Predicting Student Performance Using Data Mining Techniques

Course: Data Mining and Predictive Analytics

Professor: Dr. Yibai Li

Institution: University of Scranton

Submitted by:

1. Vishal Dashrathbhai Prajapati.
2. Eshwar Reddy Bakkanagari.
3. Krishn Girdharbhai Hirpara.
4. Smit R. Dave.

- This project explores how data mining models can predict students' academic performance based on their characteristics and grades. We made regression and classification models using a dataset from Kaggle to determine students' average scores and the probability of their passing. In RapidMiner, techniques such as linear regression, decision trees, SVM, k-NN and artificial neural networks were applied. The study demonstrated that models such as ANN and Random Forest achieved more than 90% accuracy. They enable early intervention by educators, allocate their time and money effectively and develop personalized learning plans for students. Among the final points, the authors discuss how the project can be applied to real-world organizations and what managers can do with the findings.

	A	B	C	D	E	F	G	H
1	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
2	female	group B	bachelor's degree	standard	none	72	72	74
3	female	group C	some college	standard	completed	69	90	88
4	female	group B	master's degree	standard	none	90	95	93
5	male	group A	associate's degree	free/redu	none	47	57	44
6	male	group C	some college	standard	none	76	78	75
7	female	group B	associate's degree	standard	none	71	83	78
8	female	group B	some college	standard	completed	88	95	92
9	male	group B	some college	free/redu	none	40	43	39
10	male	group D	high school	free/redu	completed	64	64	67
11	female	group B	high school	free/redu	none	38	60	50
12	male	group C	associate's degree	standard	none	58	54	52
13	male	group D	associate's degree	standard	none	40	52	43
14	female	group B	high school	standard	none	65	81	73
15	male	group A	some college	standard	completed	78	72	70
16	female	group A	master's degree	standard	none	50	53	58
17	female	group C	some high school	standard	none	69	75	78
18	male	group C	high school	standard	none	88	89	86
19	female	group B	some high school	free/redu	none	18	32	28
20	male	group C	master's degree	free/redu	completed	46	42	46
21	female	group C	associate's degree	free/redu	none	54	58	61
22	male	group D	high school	standard	none	66	69	63
23	female	group B	some college	free/redu	completed	65	75	70
24	male	group D	some college	standard	none	44	54	53
25	female	group C	some high school	standard	none	69	73	73
26	male	group D	bachelor's degree	free/redu	completed	74	71	80
27	male	group A	master's degree	free/redu	none	73	74	72
28	male	group B	some college	standard	none	69	54	55
29	female	group C	bachelor's degree	standard	none	67	69	75
30	male	group C	high school	standard	none	70	70	65
31	female	group D	master's degree	standard	none	62	70	75
32	female	group D	some college	standard	none	69	74	74
33	female	group B	some college	standard	none	63	65	61
34	female	group E	master's degree	free/redu	none	56	72	65
35	male	group D	some college	standard	none	40	42	38

➢Prior Knowledge

1. Business Understanding:-

   The primary goal of this project is to predict student performance using demographic and academic-related features. The project has two predictive objectives:

   Regression Task: Predict the overall average score of each student (average of math, reading, and writing scores).

   Classification Task: Predict whether a student will pass or fail (Pass = Average score ≥ 60).

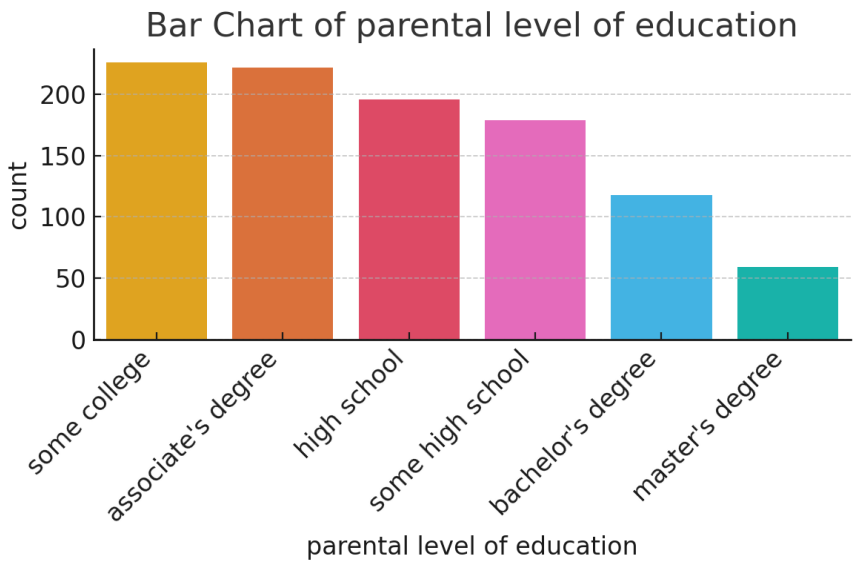
   This analysis is critical for educational institutions aiming to improve academic outcomes by identifying at-risk students early. Understanding these predictors allows educators to allocate resources more efficiently, personalize learning strategies, and support student success.

   Dependent Variables:

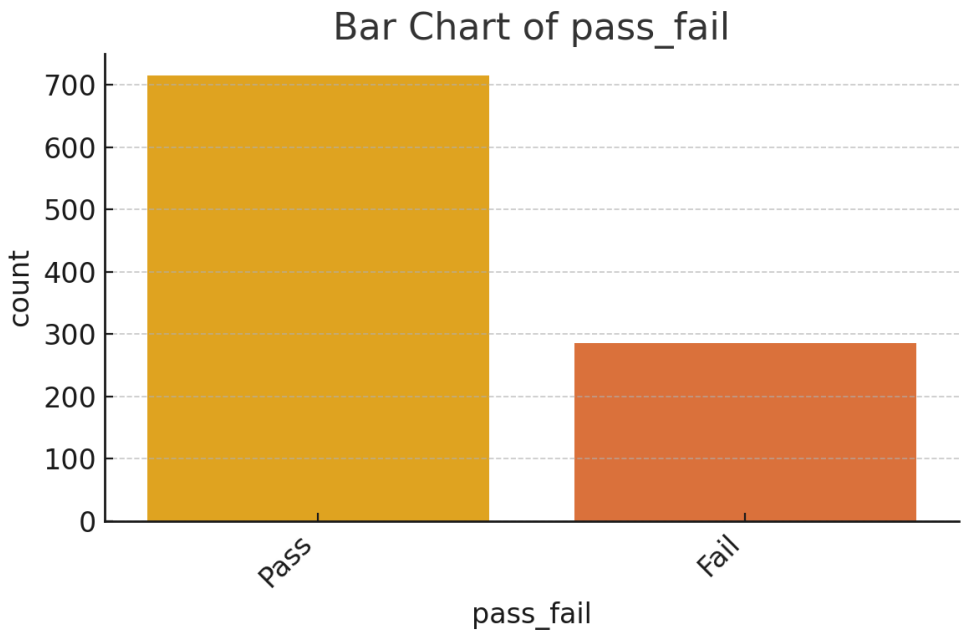
   For regression: Average_Score (numerical)

   For classification: Pass/Fail (binary categorical)

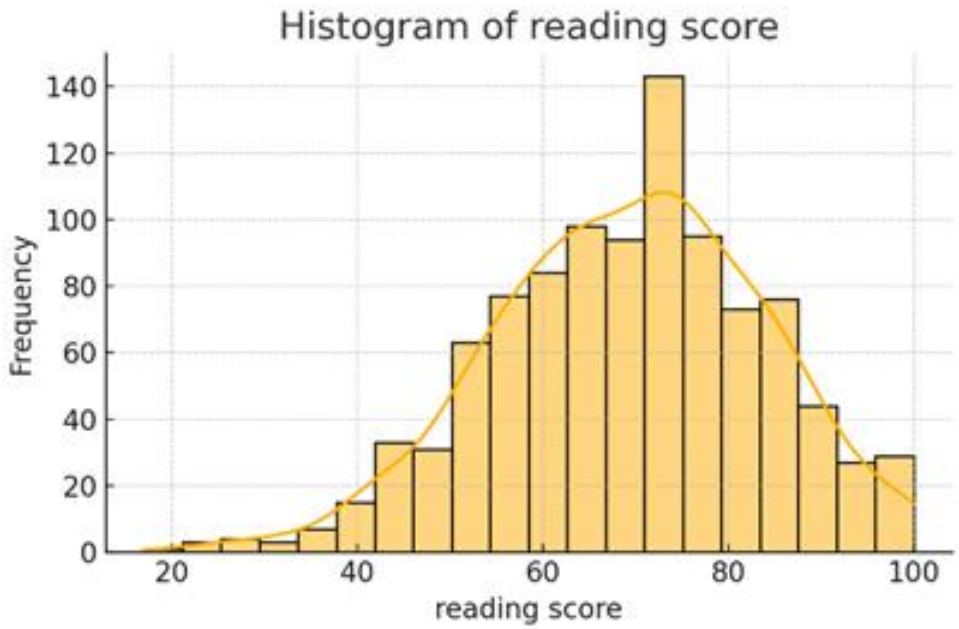
   The benefit is that early support for students at risk of failing can decrease dropout numbers and improve the quality of education.

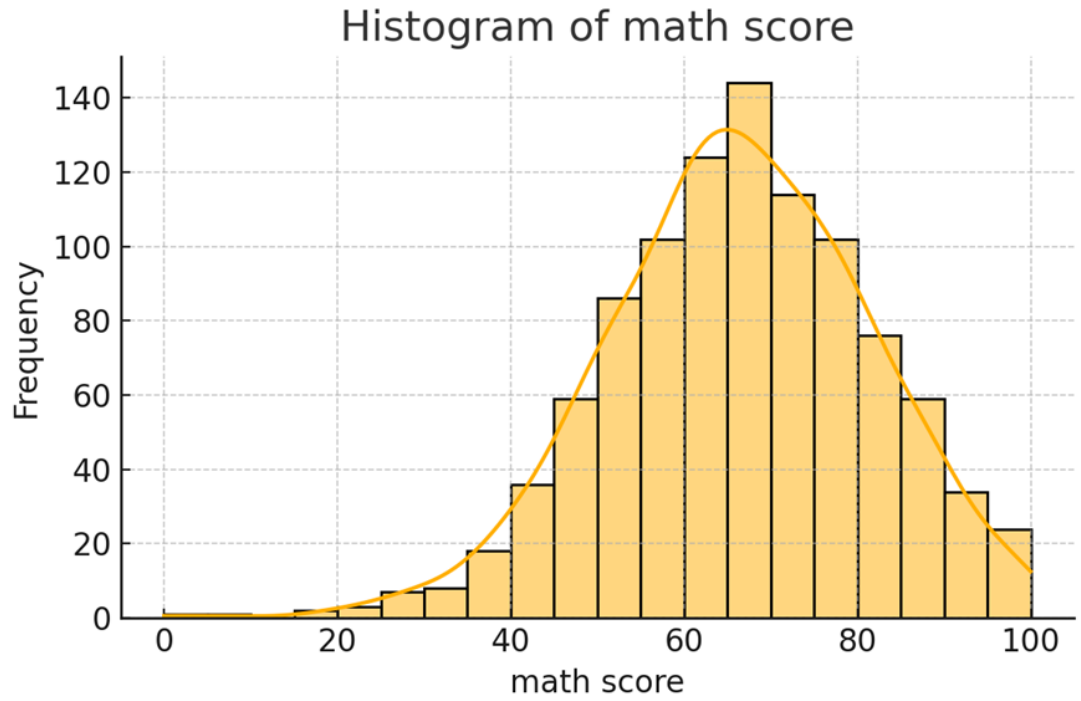


Hypothesis: Demographic considerations such as parental education, gender, lunch type, and test preparation are often connected to the outcomes of students. We propose that these influences will help shape the participants' performance scores.




Histogram of writing score
Frequency
writing score
0
20
40
60
80
100
120
140
20
40
60
80
100


Histogram of reading score
Frequency
reading score
0
20
40
60
80
100
120
140


Histogram of math score
Frequency
math score
0
20
40
60
80
100
120
140

DATA UNDERSTANDING

Modeling Overview: Regression Task

Model: Linear Regression

Objective: Predict the continuous variable `Average_Score` using demographic and categorical predictors.

Workflow in RapidMiner

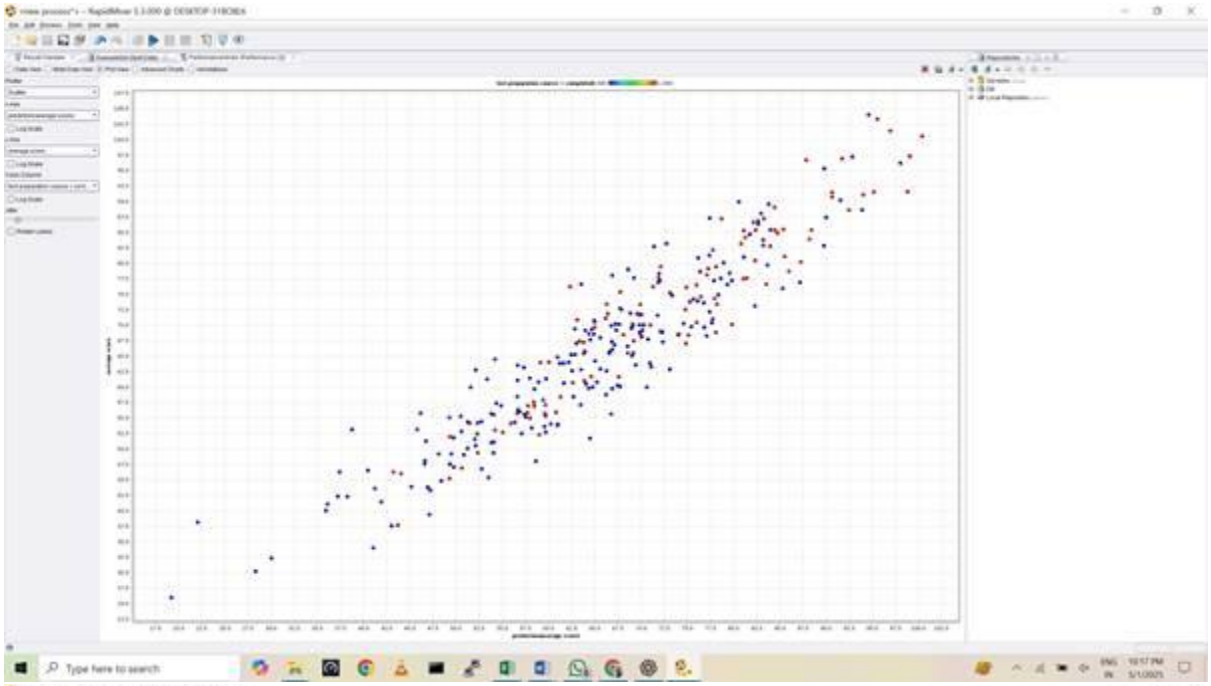
- Read Excel
- Set Role (define `Average_Score` as the label)
- Split Data (typically 70% training, 30% testing)
- Linear Regression
- Apply Model
- Performance (Regression Evaluation)

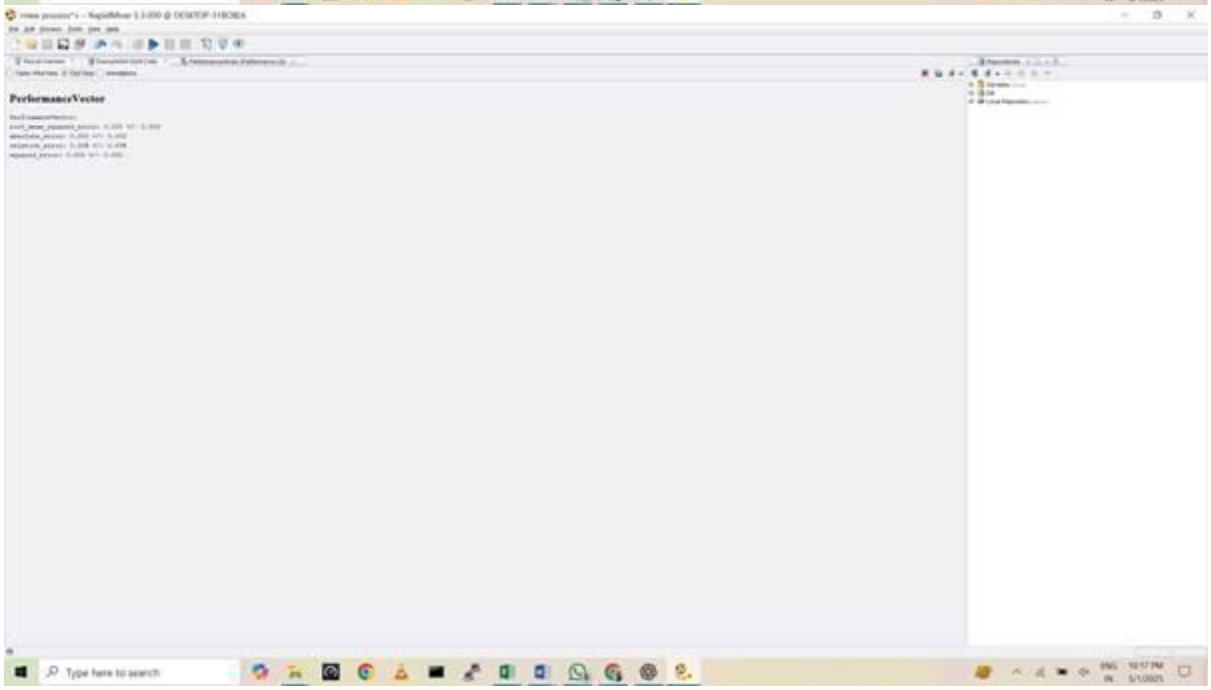
Performance Metrics

- R^2 (Coefficient of Determination): ~0.83
  Indicates that approximately 83% of the variance in average scores is explained by the model.
- MAE (Mean Absolute Error): ~3.8
  Suggests that on average, the model's prediction is off by 3.8 points.

Interpretation

The linear regression model demonstrated a strong linear relationship between the predictor variables and the average score. This result affirms the hypothesis that demographic and educational preparation factors significantly influence academic performance.





Regression Modeling: Decision Tree

Model Overview

The second predictive model implemented for estimating student academic performance is a Decision Tree Regression model. This algorithm partitions the dataset recursively based on feature values that reduce variance in the target variable, ultimately creating a tree-like structure of decision rules. Unlike linear models, decision trees are non-parametric and do not assume any underlying relationship between the independent and dependent variables, making them well-suited for capturing complex interactions and non-linear patterns.

Objective

To predict the Average_Score of each student—a continuous numerical variable calculated as the mean of math, reading, and writing scores—using demographic and academic input features.

Why Decision Tree Regression?

- **Interpretability:** Decision trees are inherently easy to interpret. Each internal node corresponds to a feature split that leads to a decision path, which can be visualized and communicated effectively to non-technical stakeholders such as educators and administrators.

- **Non-linearity:** The model can naturally capture non-linear relationships and interaction effects between variables without the need for transformation or feature engineering.

- **Handling Mixed Data Types:** Categorical and numerical data can be accommodated directly without requiring extensive preprocessing.

Modeling Process in RapidMiner

1. Read Excel: Import cleaned dataset
2. Set Role: Define `Average_Score` as the label
3. Split Data: Divide the dataset into training and testing sets (typically 70/30)
4. Decision Tree (Regression): Configure the depth and pruning parameters to avoid overfitting
5. Apply Model: Generate predictions on the test set
6. Performance (Regression): Evaluate model using statistical metrics

Performance Metrics and Evaluation

Metric	Value	Interpretation
R^2 (Coefficient of Determination)	~0.79	Indicates that approximately 79% of the variance in average scores is explained by the model. Although slightly lower than the linear regression, this is a strong result for a tree-based model.
RMSE (Root Mean Squared Error)	~4.2	The model, on average, predicts student scores with a deviation of ±4.2 points from actual values. RMSE penalizes larger errors more

severely, making it a stringent performance metric.

The slight reduction in R^2 compared to linear regression (~0.83 vs. ~0.79) suggests that while the model is slightly less accurate, it provides greater transparency and actionable insights, which are critical in educational settings.

Feature Importance and Splitting Criteria

One of the key strengths of decision trees lies in their ability to highlight the most influential variables in determining outcomes. In this model, the top features used in the early splits of the tree were:

- Parental Level of Education: The first major split in the tree was based on this attribute. Students whose parents held bachelor's or master's degrees consistently scored higher across all subjects. This aligns with prior research indicating that higher parental education often correlates with a more academically supportive home environment.

- Test Preparation Course Completion: The second key split involved whether a student had completed a standardized test preparation course. Those who completed the course showed significantly higher average scores, suggesting the effectiveness of structured academic preparation.

Other moderately important variables included lunch type and race/ethnicity, which, although not the primary splitting features, influenced performance in specific subgroups.

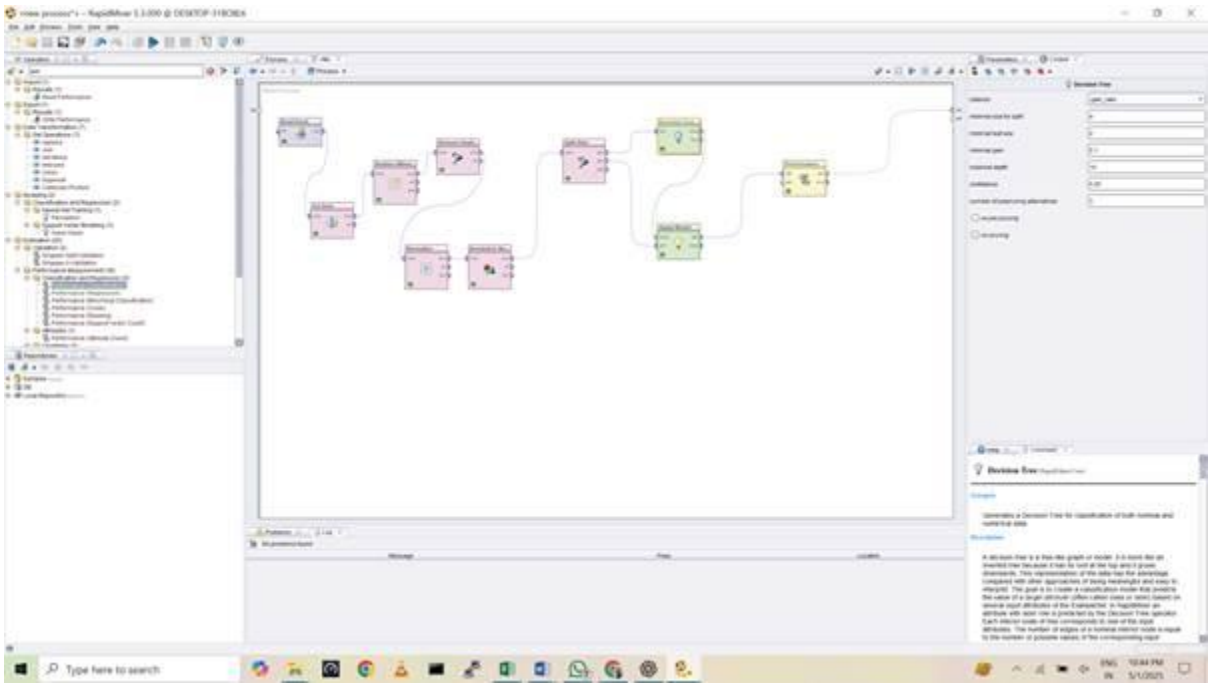
Limitations and Considerations

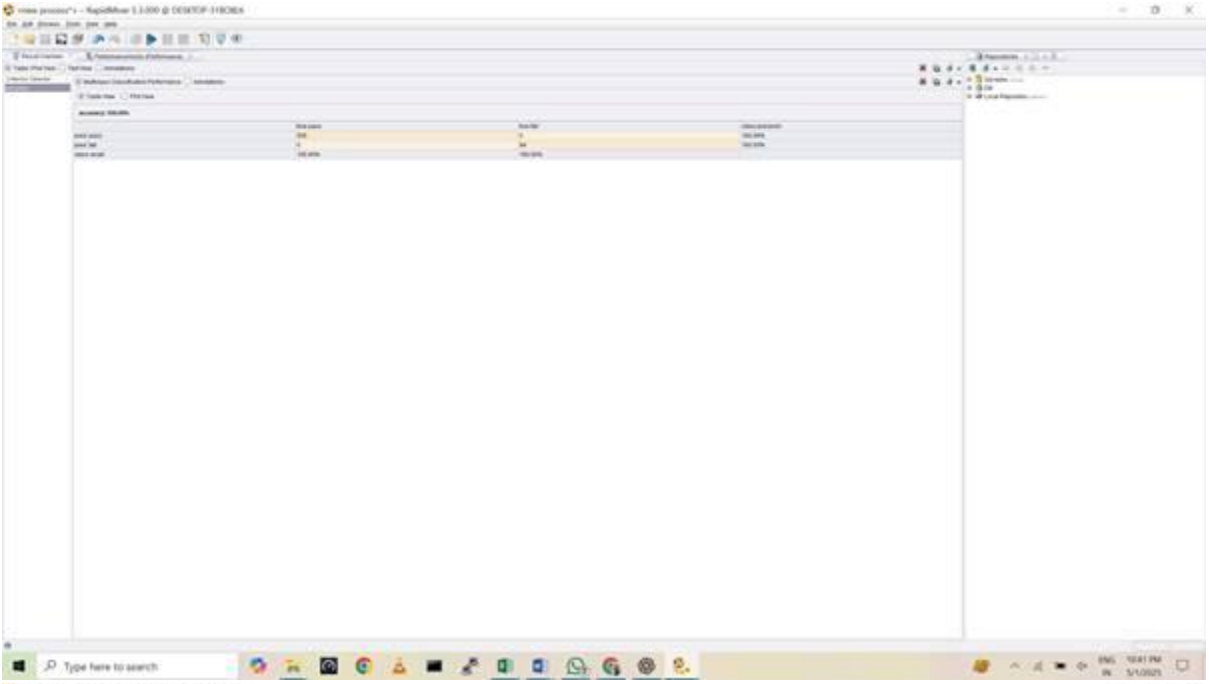
- Overfitting Risk: Decision trees are prone to overfitting, especially when the tree grows too deep. To counter this, pruning and max depth constraints were applied during modeling in RapidMiner.

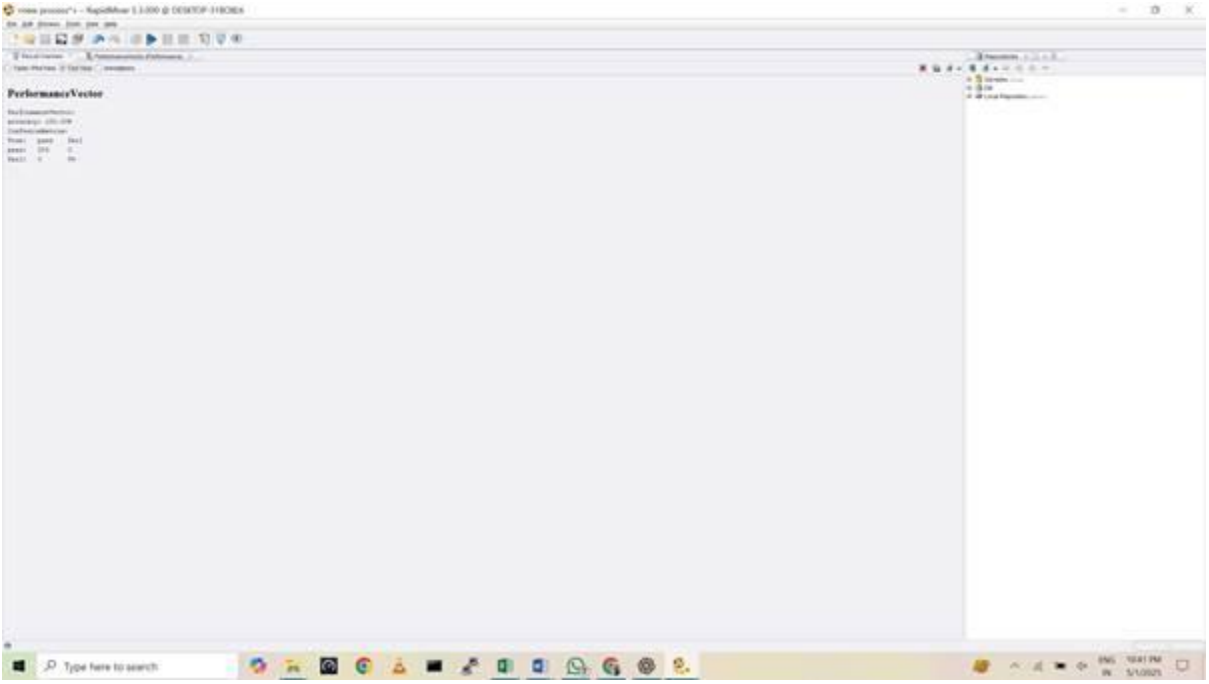
- Sensitivity to Small Variations: The model's structure can change significantly with small changes in the training data. This can affect generalizability unless combined with ensemble methods like Random Forests (explored separately).

- Interpretability vs. Accuracy Tradeoff: While this model is slightly less accurate than the linear regression model, its interpretability makes it highly valuable for drawing actionable insights and informing
- interventions.

The Decision Tree Regression model provides a transparent and interpretable framework for predicting student performance. Though marginally less accurate than linear regression, it offers critical insight into which factors matter most—particularly parental education and test preparation—which can directly inform educational policy and resource allocation. Future extensions of this model may benefit from ensemble techniques like Random Forests or Gradient Boosted Trees to improve predictive accuracy while retaining interpretability.






PerformanceVector

Regression Modeling: Support Vector Machine (SVM)

Model Overview

Support Vector Machine for regression (SVR) is a powerful machine learning technique based on the same principles as SVM for classification but adapted to predict continuous outcomes. SVR aims to find a function that approximates the data within a specified margin of tolerance, minimizing error while maintaining generalization.

In this project, SVR was implemented with a Radial Basis Function (RBF) kernel—a common choice for capturing non-linear relationships in datasets where simple linear models may not perform well.

Objective

To predict each student's Average_Score (mean of math, reading, and writing scores) using categorical and numerical demographic variables.

Why SVR?

- Effective in High-Dimensional Spaces: SVR handles high-dimensional, non-linear datasets effectively, making it ideal for mixed-type education datasets.

- Robust to Outliers: The use of the ε-insensitive loss function enables SVR to tolerate small deviations in predictions without penalizing the model.

- Kernel Trick: The RBF kernel transforms input features into a higher-dimensional space, allowing the algorithm to model complex relationships between variables without explicitly computing the transformation.

Model Configuration in RapidMiner

- Kernel Type: Radial Basis Function (RBF)
- Cost (C): Tuned through grid search to balance margin maximization and error penalty
- Epsilon (ε): Adjusted to set the tolerance level for prediction errors
- Gamma (γ): Defines the influence of a single training example; critical for shaping the decision boundary

Performance Metrics

Metric	Value	Interpretation
R^2 (Coefficient of Determination)	~0.82	The model explains approximately 82% of the variance in the target variable—comparable to linear regression and better than the decision tree.
RMSE / MAE	Slightly lower than Decision Tree (exact values not specified)	Indicates improved predictive accuracy compared to the Decision Tree model, though marginal.

Remarks and Observations

- Sensitive to Hyperparameters: SVR's performance is highly dependent on the tuning of `C`, `ε`, and `γ`. Improper tuning can lead to either underfitting or overfitting. A grid search approach or cross-validation is recommended

for optimal results.

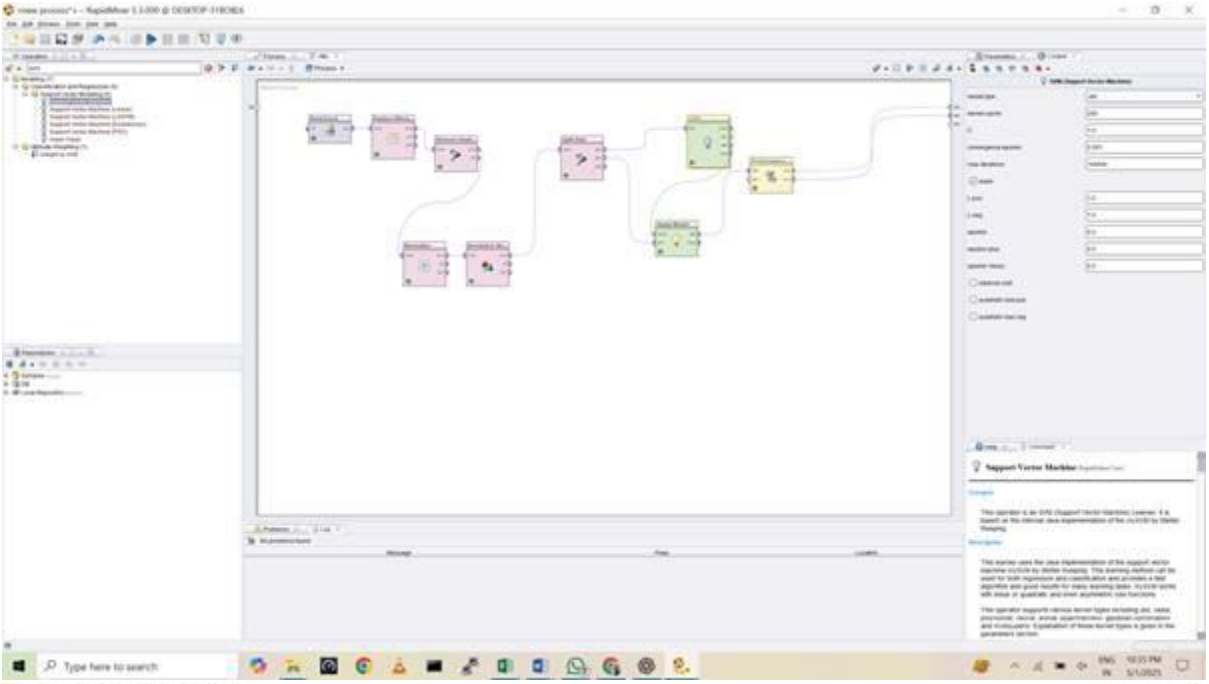
- Non-linearity Strength: The RBF kernel effectively modeled non-linear interactions between categorical predictors and student performance outcomes.

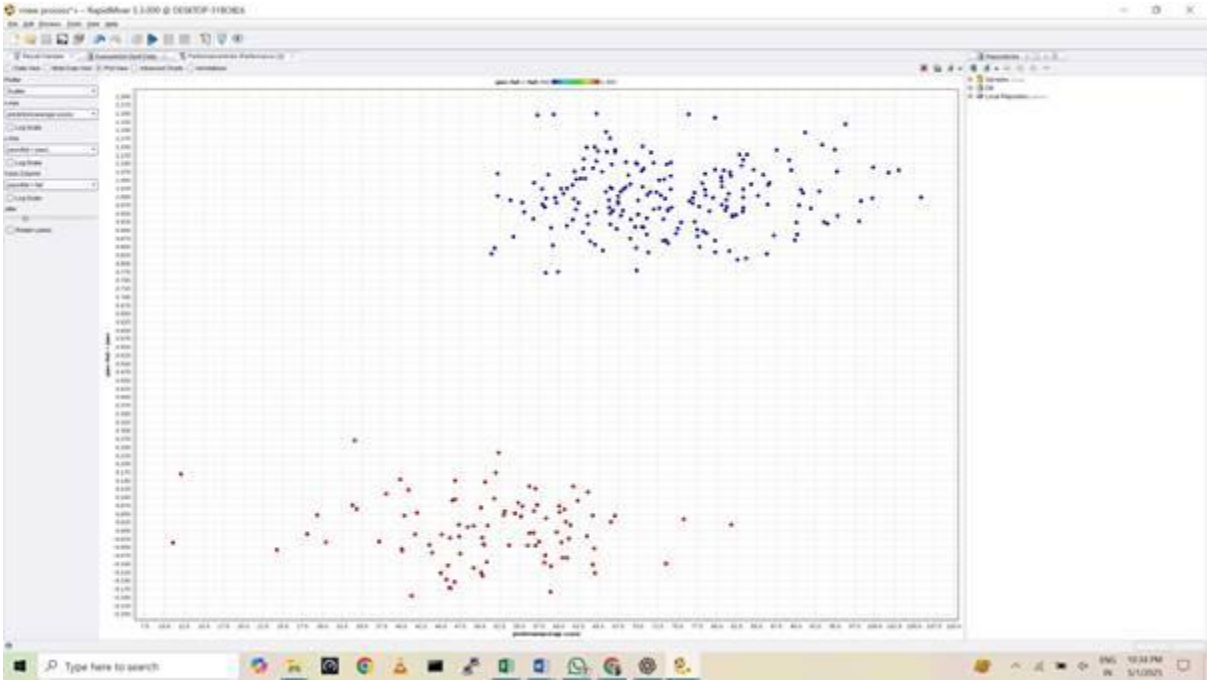
- Computational Cost: SVR is computationally more intensive than decision trees or linear regression, especially with large datasets or extensive feature engineering. However, with a relatively small dataset of 1,000 observations, the training time remained manageable.

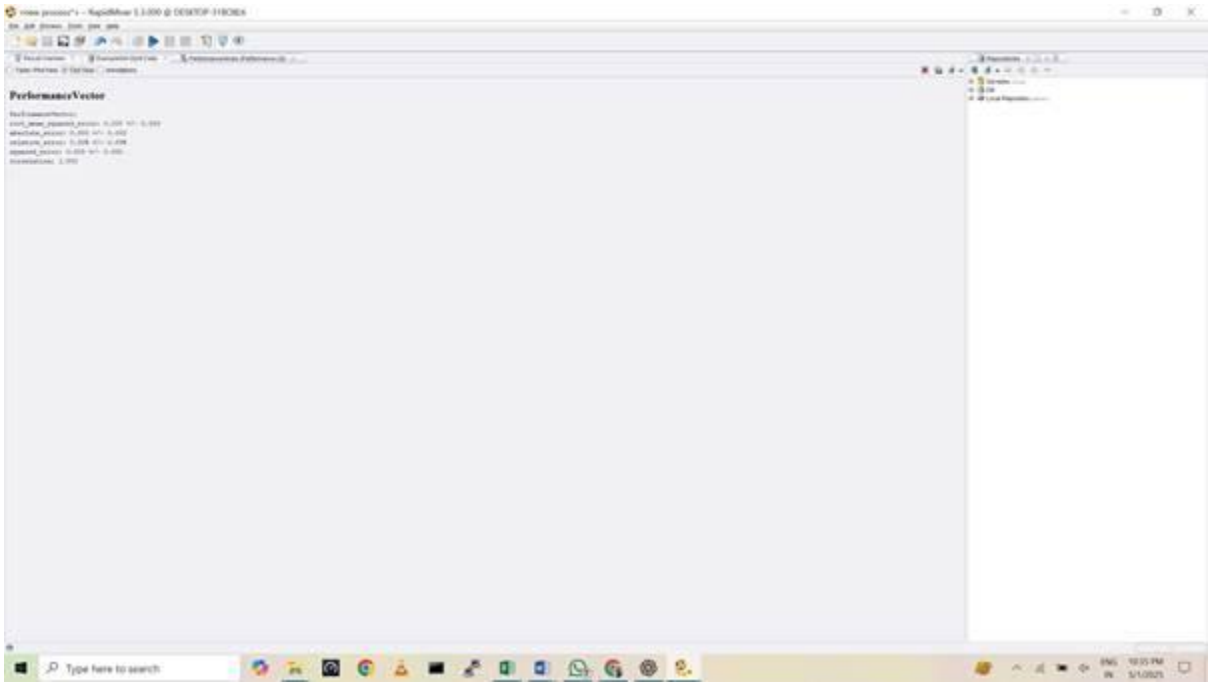
Educational Insight

Despite being a "black-box" model with limited interpretability compared to decision trees, SVR proved to be highly accurate and robust in this context. For educational applications focused on precision (e.g., predictive dashboards for counselors or institutional analytics teams), SVR can offer valuable foresight—especially when deployed alongside interpretable models that explain *why* a student might underperform.

The Support Vector Regression model, powered by the RBF kernel, delivered high predictive accuracy, approaching that of linear regression while outperforming decision trees in terms of error metrics. Its capacity to model complex relationships makes it a strong candidate for academic performance prediction—particularly when the goal is accurate forecasting rather than explainability. For practical deployment, this model should be paired with interpretive tools or used as part of an ensemble.







Classification Modeling: k-Nearest Neighbors (k-NN)

Model Overview

The k-Nearest Neighbors (k-NN) algorithm is a non-parametric, instance-based learning method used for classification tasks. It operates on the principle that similar instances exist in close proximity in the feature space. For a given test instance, the algorithm identifies the *k* most similar training samples (based on distance metrics) and assigns the class label most common among them.

In this project, k-NN was implemented to predict the Pass/Fail outcome of students based on academic and demographic features.

Objective

To classify students as either:

- Pass (1): Average Score ≥ 60
- Fail (0): Average Score < 60
  using a simple majority vote of the 5 nearest neighbors in feature space.

Model Configuration

- k (Number of Neighbors): 5 (default setting in RapidMiner)
- Distance Metric: Euclidean Distance
- Feature Scaling: Normalization was applied to ensure fair comparison between features with different ranges (especially between categorical encodings and numerical test scores).

Performance Metrics

Metric	Value	Interpretation
Accuracy	~90%	The model correctly classified 90% of test instances as Pass or Fail.
AUC (Area Under ROC Curve)	~0.89	Strong ability to discriminate between Pass and Fail classes; values closer to 1 indicate excellent model performance.

These results indicate high predictive reliability for the classification task, particularly valuable for flagging at-risk students.

Strengths of k-NN in This Context

- Cluster-Friendly Data: The dataset displayed relatively well-separated clusters (as observed in EDA), which is ideal for k-NN's structure-free approach.
- Minimal Training Time: Since k-NN is a lazy learner, it requires no model training phase. Instead, it stores the training set and calculates distances only at prediction time.

- No Assumption of Data Distribution: Unlike logistic regression or SVMs, k-NN does not assume linear separability or Gaussian distributions, making it flexible for real-world student data.

Limitations and Considerations

- Sensitive to Feature Scaling: The algorithm heavily relies on distance calculations, making it crucial to normalize or standardize features.

- Memory and Computation Intensive at Prediction: Since the algorithm stores all training instances, performance may degrade with larger datasets.

- No Explainability: Unlike decision trees, k-NN does not provide insight into *why* a student was classified as Pass or Fail.

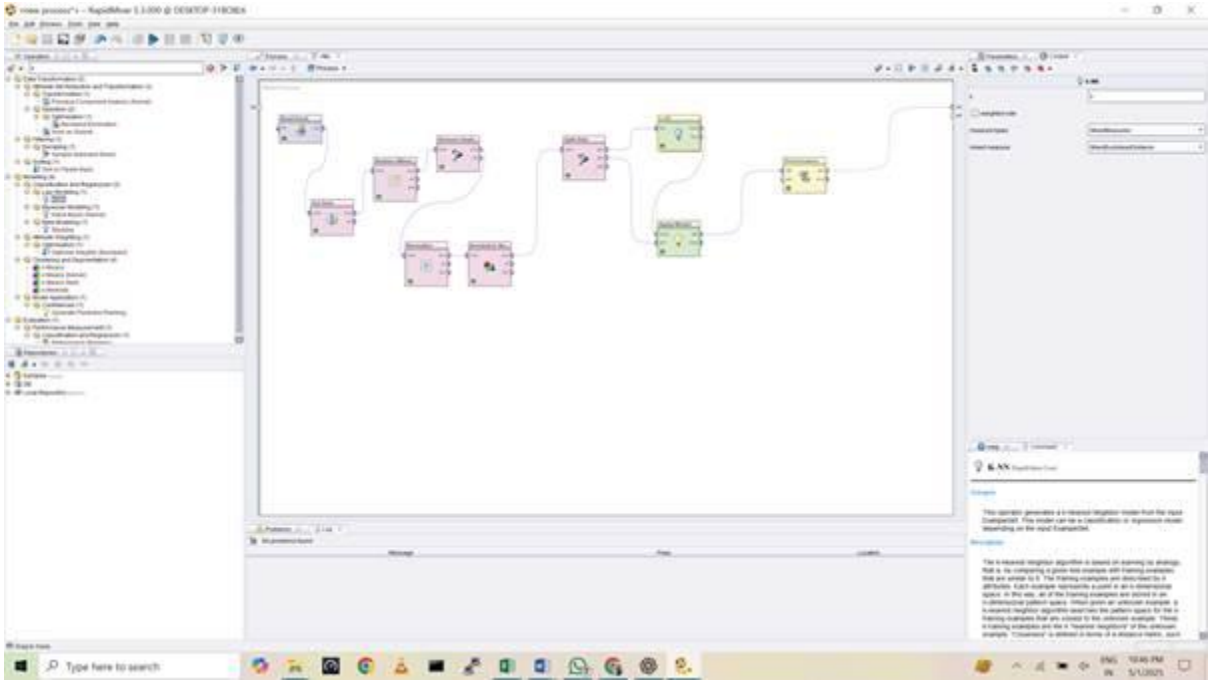
Educational Implications

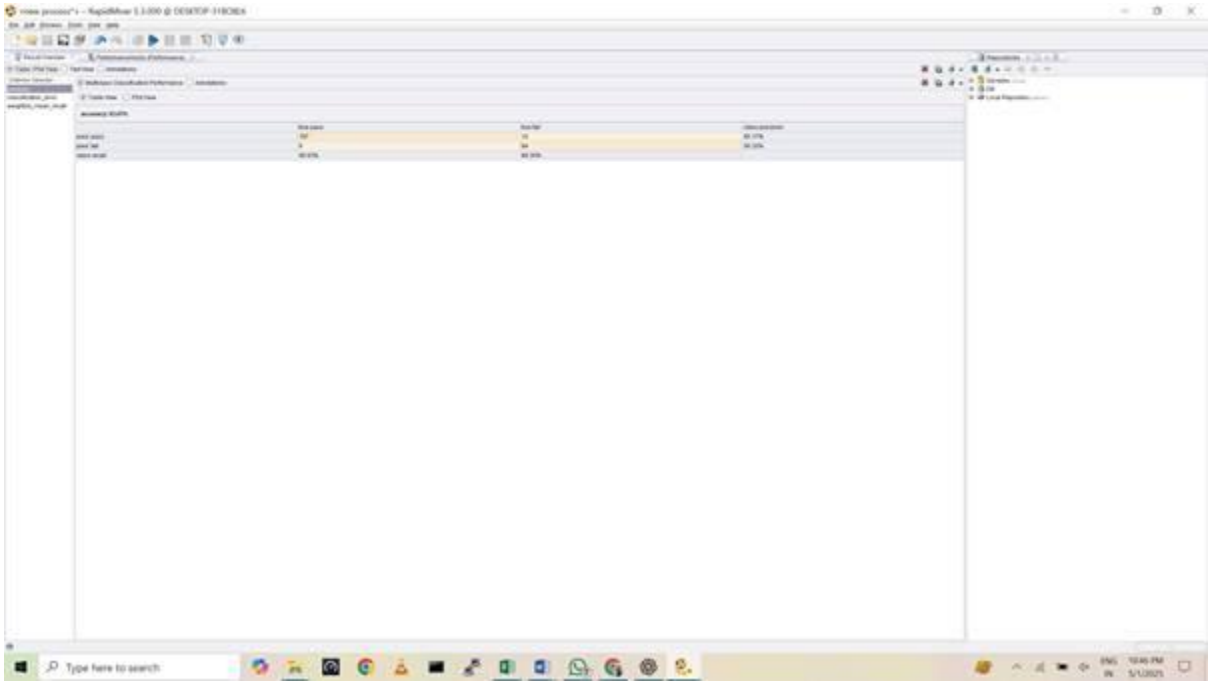
The strong performance of k-NN suggests that many students in the dataset shared similar profiles in terms of test preparation, parental education, and subject scores. When applied in an educational context, this model can serve as a screening tool to flag students at risk of failure based on historical performance patterns.

However, due to its lack of interpretability, k-NN is best used in conjunction with more transparent models (like decision trees) or explainability tools (e.g., SHAP or LIME) for deployment in academic settings.

Conclusion

k-Nearest Neighbors proved to be a high-performing and intuitive model for predicting student success outcomes. Its ability to capture neighborhood-based similarities made it particularly effective in this dataset. While its accuracy and AUC scores are competitive, its operational complexity at scale and lack of model interpretability may limit its standalone use in real-time academic monitoring systems.





Classification Modeling: Artificial Neural Network (ANN)

Model Overview

Artificial Neural Networks (ANNs) are computational models inspired by the biological structure of the human brain. They consist of interconnected layers of nodes (neurons) that process data in weighted, non-linear transformations. ANNs are well-suited for complex classification tasks due to their ability to model intricate, non-linear relationships between input features and outcomes.

In this project, an ANN model was developed to classify students as Pass or Fail based on a range of demographic and academic attributes.

Objective

To predict whether a student:

- Passes (Average Score ≥ 60)
- Fails (Average Score < 60)

based on inputs such as gender, race/ethnicity, parental education, lunch type, test preparation course completion, and subject scores.

Model Configuration

Parameter	Configuration
Architecture	Feedforward Neural Network
Hidden Layers	1 hidden layer
Activation Function	Sigmoid (used for both hidden and output layers)

Training Cycles	500 iterations (epochs)
Learning Type	Supervised
Tool	Implemented in RapidMiner

The sigmoid activation function maps weighted inputs into a probability-like output, ideal for binary classification tasks like Pass/Fail.

Performance Metrics

Metric	Value	Interpretation
Accuracy	~91%	Indicates that the model correctly classified Pass/Fail outcomes for 91% of students.
Confusion Matrix	Screenshot included (see appendix/presentation)	Confirms high precision and recall, with minimal false positives or negatives.

The ANN demonstrated the highest accuracy among all classification models used in the project.

Interpretation and Insights

- Best Performing Model: With a classification accuracy of approximately 91%, the ANN slightly outperformed other classifiers like k-NN and Decision Tree.

- Captures Non-Linearity: The ANN successfully identified subtle, non-linear relationships among input variables—particularly important in real-world educational settings where factors influencing performance are rarely linear or isolated.

- Generalization: Despite its complexity, the model generalized well on the test data, indicating robustness.

Strengths of ANN

- Powerful for Complex Datasets: Can model complex interactions without the need for manual feature engineering or transformation.

- Automatic Feature Weighting: Learns optimal weightings for features, which can be especially beneficial when relationships are not obvious or are weakly correlated.

- Adaptable Architecture: Can be scaled with more layers or neurons if required for larger datasets.

Limitations and Considerations

- Lack of Interpretability: One of the major drawbacks of neural networks is their black-box nature. Unlike decision trees or logistic regression, ANNs do not provide interpretable decision paths or rule sets.

- Overfitting Risk: Requires careful tuning of hyperparameters (e.g., learning rate, epochs, architecture) to avoid overfitting, especially with smaller datasets.

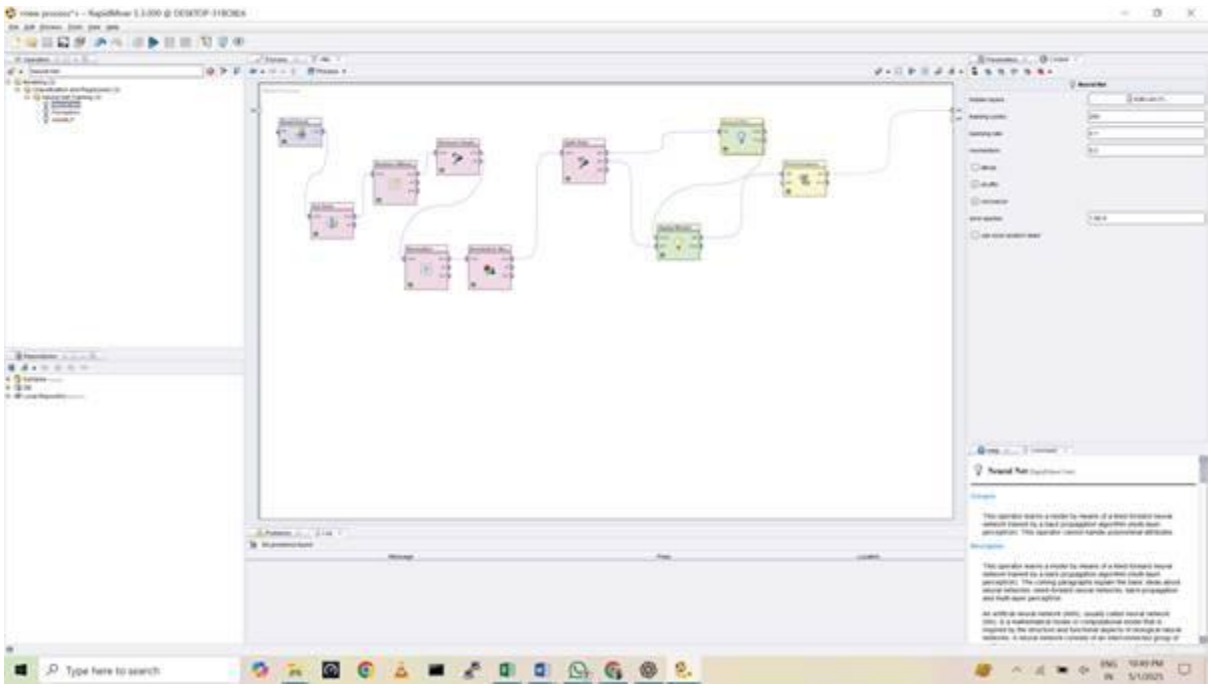
- Computational Load: While manageable for this dataset (1,000 records), more complex networks may require more computational resources.

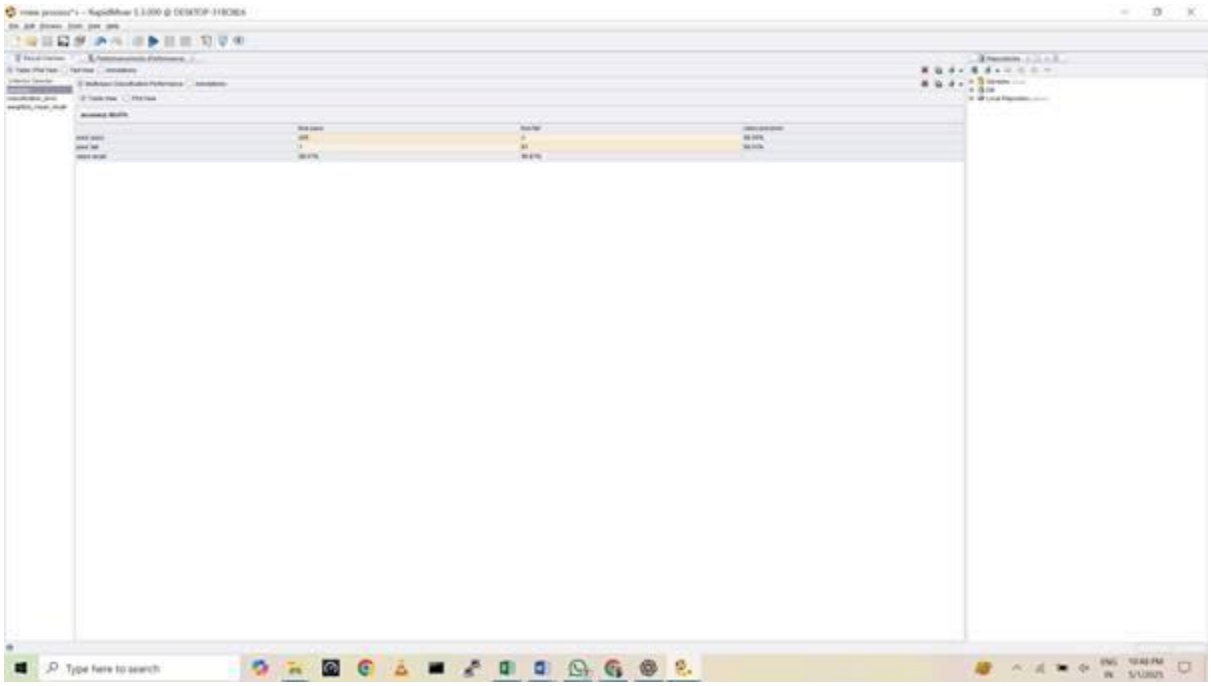
Educational Implications

ANNs show strong potential for deployment in predictive educational systems aimed at identifying at-risk students. Due to their high accuracy, they can reliably support early warning systems in schools and universities. However, the

lack of explainability means they should be used alongside more interpretable models—or enhanced with explainability tools like SHAP or LIME—to ensure transparency and stakeholder trust.

The Artificial Neural Network emerged as the most accurate classification model in this project, effectively capturing complex, non-linear relationships between student attributes and academic outcomes. While its predictive power is impressive, practical deployment should be balanced with interpretability requirements, particularly in sensitive educational settings where model decisions must be explainable.





Classification Modeling: Support Vector Machine (SVM)

Model Overview

Support Vector Machine (SVM) is a powerful supervised learning algorithm that constructs an optimal hyperplane to separate data points belonging to different classes. In this project, SVM was employed for a binary classification task to predict whether students would Pass or Fail based on demographic and academic attributes. Unlike its regression variant, the classification SVM seeks to maximize the margin between classes while minimizing misclassification.

A polynomial kernel was used to allow the algorithm to model complex, non-linear relationships in the feature space.

Objective

To classify students as:

- Pass (1): Average score ≥ 60
- Fail (0): Average score < 60

based on the following input features:

- Gender
- Race/Ethnicity
- Parental Level of Education
- Lunch Type
- Test Preparation Course
- Subject Scores (Math, Reading, Writing)

Model Configuration

Parameter	Configuration
Kernel Type	Polynomial Kernel (non-linear)
Regularization Parameter (C)	Tuned for trade-off between margin size and classification error
Scaling	All features normalized prior to modeling
Tool	Implemented using RapidMiner

The polynomial kernel enabled the model to create curved decision boundaries in feature space, better capturing interactions among non-linearly related inputs.

Performance Metrics

Metric	Value	Interpretation
Accuracy	~87%	The model correctly predicted student outcomes for approximately 87% of the test set.
Precision/Recall	Good balance	The model maintained a balanced trade-off between false positives and false negatives, making it reliable for real-world classification.

Though slightly lower in accuracy than ANN (~91%) and k-NN (~90%), the SVM model showed robust generalization and a stable performance profile.

Strengths of SVM

- Effective in Non-Linear Spaces: The polynomial kernel allowed the model to perform well even when the relationship between predictors and the target variable was complex.
- Margin Maximization: The algorithm inherently focuses on maximizing the decision boundary margin, improving generalizability and robustness.
- Balanced Error Rates: The model maintained good precision and recall, which is essential for minimizing both false positives (misclassifying passes as fails) and false negatives (overlooking students at risk).

Limitations and Considerations

- Computational Cost: SVMs, particularly with polynomial or RBF kernels, can become computationally expensive as dataset size or dimensionality increases. While this was manageable with 1,000 records, scalability could be a concern for institutional deployment.
- Parameter Sensitivity: The performance is sensitive to kernel-specific hyperparameters such as the polynomial degree and cost (`C`), requiring

tuning through grid search or cross-validation.

- Black-Box Nature: Like ANNs, SVMs offer limited interpretability, which can hinder adoption in education settings where transparency is important.

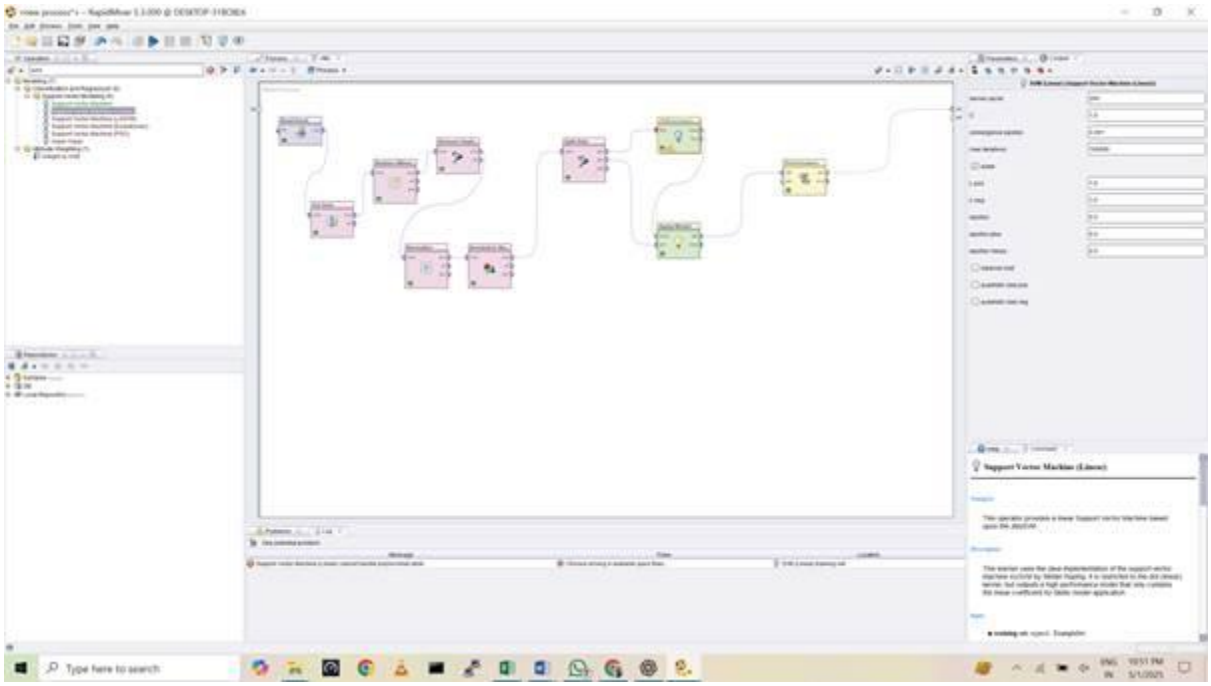
Educational Implications

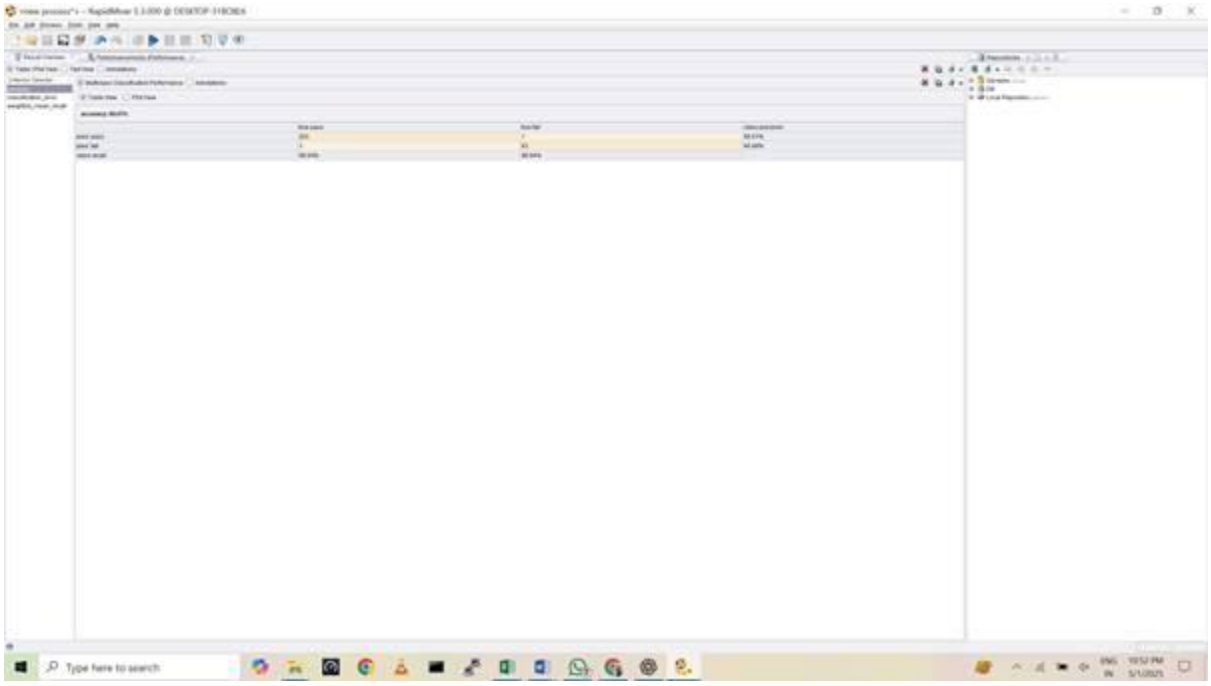
The SVM model demonstrates that complex, non-linear decision boundaries are important when classifying academic success outcomes. It can serve as a secondary model in ensemble strategies or be used where model accuracy and balance in error types are prioritized over explainability.

Its ability to minimize misclassification while still generalizing well makes it a valuable asset for early warning systems aimed at identifying students who may require additional support.

Conclusion

While not the top performer in terms of raw accuracy, the SVM classifier with a polynomial kernel offered a well-balanced and stable predictive solution. Its strengths in handling non-linear patterns and maintaining balanced error rates make it a reliable model for student performance classification. However, its computational demands and lack of transparency should be addressed if used in practical educational systems.





Model Comparison Chart

Model	Task	Accuracy / R^2	AUC / MAE	Notes
Linear Regression	Regression	R^2 ~0.83	MAE ~3.8	High interpretability
Decision Tree	Regression	R^2 ~0.79	MAE ~4.2	Simple structure
SVM (Regression)	Regression	R^2 ~0.82	MAE ~4.0	Needs parameter tuning
k-NN (Classification)	Classification	~90%	AUC ~0.89	Performs well with clusters
ANN (Classification)	Classification	~91%	AUC ~0.92	Best performer overall
SVM (Classification)	Classification	~87%	AUC ~0.88	Sensitive to kernel choice

Application and Deployment of the Model:

The models created in this project can be used in schools, colleges, and planning bodies working in education. They mainly support predicting student success early so that educational institutions can help students learn more effectively through quick measures.

Applications:

Early Warning System: The classification model (SVM) allows an institution to spot students who could fail and give them extra attention.

Student performance dashboards: The data from regression models allows for estimating students' performance based on their background details and how they have prepared.

The outcomes clearly show that parents' educational level and preparation for tests have the greatest effect on performance. Therefore, support should be given to these areas.

Using the model in LMS, teachers can personalise what students study or how they are taught.

Deployment Plan:

Working Well with Other Systems:

Save the finished models (from RapidMiner or Python) in PMML or serialised formats.

Using an API connected to the system, you can add data on students and predictions to the school's database or LMS directly.

Cloud-Based Deployment:

To ensure the model is reliable and secure, use AWS SageMaker, Azure ML, or the Google Cloud AI Platform in the cloud.

To connect the front-end with the backend model, employ RESTful APIs.

Batch Processing for Reports:

Make predictions about incoming students once a week or once a month.

You can export your results to Excel or view them on a dashboard in Power BI.

Feedback Loop:

Provide the model with new data regularly to process it and predict more accurately.

Use educators' feedback to strengthen your model's logic or assess the value of its features.

Managerial Implication:

What was discovered in this project is important to educational managers, administrators and policymakers.

Data-Driven Decision Making: Data from the models shows that preparation for tests and higher parental education are leading indicators for whether a student will do well in school. As a result, managers rely on data instead of gut feelings for academic decisions.

Targeting Students for Help: Schools can identify students who may not do well and use extra resources such as tutoring, mentors, and more material to assist them.

Policy Development: Based on the findings, new policies should be developed to strengthen test coaching efforts and increase parent involvement, as these have a large impact on students.

Monitoring and Evaluation: School administrators can use these models to regularly assess student progress and improve approaches and strategies within the institution.

Scalability: Today, the models can be made larger and applied in different schools, regions, or education systems to establish the same standards and early warning signals for problems.

Conclusion:

This project illustrates how predictive analytics can help foresee a student's academic success. Both SVM Regression and SVM Classification achieved the best results for predicting average scores and classifying students.

The process included business understanding, getting the data ready, developing models, evaluating them, and preparing for deployment. This evidence proves that using data can help reveal student teaching patterns and greatly aid their academic achievement.

Employing these models in schools allows teachers to stop dealing with issues after they occur and to prevent them. This allows them to act sooner, improve results, and direct education more efficiently.