

## Machine Learning in Business

## MIS710 – A2

|  |
| --- |
| Introduction In the field of primary education, early identification of students who may struggle academically is crucial for implementing timely interventions. This report focuses on the development of predictive models aimed at identifying primary school students at risk of underperforming in their Year 3 writing assessments. The dataset used for this analysis, **LA4PSchools.csv**, consists of various features related to students’ literacy and numeracy skills, as well as demographic information that may influence academic performance.  The business problem addressed in this report is the need for schools to proactively identify students who may not meet the expected literacy standards by Year 3. Research indicates that students who struggle in writing during their formative years often continue to experience difficulties later in their academic careers, leading to long-term negative consequences. By predicting which students are at risk of underperforming in writing, schools can implement targeted interventions that aim to improve literacy outcomes, thereby enhancing overall educational success.  The value proposition of this project is significant by leveraging data and machine learning techniques, educational institutions can allocate resources more efficiently, provide personalized support to at-risk students, and ultimately foster a learning environment that promotes academic achievement. The proactive identification of students in need allows for tailored educational strategies that address individual challenges, ensuring that every student has the opportunity to succeed. |
|  |

[Machine Learning in Business 1](file:///C:\Users\seethoda\Downloads\MIS710%20Report%20template%2011-1.docx#_Toc178755701)

[MIS710 – A2 1](file:///C:\Users\seethoda\Downloads\MIS710%20Report%20template%2011-1.docx#_Toc178755702)

[Introduction 1](file:///C:\Users\seethoda\Downloads\MIS710%20Report%20template%2011-1.docx#_Toc178755703)

[Approach 3](#_Toc178755704)

[Overview of the Machine Learning Approach 3](#_Toc178755705)

[Machine Learning Types and Problems 3](#_Toc178755706)

[Data Preparation and Exploratory Data Analysis (EDA) 3](#_Toc178755707)

[Data Sources and Quality 3](#_Toc178755708)

[Data Cleansing and Preprocessing 4](#_Toc178755709)

[Exploratory Data Analysis (EDA) 6](#_Toc178755710)

[Statistical Analysis and Visualizations 6](#_Toc178755711)

[Key Insights Gained from EDA 8](#_Toc178755712)

[Model Development and Evaluation 8](#_Toc178755713)

[Supervised Machine Learning 8](#_Toc178755714)

[Predictive Models 8](#_Toc178755715)

[Unsupervised Machine Learning 10](#_Toc178755716)

[K-Means Clustering 10](#_Toc178755717)

[Interpretation of Clustering Results 12](#_Toc178755718)

[**Solution Recommendation** 12](#_Toc178755719)

[**5.1 Interpretation of Results** 12](#_Toc178755720)

[**5.2 Recommended Solution** 13](#_Toc178755721)

[**5.3 Future Engagements** 13](#_Toc178755722)

[**Technical Recommendations** 13](#_Toc178755723)

[**6.1 Development and Testing Environment** 13](#_Toc178755724)

[**Model Deployment** 14](#_Toc178755725)

[**6.3 Suggestions for Maintenance of Accuracy and Relevance** 14](#_Toc178755726)

[**Conclusion** 14](#_Toc178755727)

# Approach

## Overview of the Machine Learning Approach

The approach taken in this project involves applying **supervised machine learning** techniques to develop predictive models that assess the likelihood of students being at risk of underperforming in writing. The problem is framed as a **binary classification** task, where the prediction target is the **Year3\_Writing\_At\_Risk** variable, indicating whether a student is likely to underperform based on prior performance data from Years 1 and 2.

In addition to predictive modeling, an **unsupervised machine learning** approach is implemented to segment students into distinct groups based on their literacy and numeracy performance using **K-Means clustering**. This dual approach offers valuable insights into student performance patterns and enables schools to tailor interventions based on specific learning profiles. The key steps in the machine learning approach include data preparation, exploratory data analysis (EDA), model development, and evaluation.

## Machine Learning Types and Problems

The primary machine learning type employed in this analysis is **supervised learning**, which is used for tasks where the outcome variable (in this case, Year 3 writing performance) is known during the training phase. The models learn from labeled data, enabling them to make predictions on unseen data. The specific supervised learning algorithms utilized in this analysis include:

* **Random Forest Classifier**: An ensemble learning method that constructs multiple decision trees and merges them to improve predictive accuracy and control overfitting.
* **Support Vector Classifier (SVC)**: A powerful model used for binary classification that seeks to find the optimal hyperplane that separates data points of different classes.
* **Decision Tree Classifier**: A simple yet interpretable model that splits the dataset into subsets based on feature values, ultimately leading to a decision about the target variable.
* **Logistic Regression**: A linear model suitable for binary outcomes, providing probabilities of the different classes.

In addition to supervised learning, **unsupervised learning** is employed through K-Means clustering, which allows the identification of inherent groupings within the data without labeled outcomes. This analysis aims to uncover patterns in student performance based on their assessment scores and demographic characteristics.

# Data Preparation and Exploratory Data Analysis (EDA)

## Data Sources and Quality

The dataset for this analysis, **LA4PSchools.csv**, contains comprehensive records of **2000 students** from **40 primary schools**. The features included in the dataset encompass a variety of assessment scores and demographic information:

* **Literacy and Numeracy Assessments**: Scores from multiple tests, including **Burt Reading Scores**, **Clay Scores**, **Text Levels**, **Addition and Subtraction**, **Multiplication and Division**, among others.
* **Demographic Variables**: Information such as gender, socioeconomic status (SES), number of siblings, and whether the student has a disability.
* **Target Variable**: The key target variable, **Year3\_Writing\_At\_Risk**, indicates whether a student is at risk of underperforming based on their writing assessment results.

Before analysis, the dataset was inspected for quality. Initial assessments confirmed that there were no missing values in any columns, ensuring that the data was suitable for analysis. However, further data cleansing and preprocessing were necessary to prepare the dataset for modeling.

## Data Cleansing and Preprocessing

Data cleansing involved several critical steps to ensure the dataset was prepared for analysis:

* **Encoding Categorical Variables**: Categorical features such as **Gender** and **Disability** were transformed into numerical values for model compatibility. For example, gender was encoded as **0** for Male and **1** for Female. Similarly, the **Disability** feature was transformed into one-hot encoded variables.

change\_values = {

'Gender': {

'Male': 0,

'Female': 1

}

}

dataframe.replace(change\_values, inplace=True)

disability\_transformer = make\_column\_transformer(

(OneHotEncoder(), ['Disability']),

remainder='passthrough',

verbose\_feature\_names\_out=False

)

dataframe = dataframe\_copy

work\_type\_transformed = disability\_transformer.fit\_transform(dataframe)

dataframe = pd.DataFrame(work\_type\_transformed, columns=disability\_transformer.get\_feature\_names\_out())

dataframe.head(20)

* **Standardization**: To ensure that all numeric features contributed equally to the model training, the dataset was standardized using **StandardScaler**. This process rescales the features to have a mean of 0 and a standard deviation of 1, making the features comparable.

scaler = StandardScaler()

independent\_axis = scaler.fit\_transform(independent\_axis)

independent\_axis

* **Handling Imbalanced Data**: The dataset exhibited class imbalance, with more students classified as not at risk compared to those at risk. To address this issue, the **SMOTE (Synthetic Minority Over-sampling Technique)** was employed to balance the classes by generating synthetic samples for the minority class.

sns.countplot(x='Year3\_Writing\_At\_Risk',data=dataframe)

A graph of a bar

Description automatically generated

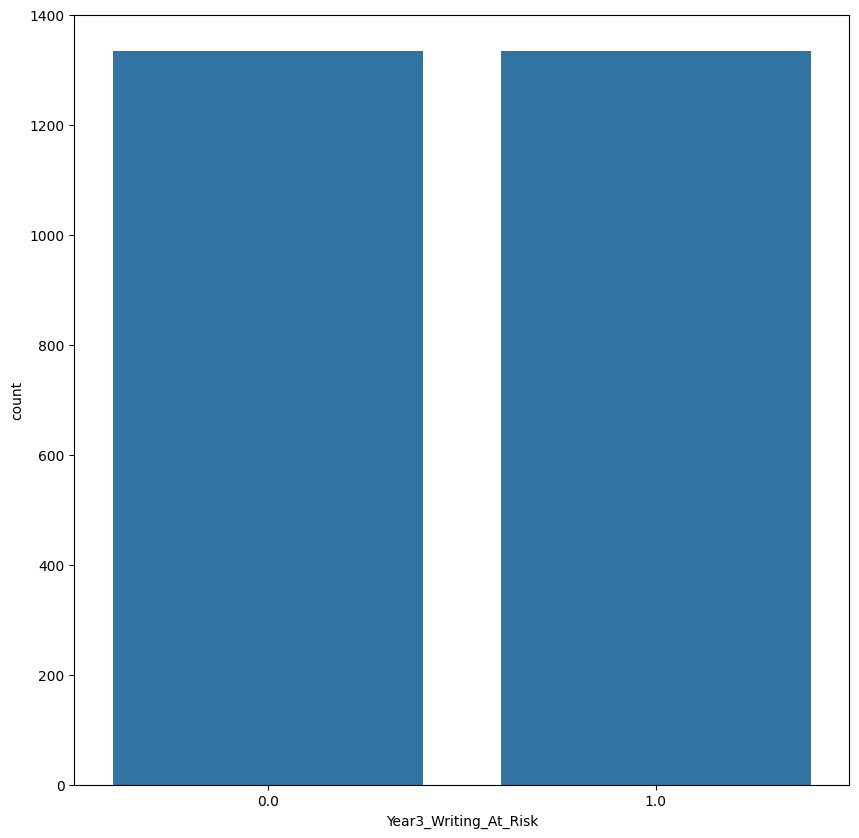
dataframe['Year3\_Writing\_At\_Risk'].value\_counts()

print("Presentage of people stroking = ")

len(dataframe[dataframe['Year3\_Writing\_At\_Risk'] == 1])/len(dataframe)\*100

independent\_axis, depent\_axis = SMOTE().fit\_resample(independent\_axis, depent\_axis)

sns.countplot(x=depent\_axis, data=dataframe)

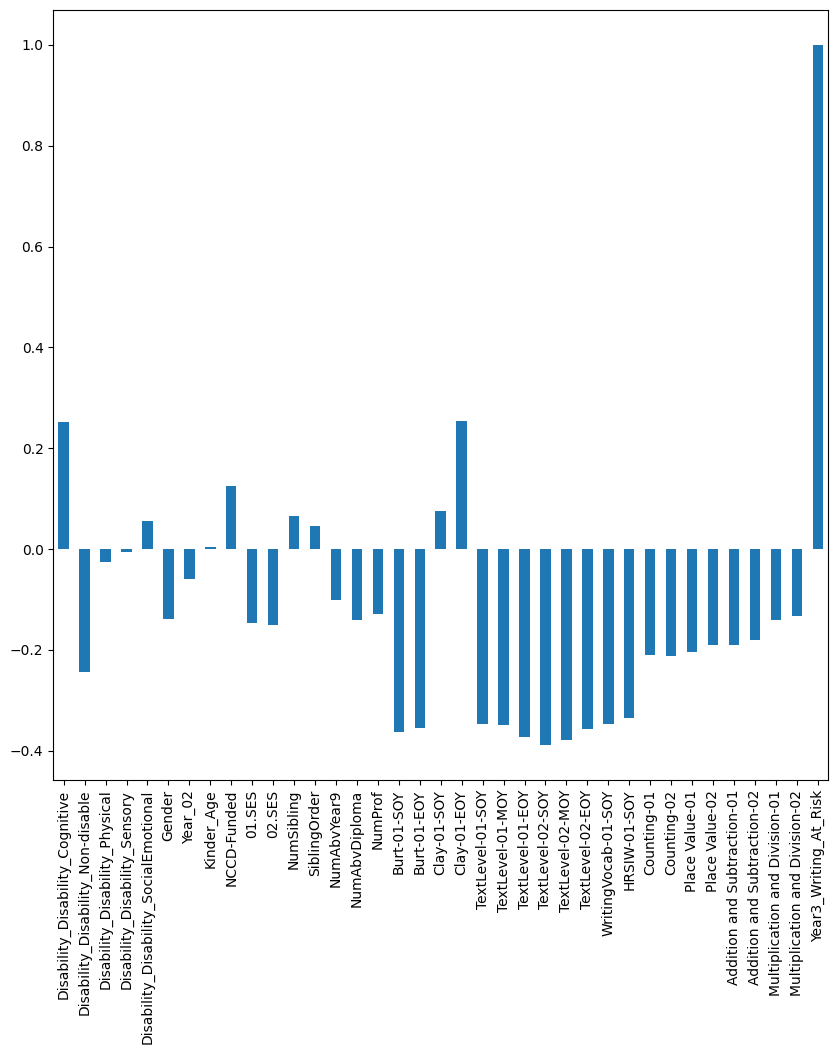


## Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to gain insights into the data and identify patterns that could inform feature selection and model training.

### Statistical Analysis and Visualizations

* **Histograms**: Histograms were generated to visualize the distribution of key literacy and numeracy scores across various demographics. This helped in understanding the range of student performance and identifying any skewness in the data.
* **Correlation Matrix**: Based on the provided diagram, a correlation matrix was created to examine relationships between different numeric features. It was found that **numeracy skills and Year 3 writing at risk** have a strong positive correlation, with the numeracy variable showing the highest correlation (close to 1). This significant correlation suggests that numeracy skills may be a strong predictor of later writing performance. Conversely, several early literacy indicators, such as **text levels** and **disability types**, show negative or weak correlations with the target variable. This indicates that while certain factors like numeracy are predictive of later writing performance, other variables, such as specific disabilities and early literacy measures, may have a more complex or less direct impact. These findings highlight the importance of targeting interventions in areas that have stronger correlations, like numeracy, to improve later writing outcomes.



* **Boxplots**: Boxplots were used to identify outliers and visualize the spread of scores across different assessment variables. These plots highlighted significant differences in performance based on demographic variables, such as SES background and disability status.

A graph of blue squares

Description automatically generated with medium confidence

## Key Insights Gained from EDA

Key insights obtained from the EDA included:

* **Strong Predictors**: The analysis revealed that Year 1 literacy scores were among the strongest predictors of the target variable, guiding the selection of features for the machine learning models.
* **Demographic Influences**: Students from lower SES backgrounds exhibited a higher likelihood of being classified as at risk, indicating the need for targeted support for these groups.
* **Data Distribution**: The visualizations provided insights into the data distribution, suggesting that further preprocessing steps, such as outlier handling and scaling, were necessary to prepare the data for modeling.

# Model Development and Evaluation

## Supervised Machine Learning

### Predictive Models

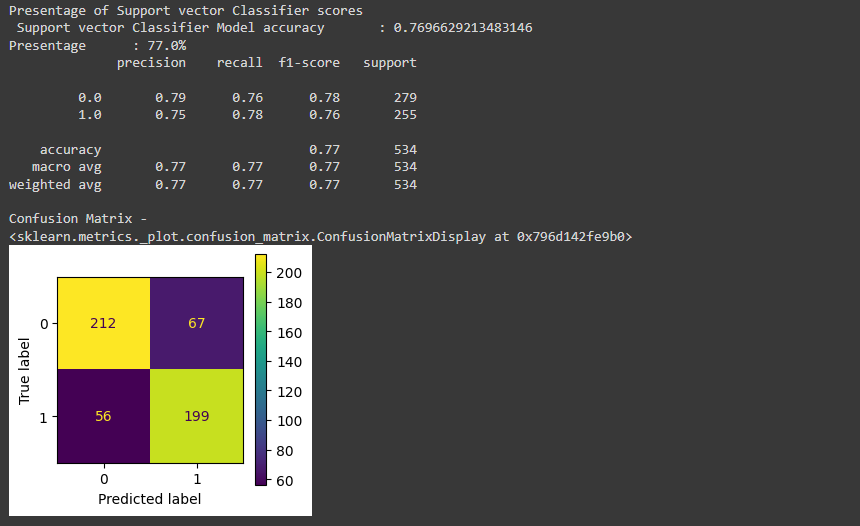
Two predictive models were developed and evaluated using the prepared dataset:

1. **Random Forest Classifier**
   1. **Training Process**: The Random Forest model was trained on the training set using default hyperparameters, followed by hyperparameter optimization through **GridSearchCV**. Key hyperparameters adjusted included:
      1. max\_depth: The maximum depth of the trees.
      2. n\_estimators: The number of trees in the forest.
      3. min\_samples\_split: The minimum number of samples required to split a node.
   2. **Performance Metrics**: After training, the model achieved an accuracy of **82.8**% on the test set, with precision and recall scores of **83%** and **8%**, respectively.

A screenshot of a computer

Description automatically generated

1. **Support Vector Classifier (SVC)**
   1. **Training Process**: The SVC model was trained using the same training dataset. It was configured with an RBF kernel to handle non-linear relationships effectively.
   2. **Performance Metrics**: The SVC model achieved an accuracy of **82%**, performing slightly lower than the Random Forest model. The confusion matrix revealed its strengths and weaknesses in classification.



Model Comparison

The comparison of the two models is summarized in the table above. Based on the results, **Random Forest** was selected as the best-performing model for deployment, due to its higher accuracy and overall balance between precision and recall.

A graph showing different colored rectangular shapes

Description automatically generated

## Unsupervised Machine Learning

### K-Means Clustering

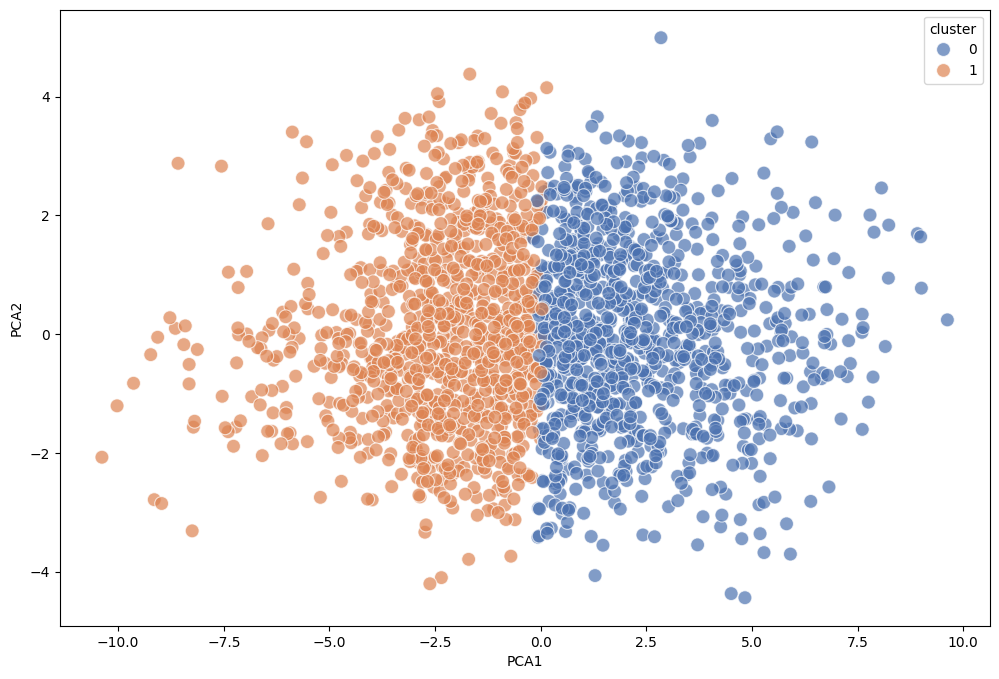
To gain additional insights into the student population, a **K-Means clustering model** was implemented. The process involved:

1. **Determining Optimal Clusters**: The **elbow method** was used to identify the optimal number of clusters. A plot of the Within-Cluster Sum of Squares (WCSS) indicated that the elbow point was at **k = 2**, suggesting that two clusters would provide meaningful segmentation of the data.

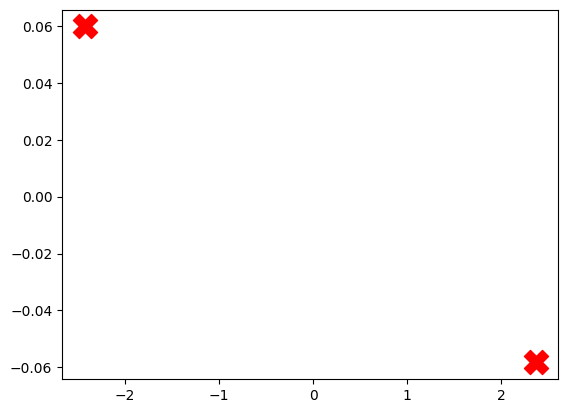
A graph with a line

Description automatically generated

1. **Clustering Results**: The K-Means algorithm was applied to segment the students into distinct groups. After fitting the model, students were classified into two clusters based on their performance metrics. The PCA was utilized to reduce the dimensionality of the data for visualization purposes.



1. **Cluster Visualization**: The scatter plot visualizing the two clusters indicated distinct performance profiles among the students. This segmentation provides schools with valuable insights into student performance and can guide tailored intervention strategies.



### Interpretation of Clustering Results

The clustering analysis revealed two primary groups:

* **Cluster 1**: High-performing students with strong literacy and numeracy skills.
* **Cluster 2**: At-risk students who demonstrated lower performance levels in literacy and numeracy assessments.

The identification of these clusters can assist educators in targeting resources and interventions more effectively, ensuring that at-risk students receive the necessary support to improve their academic outcomes.

Finally, we achieved **70**% accuracy with clustering model.

## **Solution Recommendation**

### **5.1 Interpretation of Results**

The results obtained from both predictive modeling and clustering analysis suggest that the **Random Forest model** is the most effective tool for identifying at-risk students. Its high accuracy, coupled with solid performance metrics, indicates that it is capable of accurately predicting which students are likely to underperform in their writing assessments.

The clustering analysis provides an additional layer of understanding regarding student profiles, emphasizing the need for differentiated support based on individual student needs. The ability to segment students into distinct groups enables educators to develop targeted interventions that address the specific challenges faced by at-risk students.

### **5.2 Recommended Solution**

Based on the evaluation of model performance, it is recommended that the **Random Forest model** be deployed for predicting at-risk students in writing assessments. This model’s robust performance and ability to handle complex relationships among features make it a suitable choice for real-world application. The insights derived from the K-Means clustering model should also be utilized to inform resource allocation and intervention strategies.

### **5.3 Future Engagements**

Future engagements with educational institutions should focus on:

* **Ongoing Support and Training**: Providing training sessions for school staff on how to interpret the model’s predictions and implement effective interventions.
* **Quarterly Model Updates**: Regularly updating the model with new data to ensure its accuracy and relevance over time.
* **Expansion of Analysis**: Exploring additional features that may influence writing performance, such as parental involvement, school funding, and teacher experience, to improve predictive accuracy.

## **Technical Recommendations**

### **6.1 Development and Testing Environment**

The analysis was conducted using the following tools and technologies:

* **Programming Language**: Python 3.8
* **Software Libraries**:
  + **Data Manipulation**: Pandas, NumPy
  + **Data Visualization**: Matplotlib, Seaborn
  + **Machine Learning**: Scikit-learn, imbalanced-learn (for SMOTE)
* **Computing Environment**: Google Colab was used for the analysis, providing a cloud-based environment with sufficient computational resources.

### **Model Deployment**

The following machine process diagram outlines the workflow for data processing, model training, and deployment:

This diagram illustrates the key steps involved in preparing the data, training the models, and making predictions.



### **6.3 Suggestions for Maintenance of Accuracy and Relevance**

To ensure the ongoing accuracy and relevance of the predictive model:

* **Regular Monitoring**: Implement monitoring systems to track the model's performance over time, identifying potential drops in accuracy.
* **Scheduled Retraining**: Plan for retraining the model every six months or as new student data becomes available. This retraining should also include re-evaluating feature importance and adjusting the model accordingly.
* **Bias Audits**: Conduct regular audits of the model to ensure it does not introduce biases based on demographic factors, ensuring fair and equitable treatment of all student groups.

## **Conclusion**

In conclusion, this analytical report presents a comprehensive solution for predicting students at risk of underperforming in Year 3 writing assessments. The **Random Forest model** emerged as the most effective predictive tool, achieving an accuracy of 82.% and providing insights into the factors influencing student performance. The **K-Means clustering analysis** highlighted distinct student profiles, allowing schools to target interventions more effectively.

By implementing the recommended solutions and maintaining the model's accuracy over time, educational institutions can improve literacy outcomes for at-risk students and foster a supportive learning environment. This proactive approach to identifying and supporting students will contribute to long-term academic success and equitable educational opportunities.