Curriculum Optimization for Ontario Public Schools

By: Teamar Samison and Chiedza Magumbe

Problem Statement

Elementary schools across the province must ensure that their students are able to meet the required provincial reading standard

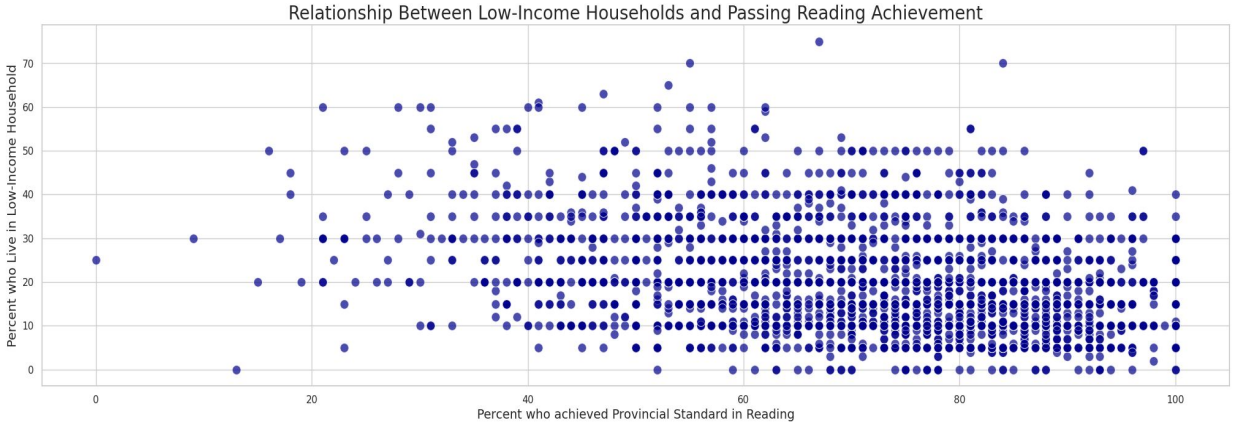
The data analysis will determine the primary challenges that prevent students from succeeding and predict which attributes improve reading scores.

By identifying these components, the provincial government can launch targeted programs to strategically support elementary schools across Ontario.

Proposed Solution:

Building Predictive Model

- **Goal:** Use machine learning to identify schools where students may underperform due to socioeconomic challenges.
- **Insight:** Schools with more low-income students tend to have lower reading achievement scores.
- **Next Steps:** Explore additional socioeconomic factors to create a model for targeted interventions to improve reading scores.



Methodology:
Logistic Regression and KNN

The project applied different classification models, to achieve several key outcomes:

- **Accurate Student Performance Prediction:** Schools can identify students at risk of failing their grade 3 reading test based on key factors such as enrollment, parental education, and socioeconomic status.
- **Targeted Interventions:** Educational resources can be allocated more effectively to schools with students that have higher predicted rates of failure, allowing for more support to students.
- **Data-Driven Decision Making:** Insights from the model can inform policy adjustments and strategies aimed at improving reading performance across different regions and demographics

Methodology:
Choosing our features

Independent Variables

- Enrolment
- Board Type
- Percentage of Students Whose First Language Is Not English
- Percentage of Students Whose First Language Is Not French
- Percentage of Students Who Are New to Canada from a Non-English Speaking Country
- Percentage of Students Receiving Special Education Services
- Percentage of Students Identified as Gifted
- Percentage of School-Aged Children Who Live in Low-Income Households
- Percentage of Students Whose Parents Have No Degree Diploma or Certificate

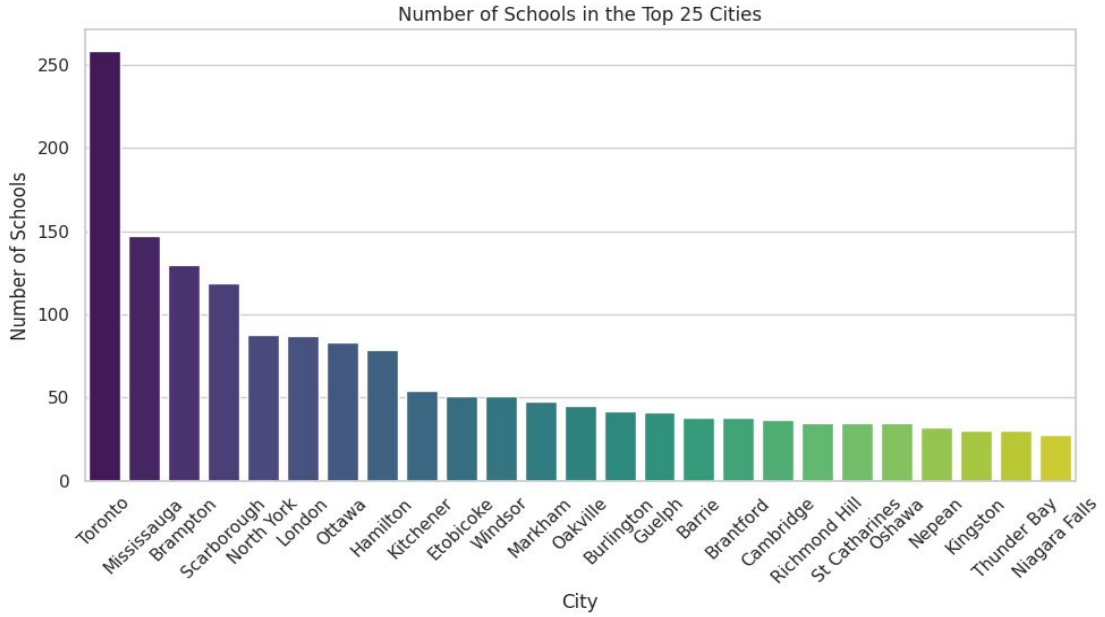
Target Variable

- Percentage of Grade 3 Students Achieving the Provincial Standard in Reading

These variables were chosen to understand how they influence the reading performance of Grade 3 students

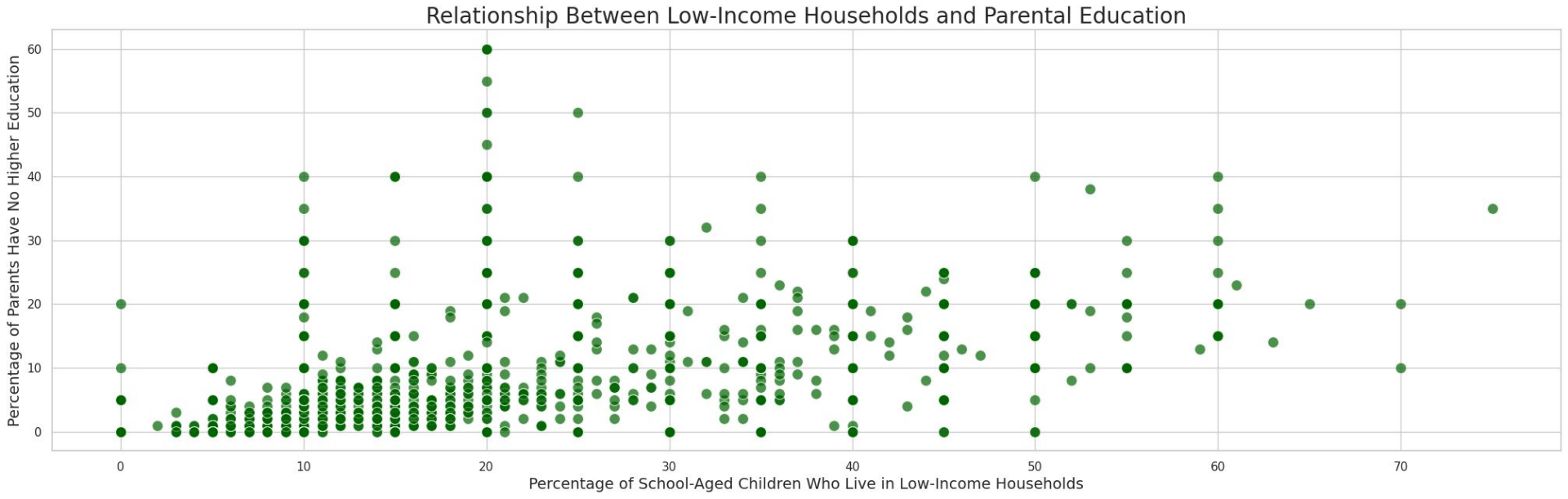
Data Understandi

- **Dataset Overview:** Our dataset contains information about schools located in Ontario
- **Key Insight:** The chart highlights the number of schools in the top 25 cities in our data set, showing a significant concentration in major urban areas.



Data Understandin

- This scatter plot showcases the relationship between low-income households and parental education levels across schools.
- Key Observation:
  - There appears to be a clustering of schools where a higher percentage of children from low-income households correlates with a lower percentage of parents with higher education.
  - This relationship may highlight the socioeconomic challenges affecting educational performance and how parental education might influence student outcomes.



Data Preprocessing:
Data Cleaning

Before moving forward with implementing our machine learning models data pre-processing and cleaning was required to convert the raw data into a cleaned dataset.

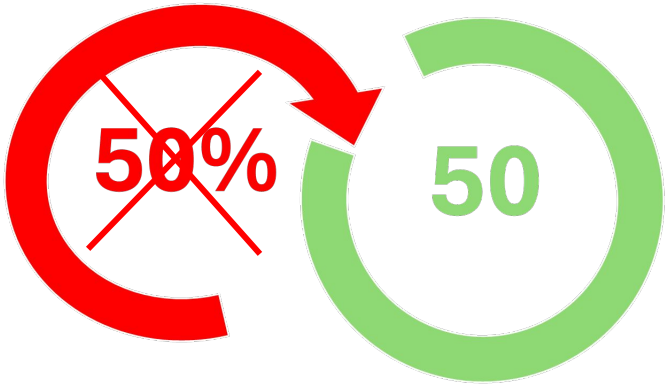
Changes Made:

- Dropped columns that were unrelated to the analysis
- Dropped null values
- Removed percent signs in variables of interest
- Converted numeric values from object to float

Data Preprocessing:
Data Cleaning

Cleaning Text Values

- Some of the data in our dataset was registering as text characters instead of numbers
- We had to perform further cleaning to remove characters such as percent signs
- We converting each individual item in a column to numbers for columns where we expected numbers to help generate our visualizations and create our models

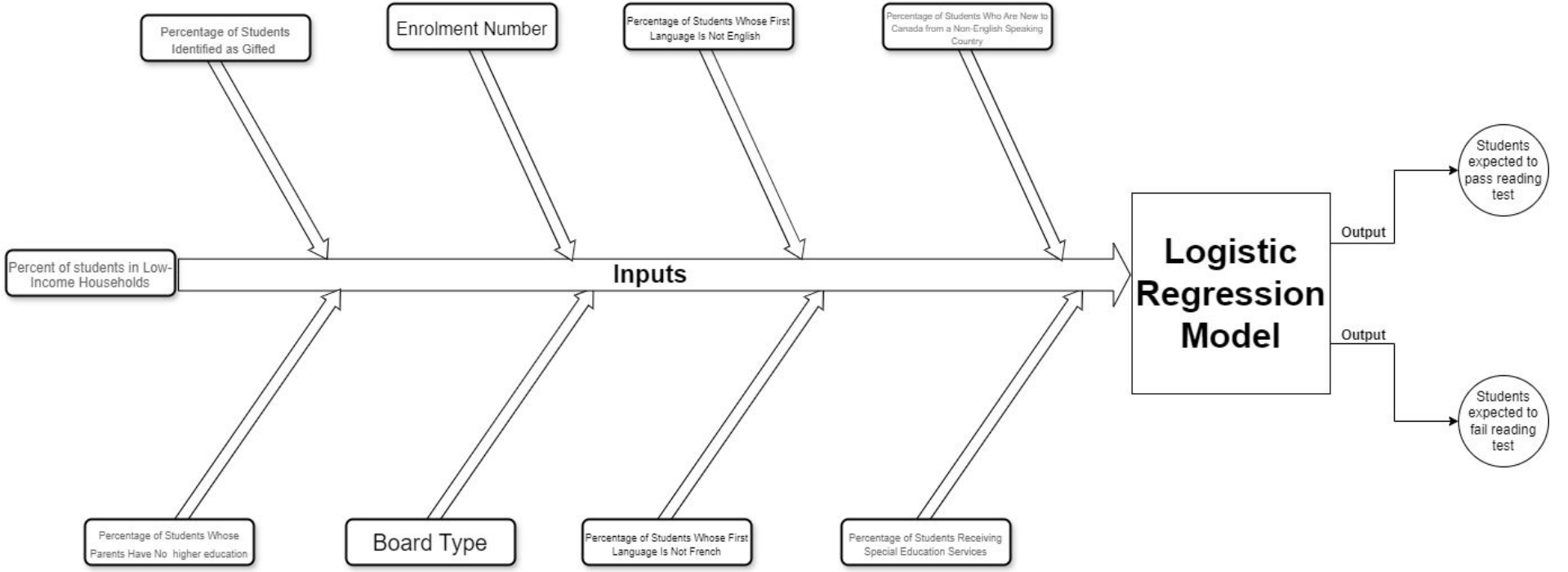


Model Evaluation: *Logistic Regression*

- We chose Logistic Regression as our first model due to its simplicity, interpretability, and effectiveness in binary classification tasks. To align with this approach, we converted the percentage-based reading scores into a binary format (pass = 1, fail = 0), as applying logistic regression directly to continuous percentage data produced poor results.
- Logistic regression allows us to clearly understand the relationship between our independent variables, such as school demographics and socio-economic factors, and the target variable—reading performance. Additionally, its support for regularization techniques (Lasso and Ridge) helps manage overfitting, making it suitable for datasets with high dimensionality or noise.

Model Evaluation:
Logistic Regression

Visualizing the Inputs and Outputs: This diagram shows how various socioeconomic and educational factors are fed into the logistic regression model to predict whether students will pass or fail their reading test.



Model Evaluation: *Logistic Regression Results*

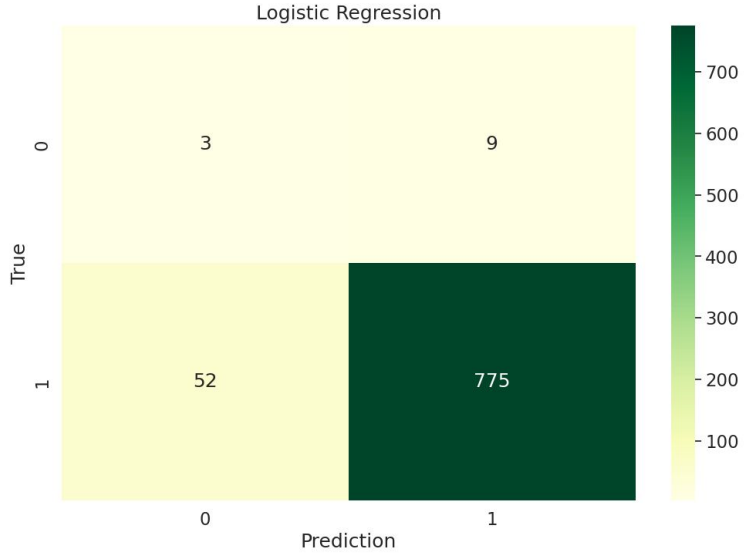
```
Accuarcy Score:  92.72943980929678
F1 Score:  96.2135319677219
Percsion Score:  93.71221281741234
Recall Score:  98.85204081632652
```

This shows how well our logistic regression model performed in predicting whether students will pass or fail their reading test. The confusion matrix breaks down the model's predictions:

- **0** represents students who are expected to **fail** the test.
- **1** represents students who are expected to **pass** the test.

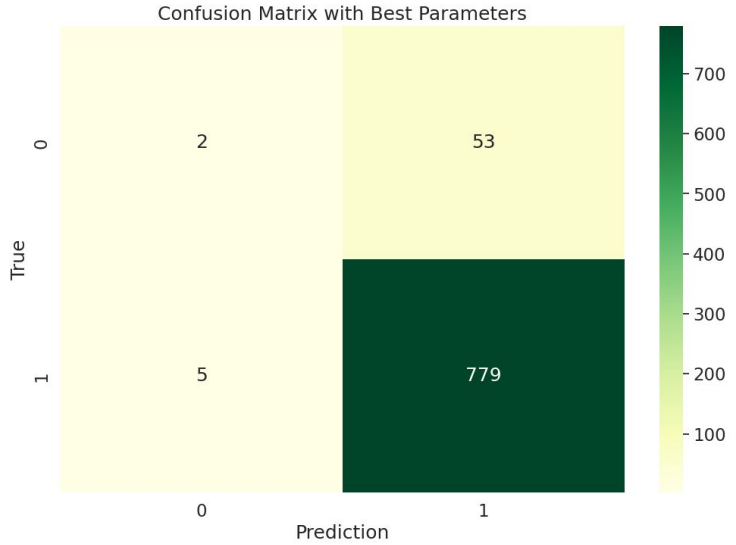
The true values are represented vertically, and the predicted values are represented horizontally.

- For example, we can see that **775 students** were correctly predicted to pass the test. The prediction value outputted by the model was 1(*a pass*) and the true value was 1.



Model Evaluation: *Logistic Regression Tuning*

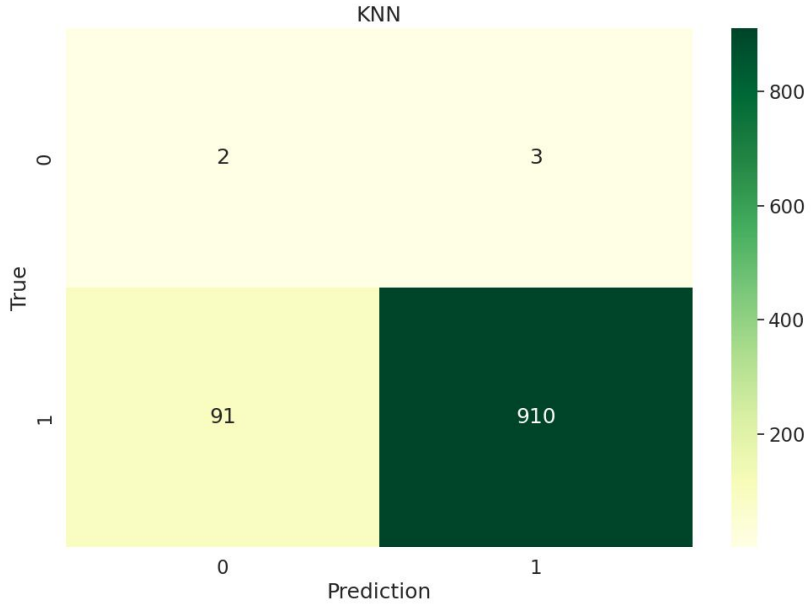
After applying Grid Search to tune the Logistic Regression model, we observed an increase in accuracy, F1 score, and recall. This improvement highlights the importance of hyperparameter tuning in enhancing the model's ability to correctly identify students who are likely to pass their reading test.



	Model	Accuracy	F1 Score	Precision	Recall
0	Logistic Regression	92.73	96.21	93.71	98.85
1	Logistic Regression with Best Parameters	93.08	96.41	93.63	99.36

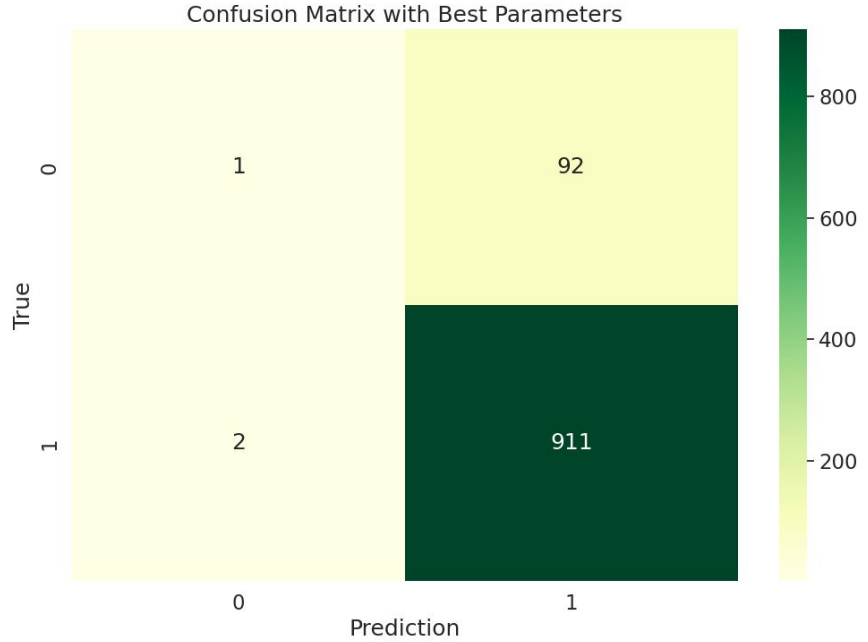
Model Evaluation: *K-Nearest Neighbours*

We chose K-Nearest Neighbours as our second model because it is effective on medium sized datasets and can provide further analysis of model performance.KNN will be able to classify new data points by analysing the labelled points closest to it, known as its nearest neighbours.



Model Evaluation: *K-Nearest Neighbours Tuning*

The tuned model provides more accurate predictions into whether grade 3 students in Ontario schools are likely to pass or fail their reading test by recognizing and classifying positive data points correctly.



	Model	Accuracy	F1 Score	Precision	Recall
0	KNN	90.36	95.09	90.91	99.67
1	KNN with Best Parameters	90.66	95.09	90.83	99.78

Best Model: *Logistic Regression*

We tested multiple algorithms, including Logistic Regression and KNN, to identify the best model. After tuning the Logistic Regression model with Grid Search, we achieved higher accuracy and recall scores, making it the most reliable option for predicting student reading outcomes.

Logistic Regression Results

	Model	Accuracy	F1 Score	Precision	Recall
0	Logistic Regression	92.73	96.21	93.71	98.85
1	Logistic Regression with Best Parameters	93.08	96.41	93.63	99.36

K-Nearest Neighbors Results

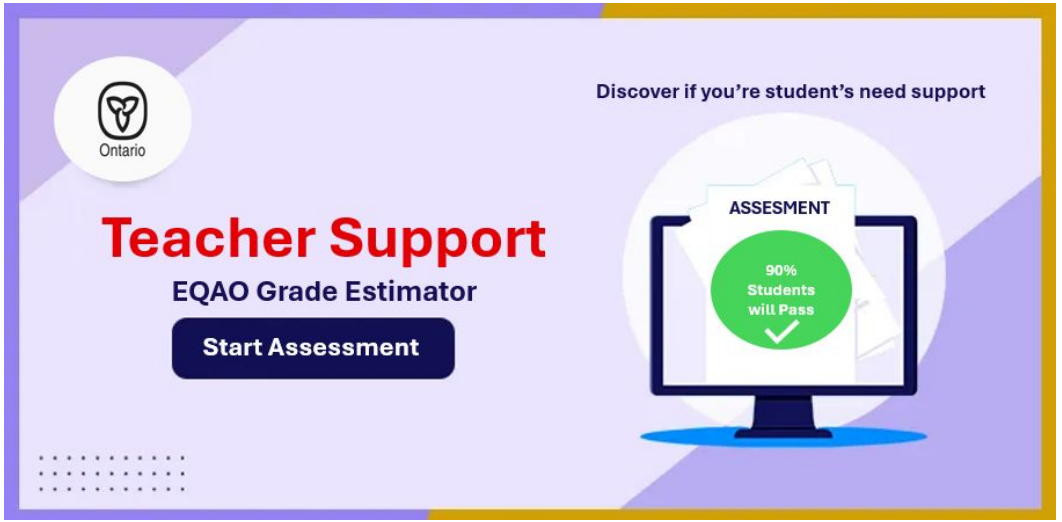
	Model	Accuracy	F1 Score	Precision	Recall
0	KNN	90.36	95.09	90.91	99.67
1	KNN with Best Parameters	90.66	95.09	90.83	99.78

Results

The insights gained from the model revealed which factors were most influential in predicting whether students are likely to pass or fail their reading test

Next steps

- **Build an Application:** Develop an application for teachers in Ontario to assess the risk of their students failing the EQAO reading test.
- **Further Model Refinement:** Build a linear regression model to predict the actual percentage of students expected to pass based on various factors
- **Resource Allocation:** Use the model's predictions to allocate additional learning resources to schools that are identified as high-risk for underperformance.



Conclusion

Overall, the model demonstrates strong potential for aiding schools in making informed decisions but can benefit from further refinement and expansion in data features to enhance its predictive power.

Limitations:

- Binary nature of the current outcome variable (pass/fail), which oversimplifies the nuance of educational performance
- Without data from previous years the model's ability to capture trends and improve accuracy is limited

The model's practical implications include the ability for schools and policymakers to make data-driven decisions, allowing for early interventions and resource allocation to support struggling students. This could improve educational outcomes by proactively addressing factors that influence reading performance.

By identifying and understanding these components, the provincial government can launch targeted programs to strategically support elementary schools across Ontario.