** https://www.kaggle.com/datasets/lainguyn123/student-performance-factors
- **Objective :** The purpose of this project is to build and evaluate classification models on a dataset, to classify students into **above-average** (score >= 67) or **below-average** (score < 67) performance based on exam scores. We will use various classification algorithms and perform GridSearchCV to optimize their performance.

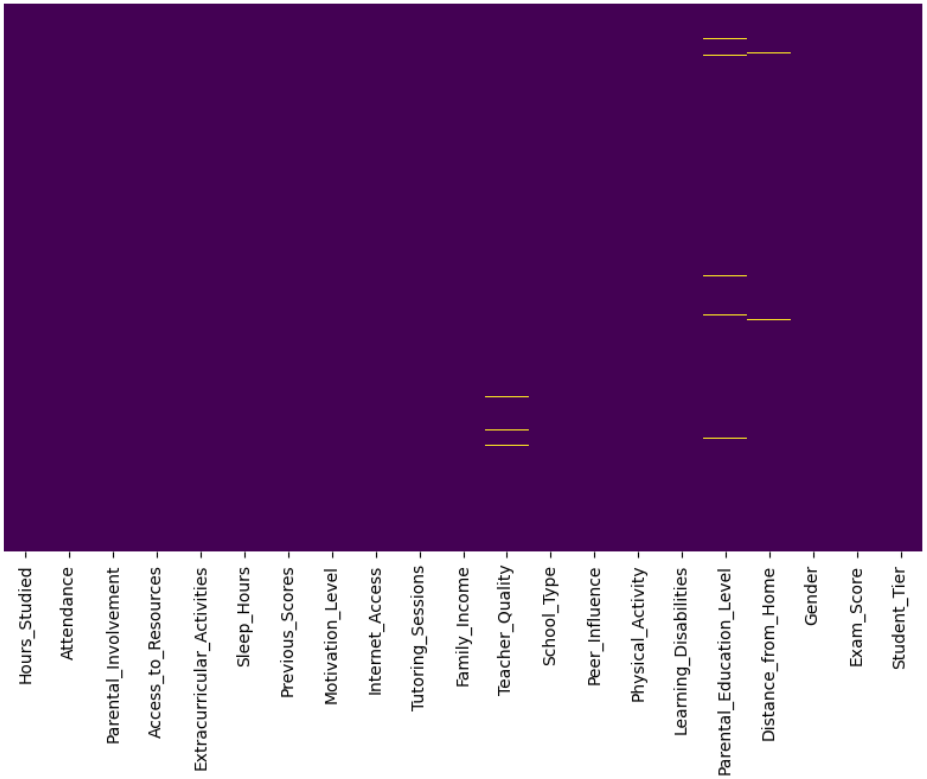
Characteristics	Name	Quantity	Values
Behavioral	Hours Studied, Sleep Hours, Physical Activity, Extracurricular Activities, Internet Access	5	Nominal, Continuous
Health Factor	Learning Disabilities	1	Nominal
Parental & Peer Influence	Parental Involvement Level, Peer Influence	2	Ordinal
Geographical	Distance from Home	1	Ordinal
Education Resources	Access to Resources, School Type, Teacher Quality, Tutoring Sessions	4	Continuous Ordinal, Nominal
Academic Performance	Attendance, Previous Scores	2	Continuous
Motivation	Motivation Level	1	Ordinal
Demographic	Gender, Family Income, Parental Education Level	3	Nominal, Ordinal

Exam Score (Target)
Above-Average (Score >= 67)
Below-Average (Score < 67)

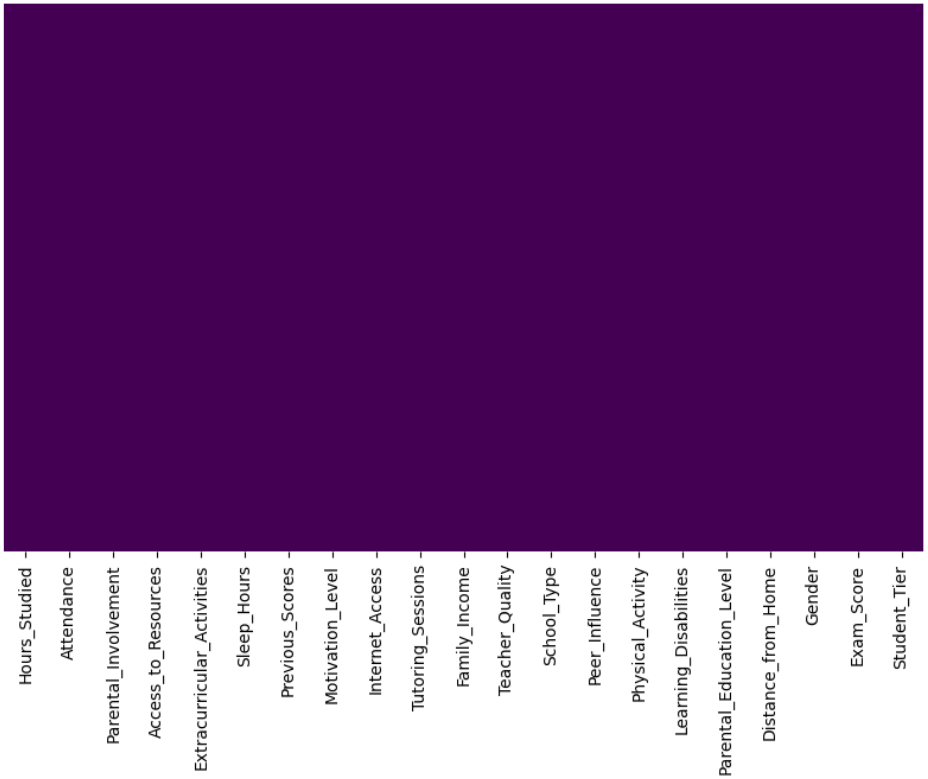
Shape
6607 Samples
20 columns

Data Cleaning & Preprocessing

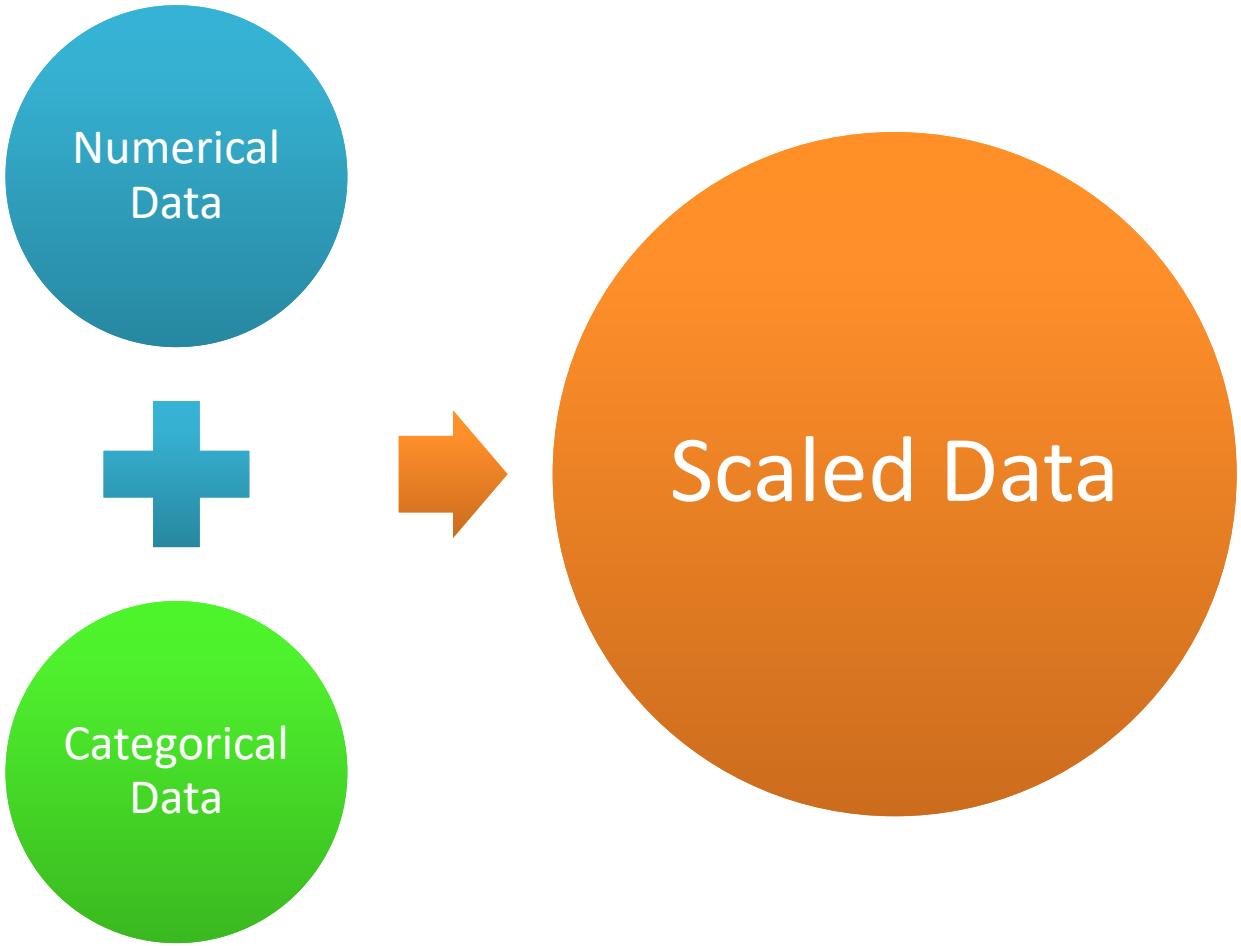
Missing Values



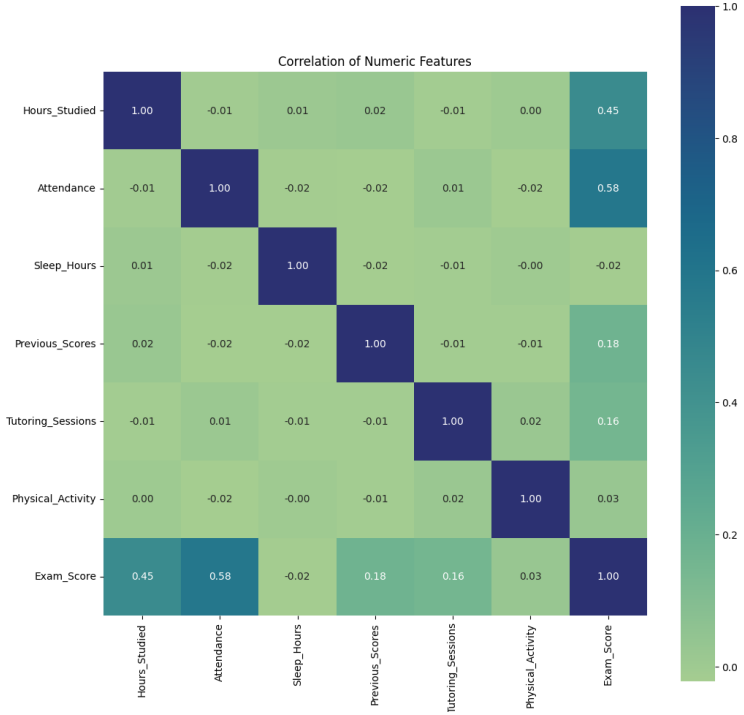
Mode Imputation

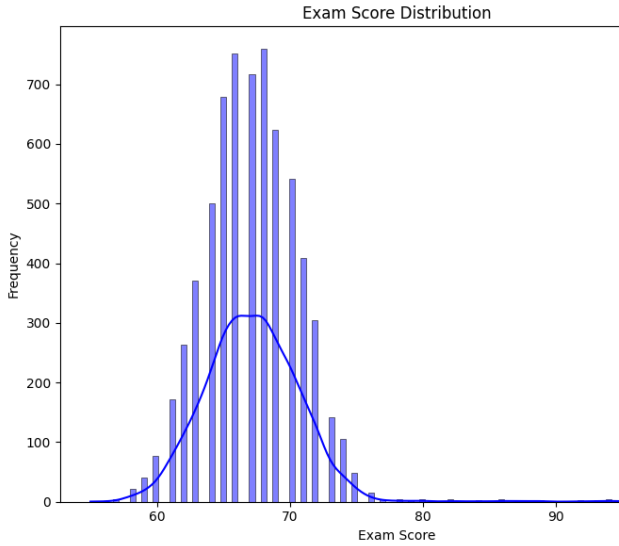


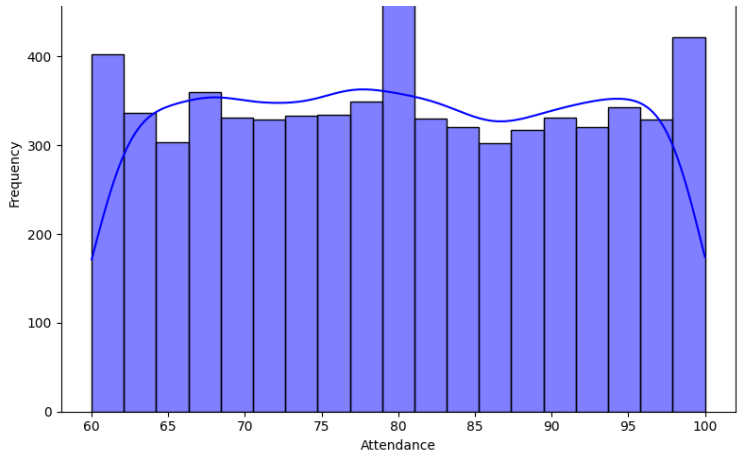
Data Cleaning & Preprocessing

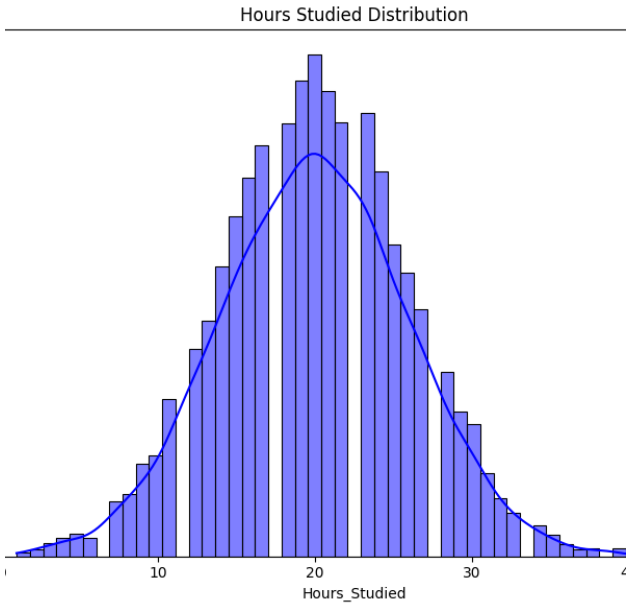


Exploratory Data Analysis



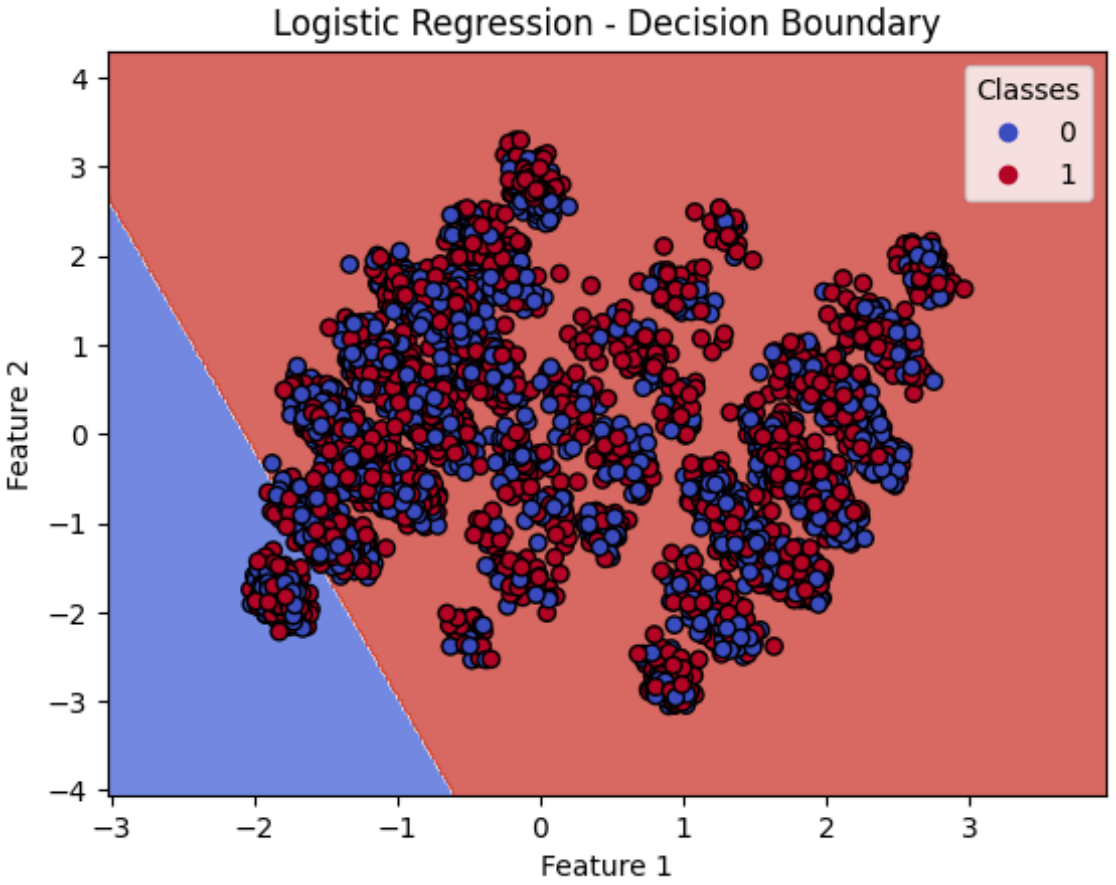






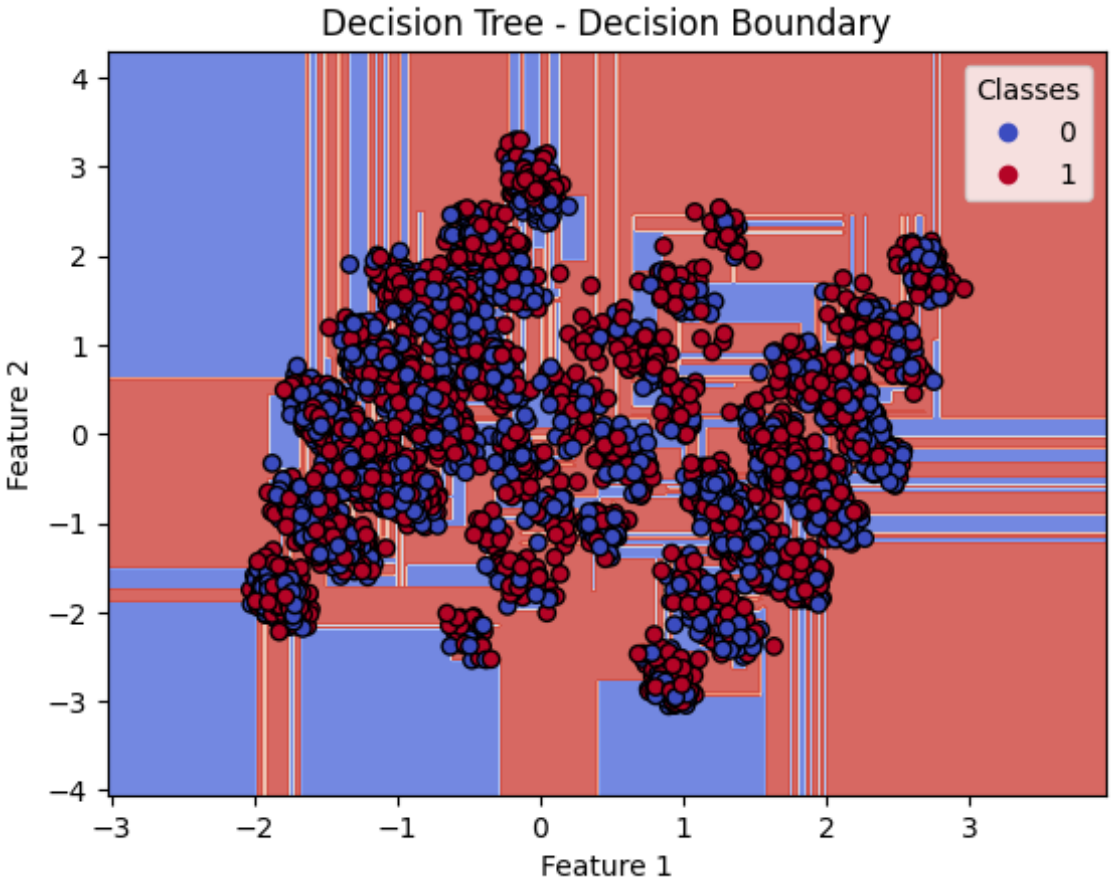
Logistic Regression

- **Logistic Regression**: A statistical model that predicts the probability of a binary outcome based on input features by modeling relationships with a logistic function.
- Logistic Regression is a linear classifier, meaning the decision boundary is a straight line or hyperplane.



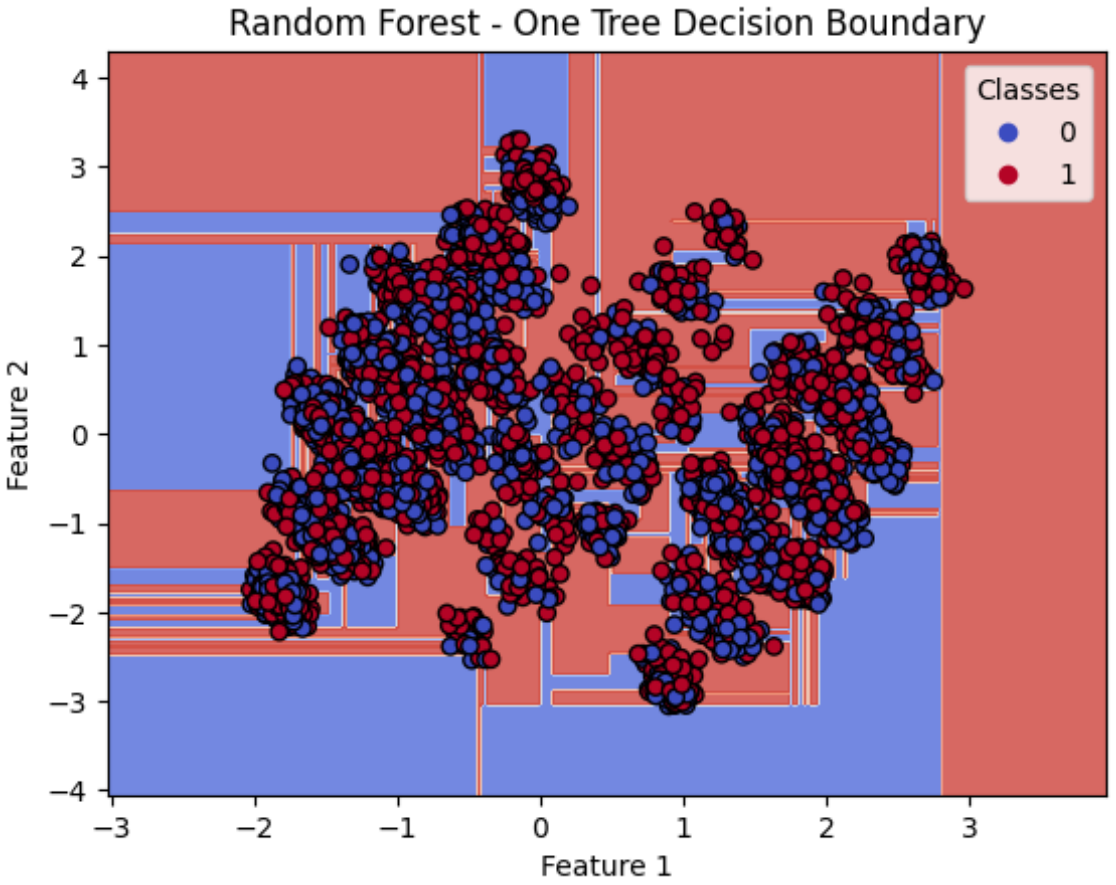
Decision Tree

- **Decision Tree**: A tree-like model of decisions and their possible consequences, splitting data based on feature conditions to classify or predict outcomes.
- The decision boundary is not linear; it consists of vertical and horizontal lines



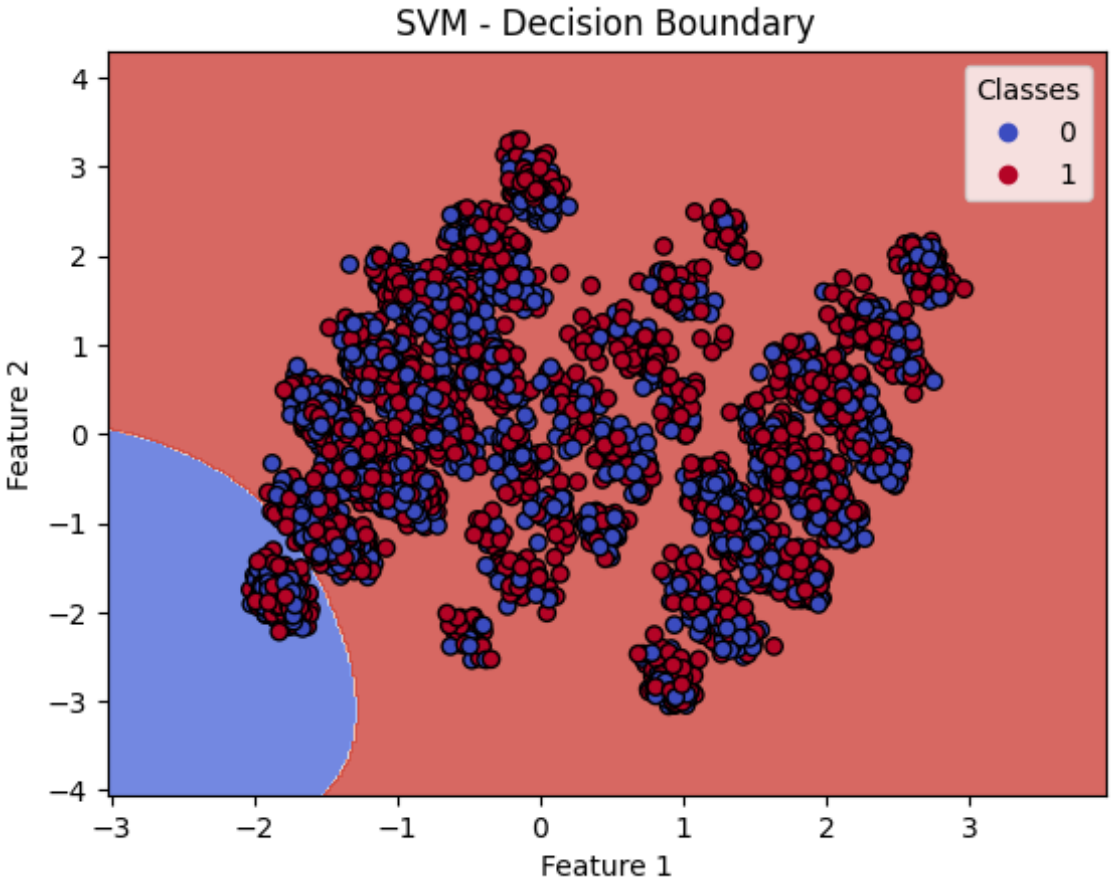
Random Forest

- **Random Forest**: An ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
- Similar to a Decision Tree, but it may show more nuanced boundaries, especially if one tree in the forest has different splits from others.



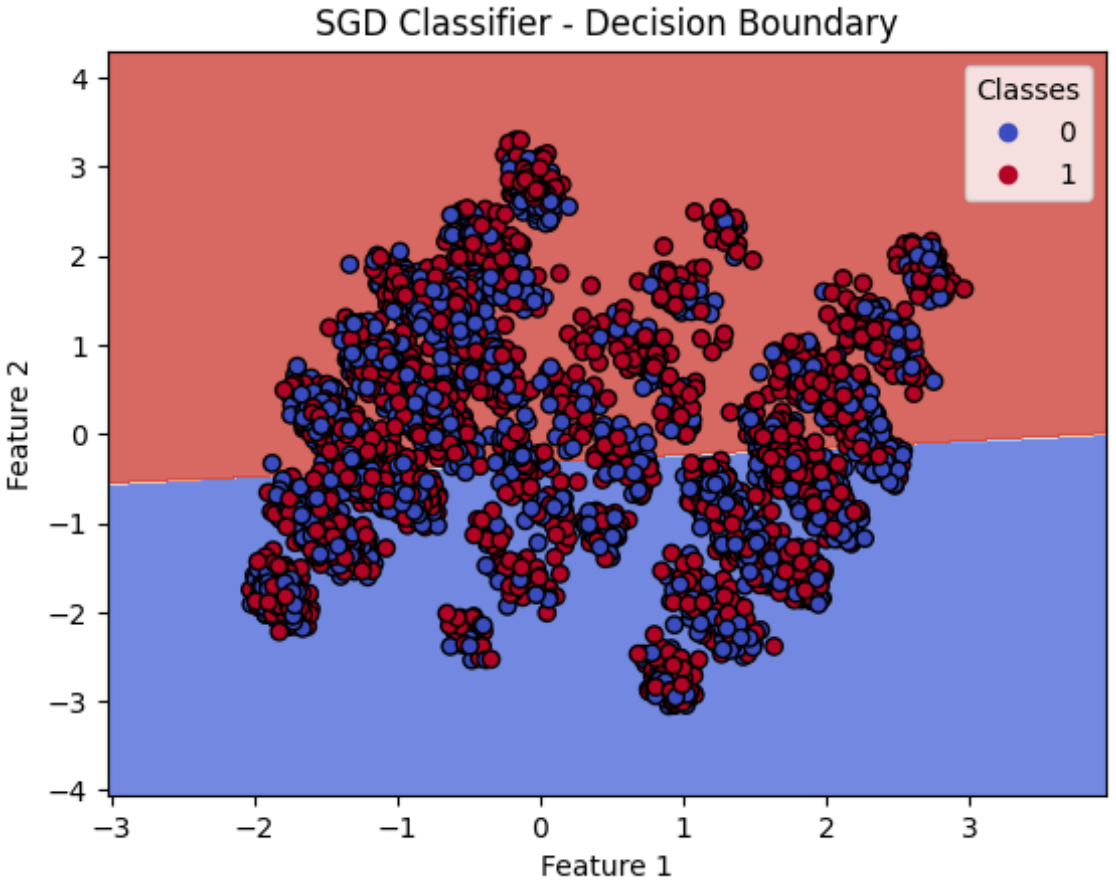
Support Vector Machine (SVM)

- **SVM (Support Vector Machine)**: A supervised learning algorithm that finds the optimal hyperplane to separate data into distinct classes, maximizing the margin between them.
- SVM typically creates a linear decision boundary if using a linear kernel, but it can also generate non-linear boundaries with other kernels (e.g., polynomial, radial basis function).



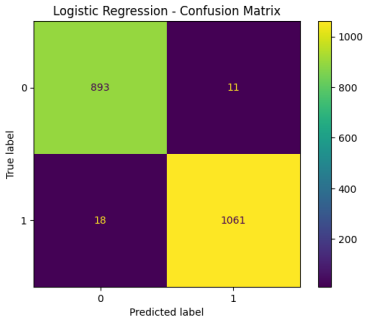
Stochastic Gradient Descent (SGD)

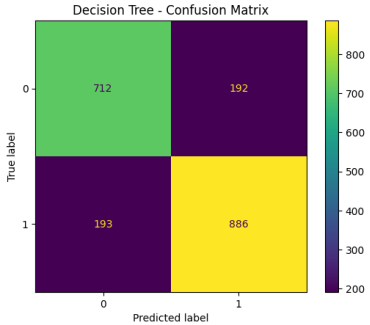
- **SGD (Stochastic Gradient Descent)**: An optimization algorithm that updates model parameters iteratively by minimizing a loss function using small random data batches.
- SGD Classifier will create a plot showing the decision boundary tree.

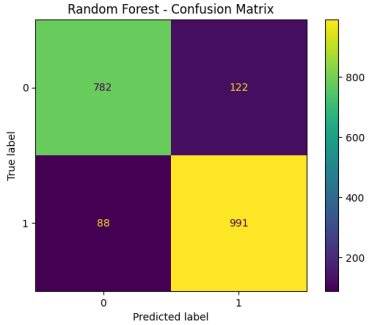


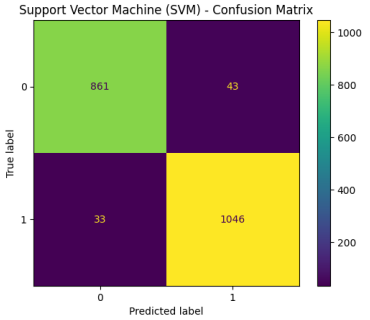
Model Evaluation

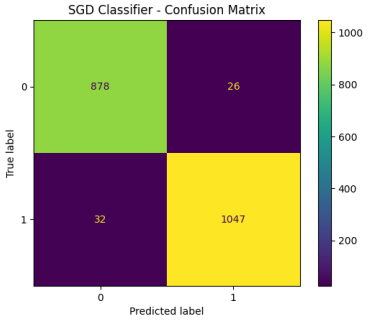
	Logistic Regression	Decision Tree	Random Forest	SVM	SGD
Accuracy	0.99	0.81	0.90	0.96	0.97
ROC	0.99426	0.80999	0.96458	0.99155	-



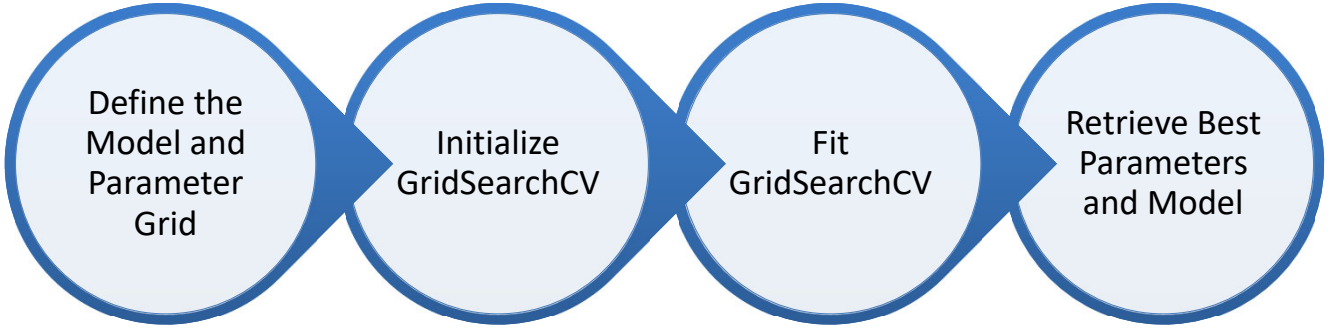








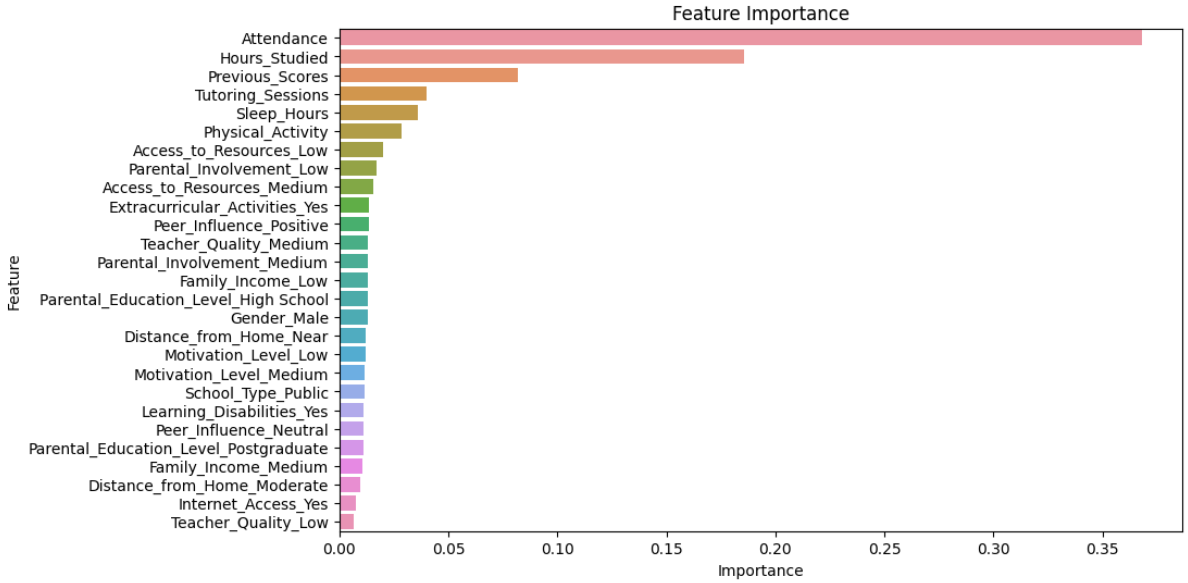
Tuning Hyperparameters - GridSearchCV



	Logistic Regression	Decision Tree	Random Forest	SVM	SGD
Hyper-Parameters	{'C': 1, 'max_iter': 500, 'solver': 'lbfgs'}	{'criterion': 'entropy', 'max_depth': 5, 'min_samples_split': 2}	{'max_depth': None, 'min_samples_split': 5, 'n_estimators': 200}	{'C': 1, 'kernel': 'linear'}	{'alpha': 0.0001, 'loss': 'hinge', 'max_iter': 500}
Default	0.99	0.81	0.90	0.96	0.97
Grid Search	0.98	0.85	0.90	0.98	0.94

Feature Selector

	Default	Grid Search	Pearson	Chi-Square	Random Forest	RFE
Logistic Regression	0.99	0.98	0.93	0.93	0.93	0.63
Decision Tree	0.81	0.85	0.82	0.82	0.82	0.58
Random Forest	0.90	0.90	0.89	0.89	0.89	0.59
SVM	0.96	0.98	0.92	0.92	0.92	0.61
SGB	0.97	0.94	0.92	0.91	0.91	0.61



Conclusion

- “Attendance" and “Hours_studied" are the most significant predictors of performance.
- Students with higher tutoring hours, access to Resources also performed better.
- The more a student spends time on academics, the higher their performance.
- Logistic Regression and Support Vector Machine (SVM) outperformed all other models in terms of decision making, accuracy and generalizability.

What’s next

- Exam score can be from 0 to 100 not just from 60 to 100
- Include new useful features incorporating additional features like real-time attendance tracking, tracking Assignments and Mental Health
- An expert opinion is essential to validate feature selection and data preprocessing.