1. Data Overview and Cleaning

The project focuses on analysing and predicting chronic absenteeism among students. Data from the schools.csv file has already been cleaned beforehand. The logic is similar to the last task report. For the **Fluency** column, unknown values were replaced with the first valid value from matching **ANON_IDs**, and a similar approach was applied to **SED**. Student **Age** was calculated using the difference between their birth year and the end year of the school year. Missing values in **Susp** were filled with 0, while the **Gen** column was standardized by converting lowercase **"m"** to uppercase **"M."** Lastly, remaining missing values, including those in **AttRate, DaysEnr, and DaysAbs**, were filled with 0 to ensure consistency.

2. Data Preparation

To predict chronic absenteeism for the **2023-2024** school year, the dataset was filtered to include only students with chronic absenteeism status for that year. Data from previous years **(2017 to 2023)** was checked, and if available, it was added as new columns corresponding to each year. Numeric data such as **AttRate, DaysAbs, and DaysEnr** were renamed to reflect the time lag, e.g., **AttRate_1YearAgo** for the **2022-2023** school year, following a similar pattern for other years. To prevent data leakage, values for **AttRate, DaysAbs, DaysEnr, and Birthdate** from the current year **(2023-2024)** were excluded from the predictive features.

3. Model 1: Below Grade 6 Model to Predict 2023-2024 Chronic Absenteeism Status

To predict chronic absenteeism for students in grades below 6, the data was split into training and testing sets. The model utilized both categorical features such as **Fluency, Gen, SED, SpEd, Eth, and School,** which were processed using **One-Hot Encoding**, and numerical features such as **AttRate, DaysEnr, and DaysAbs** from past years, along with **Age and Susp,** which were standardized. A **Random Forest Classifier (rf1)** was trained on the prepared dataset, achieving a **test accuracy of 72.6%** and a **training accuracy of 82.5%**, indicating a good fit with minimal overfitting.

```
Test Accuracy: 0.7258064516129032

Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.74      0.74        98
           1       0.71      0.70      0.71        88

    accuracy                           0.73       186
   macro avg       0.73      0.72      0.72       186
weighted avg       0.73      0.73      0.73       186
```

Figure 1

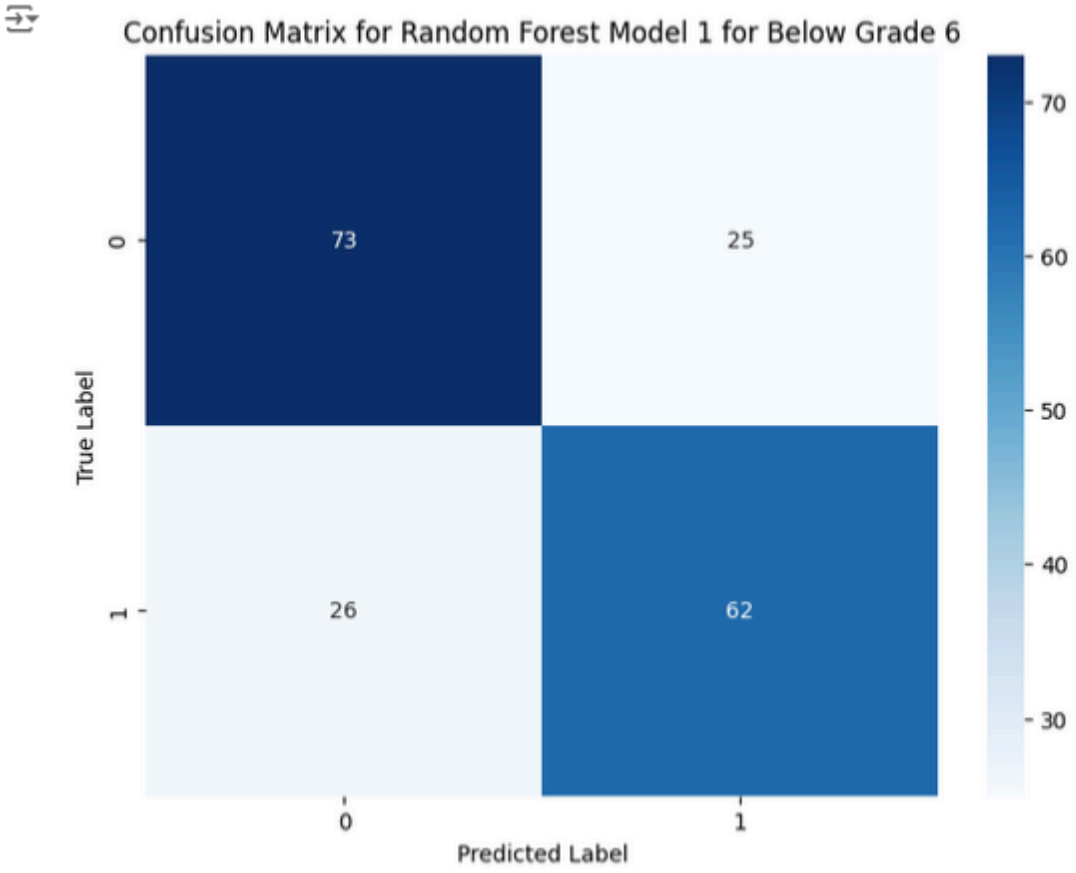

Figure 2

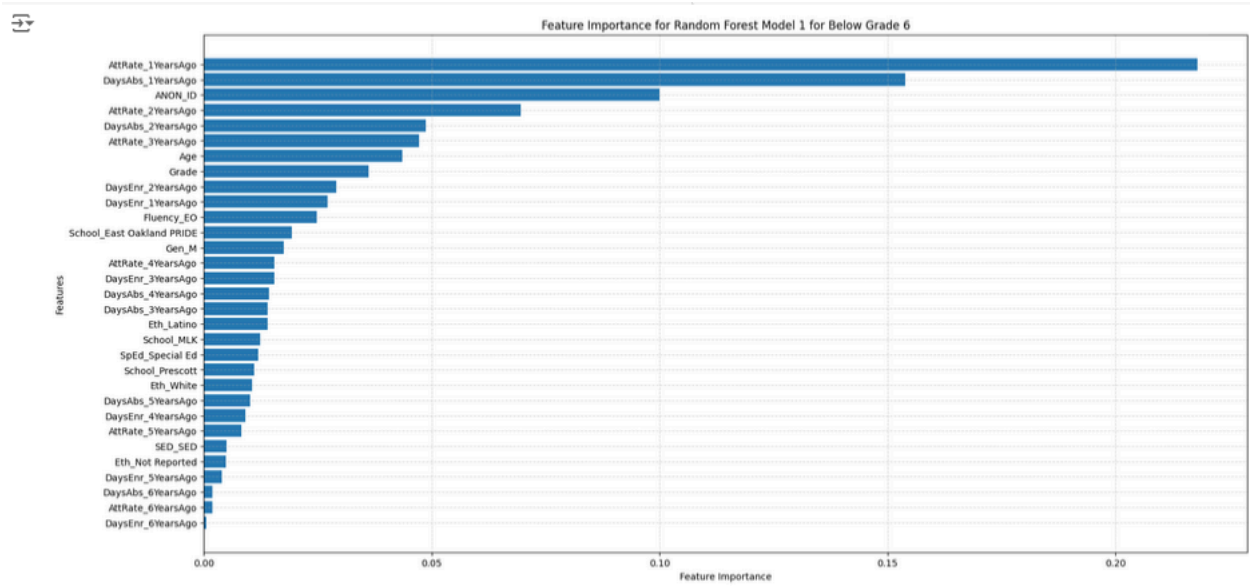

Figure 3

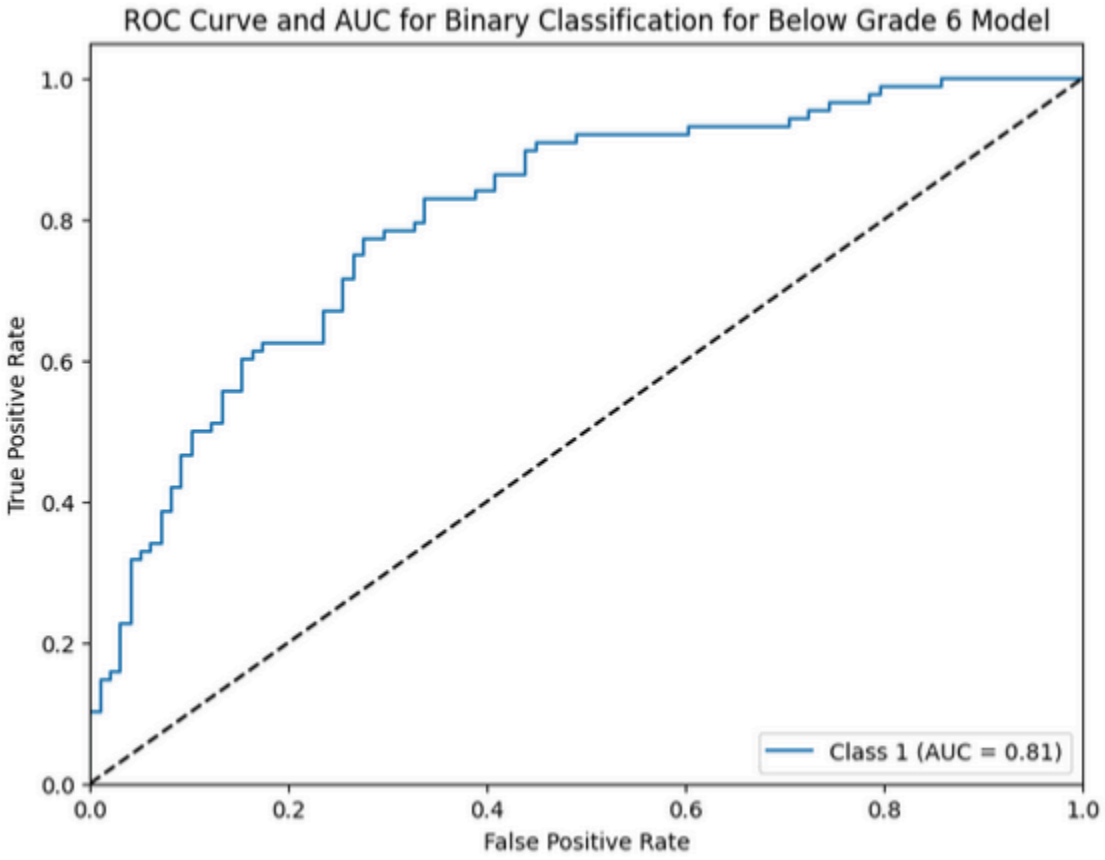

Figure 4

The classification report (*Figure 1*) highlights a **test accuracy of 72.6%**, with precision, recall, and f1-scores for both classes being approximately **0.71 to 0.74.** This balance between precision and recall suggests the model performs consistently across both classes. The macro and weighted averages for precision, recall, and f1-score confirm the model's stability and reliability in handling the dataset.

The confusion matrix *(Figure 2)* shows the classification results, with **73 true negatives and 62 true positives**, indicating the model correctly identified most cases of chronic absenteeism and non-absenteeism. However, there were **25 false positives and 26 false negatives**, which could indicate room for improvement in balancing sensitivity and specificity.

Feature importance *(Figure 3)* provides insights into the variables driving the predictions. The most influential features include **AttRate_1YearsAgo, DaysAbs_1YearsAgo, and ANON_ID,** indicating that attendance rates and days absent from the previous year were critical predictors. Other important features include historical attendance data (**AttRate_2YearsAgo, DaysAbs_2YearsAgo)** and demographic variables like **Age and Grade.**

The ROC curve for the below-grade-6 model (*Figure 4*) further validates the model's performance in distinguishing between cases of chronic absenteeism and non-absenteeism. The **Area Under the Curve (AUC) is 0.81**, indicating strong predictive ability. This score suggests that the model has a good balance between sensitivity (true positive rate) and specificity (false positive rate) and is effective in distinguishing between the two classes. The curve's shape shows the model performs significantly better than random chance, represented by the diagonal line **(AUC = 0.5)**. This evaluation confirms the model's robustness and reliability in making accurate predictions for chronic absenteeism.

4. Model 2: Above Grade 6 Model to Predict 2023-2024 Chronic Absenteeism Status

The model 2 for above grade 6 data follows similar feature engineering steps as the below-grade-6 model 1, with the addition of GPA-related features, including **current GPA and historical GPA values**. These numeric features are also standardized.

```
Test Accuracy: 0.7972665148063781

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.76      0.78       209
           1       0.79      0.83      0.81       230

    accuracy                           0.80       439
   macro avg       0.80      0.80      0.80       439
weighted avg       0.80      0.80      0.80       439
```

Figure 5

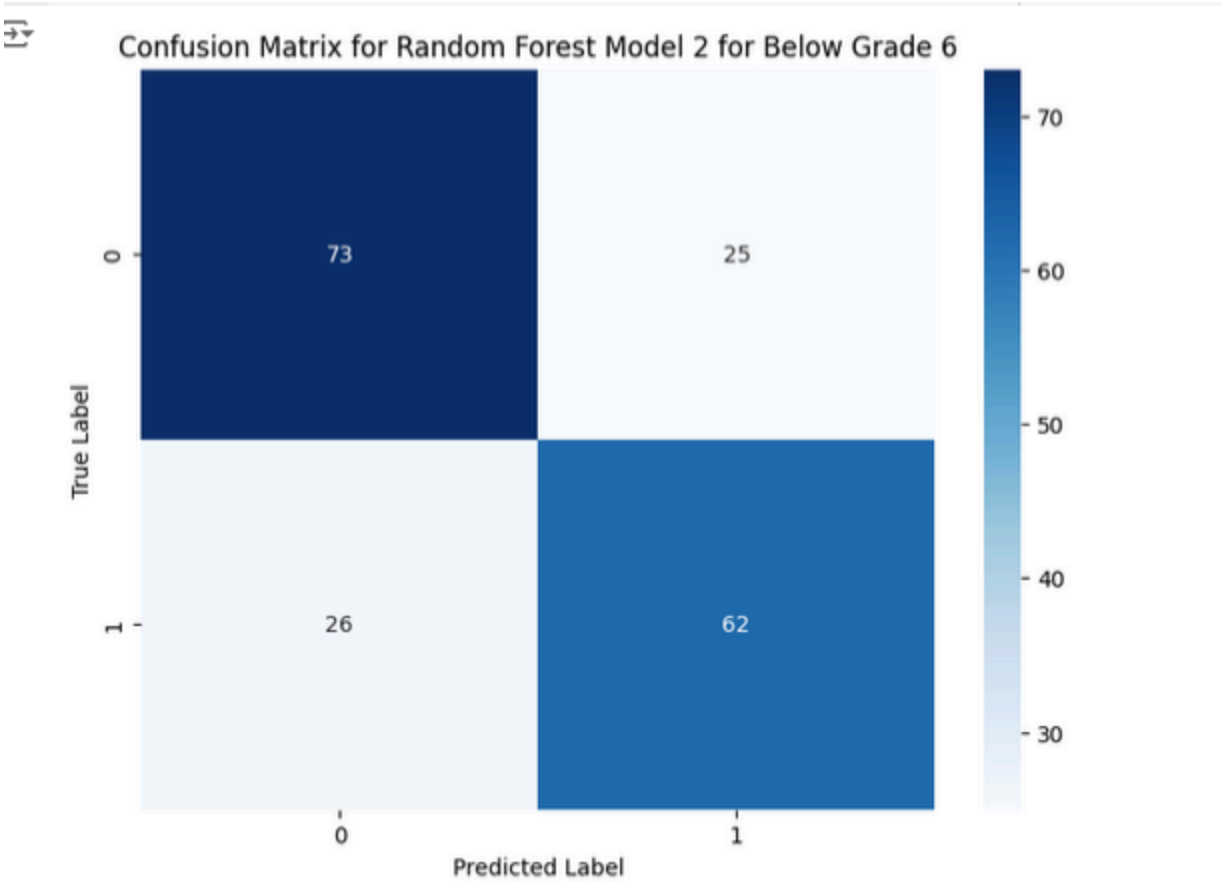

Figure 6.

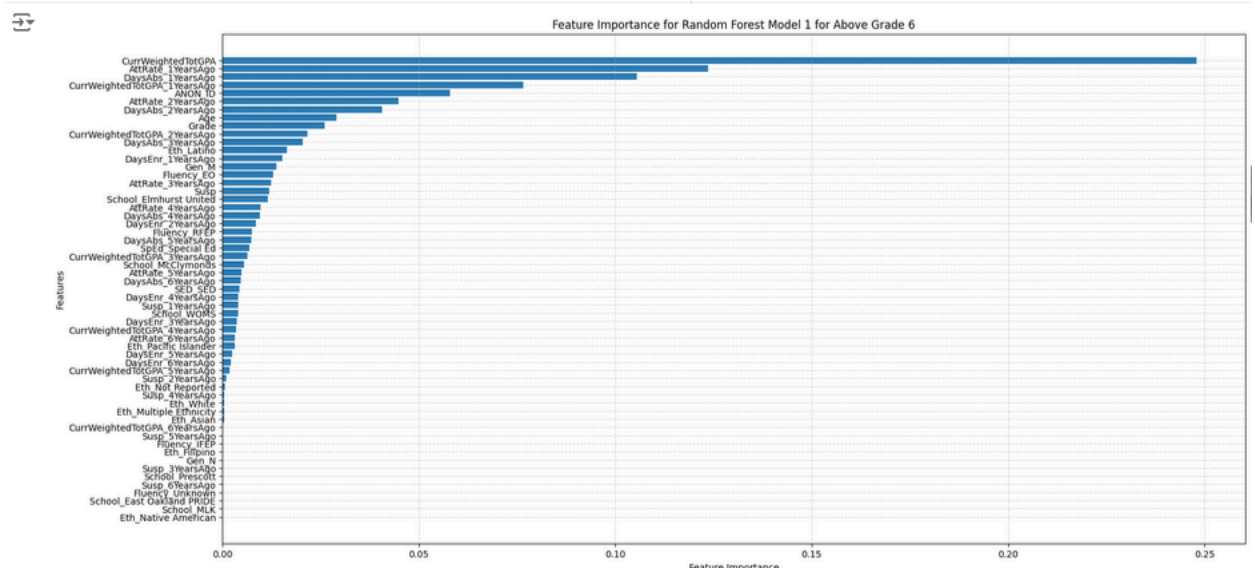

Figure 7

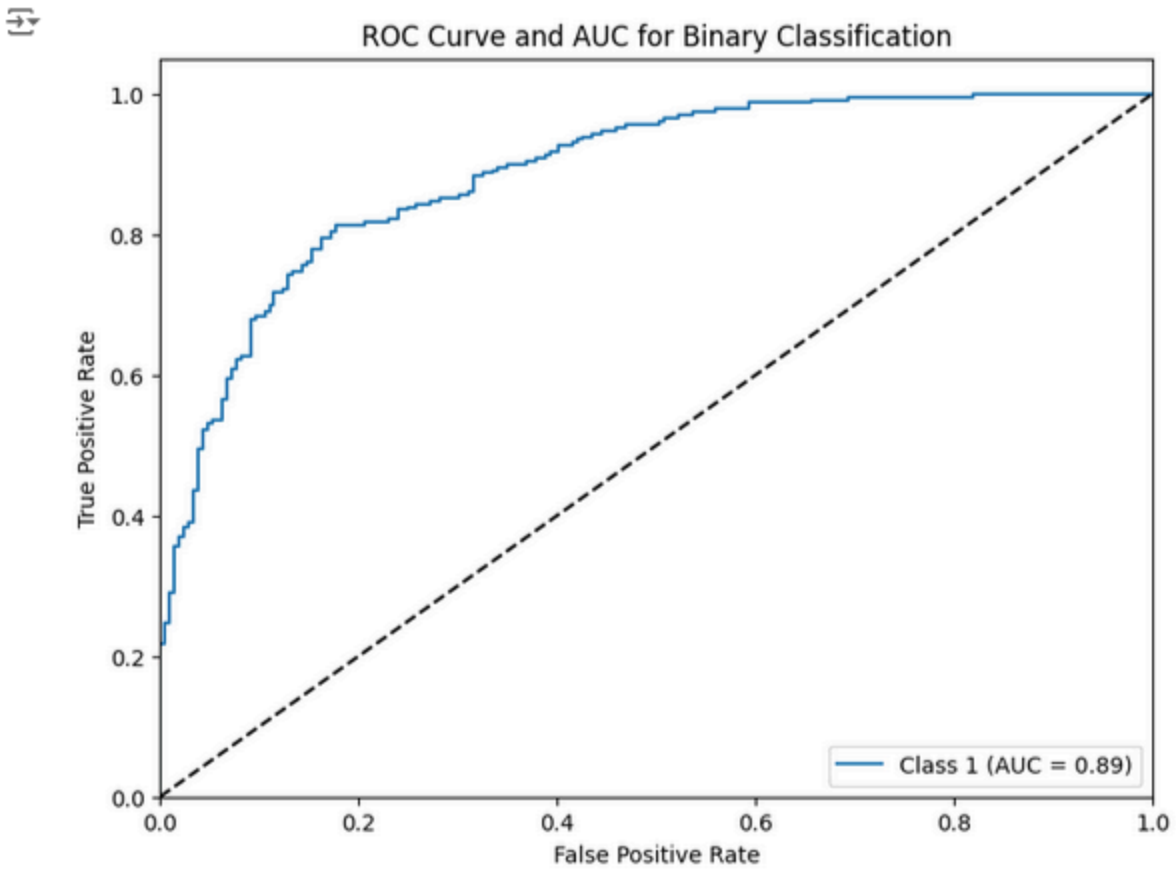

Figure 8

The classification report (*Figure 5*) shows a **test accuracy of 79.7%**, with precision, recall, and f1-scores for both classes hovering around **0.80**. These results indicate balanced model performance for predicting both chronic absenteeism and non-absenteeism.
The confusion matrix (*Figure 6*) highlights the model's ability to correctly classify the majority of cases, with a relatively small number of **false positives and false negatives**. This balance further reflects the model's reliability in handling the dataset.
The feature importance chart (*Figure 7*) demonstrates that **GPA-related features, including CurrWeightedTotGPA from various years**, were the most influential predictors, followed by **attendance and absence rates**. This finding highlights the significance of academic performance and attendance behavior in predicting chronic absenteeism.
The ROC curve *(Figure 8*) shows the model's ability to distinguish between the two classes, with an **AUC of 0.89**. This high AUC value indicates strong discriminatory power, with the model achieving a good balance between sensitivity and specificity.

5. Predicting Chronic Absenteeism for 2024-2025

The trained models from the previous sections, **Model 1 (below-grade-6) and Model 2 (above-grade-6)**, were utilized to predict chronic absenteeism for the 2024-2025 school year.

To achieve this, the training data was carefully prepared to match the features and structure of the models for compatibility. For categorical features, the same variables as in the original models were included, ensuring consistency in encoding and representation. For numerical features, columns such as A**ttRate, DaysEnr, and DaysAbs from the 2023-2024** school year were shifted to **"_1YearAgo,"** while data from prior years was similarly adjusted to reflect a one-year shift. For **Model 2, GPA-related** features, including the **current GPA and historical GPAs**, were processed similarly to align with the feature engineering steps used during training. Additionally, the feature set was reviewed and double-checked to ensure it matched the structure of the original models, maintaining compatibility and avoiding issues during prediction.

After transforming the data features, the prepared dataset was fed into both models to predict chronic absenteeism for the **2024-2025 school year**.Then, the original dataset was concatenated with additional columns containing the actual chronic absenteeism status for the **2023-2024 school year**, as well as the **models' predictions for both 2023-2024 and 2024-2025**. This new dataset provided a comprehensive view of the models' predictive capabilities over time.

To analyze the models' insights, the data was filtered to identify **students whose actual and predicted chronic absenteeism status for 2023-2024 was non-chronically absent but were predicted to be chronically absent in 2024-2025**. A closer look at these cases revealed a pattern: students with **attendance rates in 2023-2024 at the borderline level (around 90%-92%), low GPAs (below 2.0), or a combination of both** were frequently predicted to become chronically absent in the following school year. This finding highlights the importance of borderline attendance and academic performance as early indicators of future absenteeism, offering actionable insights for early interventions.

```
# Look at prediction not not the same as 2324 predict and true label of 2324 and actual label for 23-24 is 0
data_with_pred[(data_with_pred['ChroAbs 2425 pred'] != data_with_pred['ChroAbs']) &
 (data_with_pred['ChroAbs 2425 pred'] != data_with_pred['ChroAbs 2324 pred'])&
              (data_with_pred['ChroAbs'] == 0)]
```

	ANON_ID	Birthdate	Gen	Eth	Fluency	SpEd	Grade	AttRate	DaysEnr	DaysAbs	Susp	CurrWeightedTotGPA	SED	School	Year	ChroAbs	Age	ChroAbs 2324 pred	ChroAbs 2425 pred
104	10750	2008-10-11	F	Latino	RFEP	Not Special Ed	9	1.0000	16.0	0.0	0.0	0.00	SED	Castlemont	23-24	0	15	0	1
169	15595	2007-01-23	F	Latino	EL	Not Special Ed	11	0.9056	180.0	17.0	0.0	1.50	SED	Castlemont	23-24	0	16	0	1
170	15603	2006-04-15	F	Latino	EL	Not Special Ed	10	0.9000	180.0	18.0	0.0	2.38	SED	Castlemont	23-24	0	17	0	1
226	20213	2008-05-08	M	Latino	RFEP	Not Special Ed	10	0.9000	180.0	18.0	0.0	2.13	SED	Castlemont	23-24	0	15	0	1
313	26338	2008-08-06	M	Latino	RFEP	Not Special Ed	9	1.0000	17.0	0.0	0.0	0.00	SED	Castlemont	23-24	0	15	0	1
392	34849	2008-04-12	M	Latino	EL	Not Special Ed	9	0.9000	180.0	18.0	0.0	1.86	SED	Castlemont	23-24	0	15	0	1
513	47225	2007-01-29	M	Latino	EL	Not Special Ed	10	0.9111	180.0	16.0	0.0	2.25	SED	Castlemont	23-24	0	16	0	1
558	50760	2009-06-08	M	Multiple Ethnicity	EO	Not Special Ed	9	0.9278	180.0	13.0	0.0	2.25	SED	Castlemont	23-24	0	14	0	1
559	50761	2009-06-08	M	Multiple Ethnicity	EO	Not Special Ed	9	0.9278	180.0	13.0	0.0	2.25	SED	Castlemont	23-24	0	14	0	1
657	58753	2007-03-12	M	Latino	RFEP	Not Special Ed	11	0.9000	180.0	18.0	0.0	1.83	SED	Castlemont	23-24	0	16	0	1
659	59107	2006-08-04	F	Latino	RFEP	Not Special Ed	12	0.9056	180.0	17.0	0.0	4.00	SED	Castlemont	23-24	0	17	0	1
678	60377	2007-03-07	M	Multiple Ethnicity	EO	Not Special Ed	11	0.9222	180.0	14.0	0.0	1.75	SED	Castlemont	23-24	0	16	0	1
680	60382	2007-03-07	M	Multiple Ethnicity	EO	Not Special Ed	11	0.9222	180.0	14.0	0.0	1.75	SED	Castlemont	23-24	0	16	0	1