Probing Influences on Singaporean Academia

a user guide to the journey

HOME PAGE

- Overview and Objectives of Project PISA
- Variable List
  Details of all the predictors available for exploration and analysis
- Key Statistics
  Includes distribution of subject scores, breakdown of gender and school type

DATA EXPLORATION

- Variable Overview
  For univariate and bivariate analysis. Explore distribution, proportions and key summary statistics of each variable
- Sankey Diagram
  Visualise the flow from one variable to another

CONFIRMATORY ANALYSIS

- Analysis of Scores
  Evaluate if a user-defined value significantly deviates from the population mean
- Analysis of Predictors
  Ascertain whether there exists a statistically significant distinction among two or more categorical groups

CLUSTER ANALYSIS

- Association between Predictors
  Measure dependency/association between two variables
- Latent Class Analysis
  Show the distribution of categorical variables within each latent class.

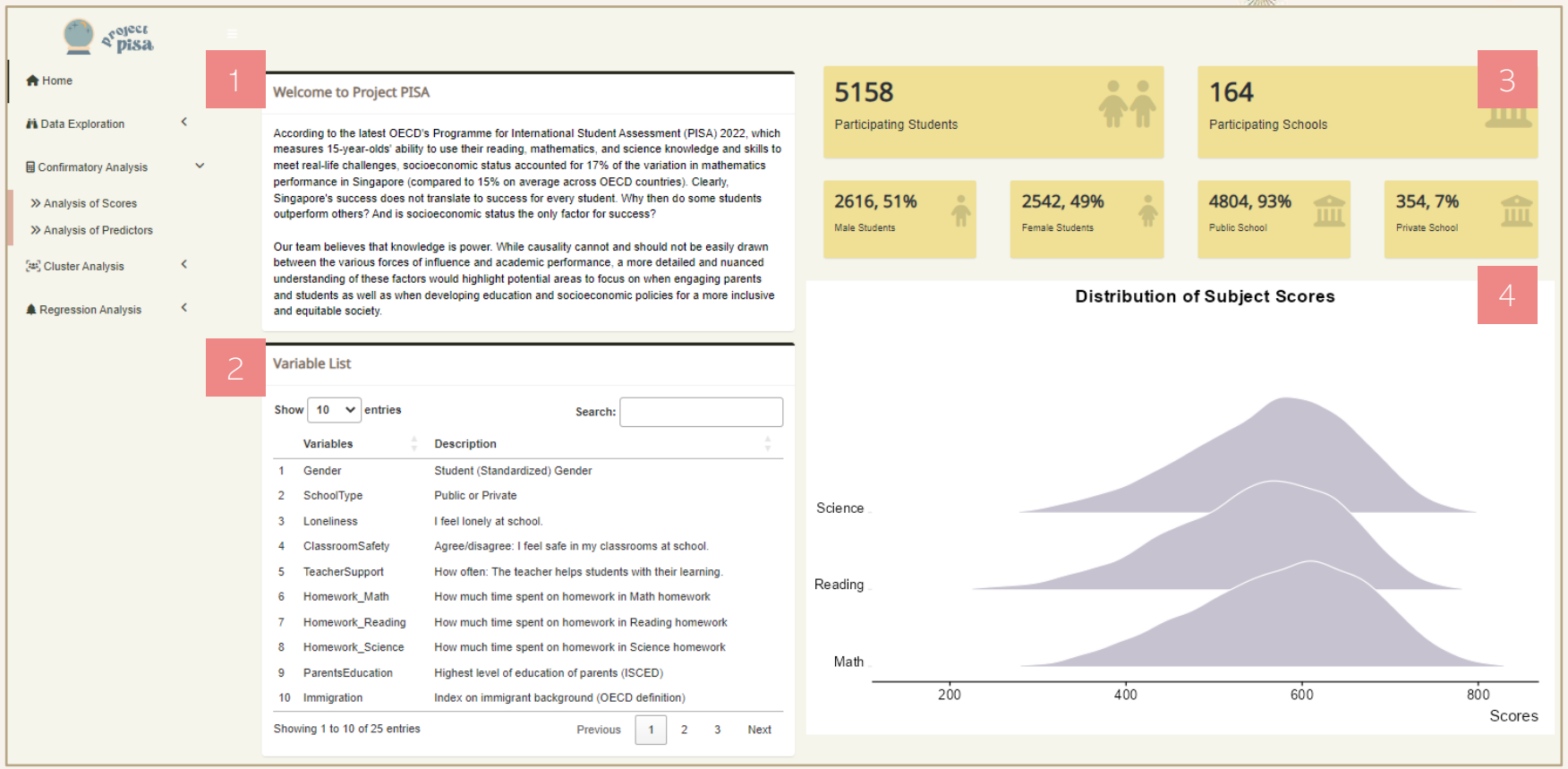
REGRESSION ANALYSIS

Users can select target and predictor variables as well as the train-test partition split to be applied across the three available models – regression tree, random forest, and gradient boosting

Within each model, the user can further calibrate the model, and select resampling options

Model evaluation methods include plots of model fit as well as statistical metrics such as R-square, Root Mean Square Error, and Mean Absolute Error.

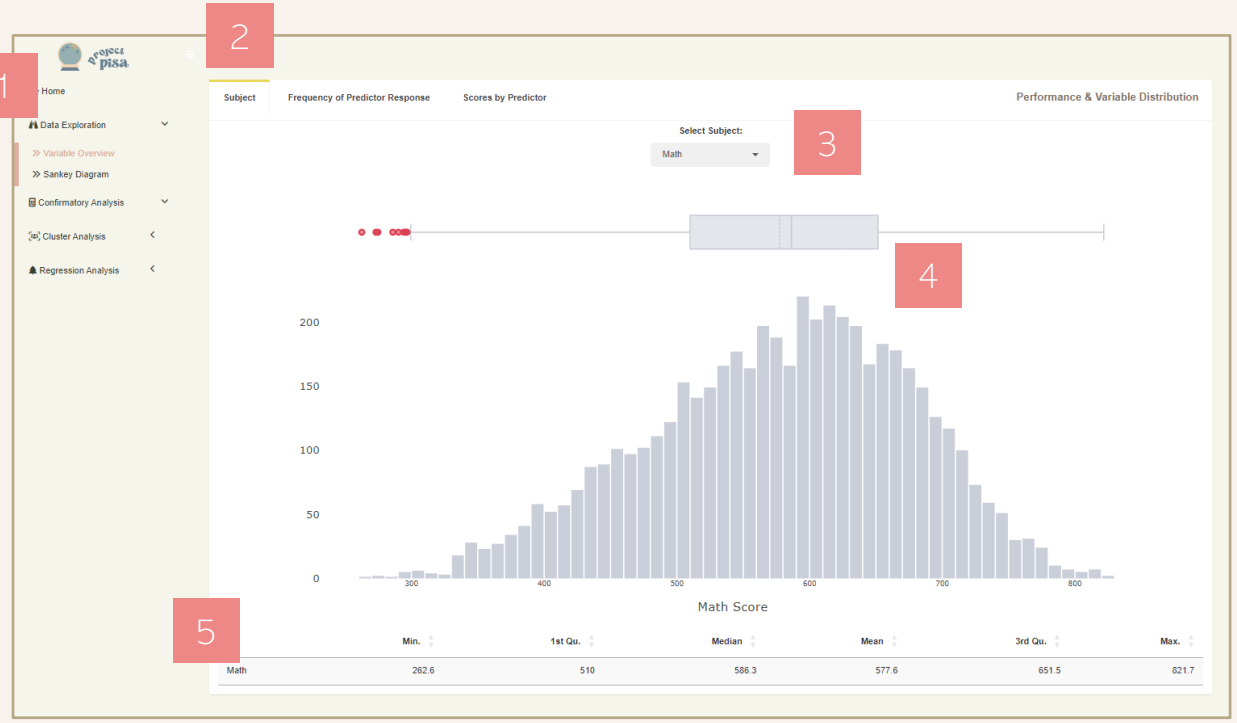
1. Home Page

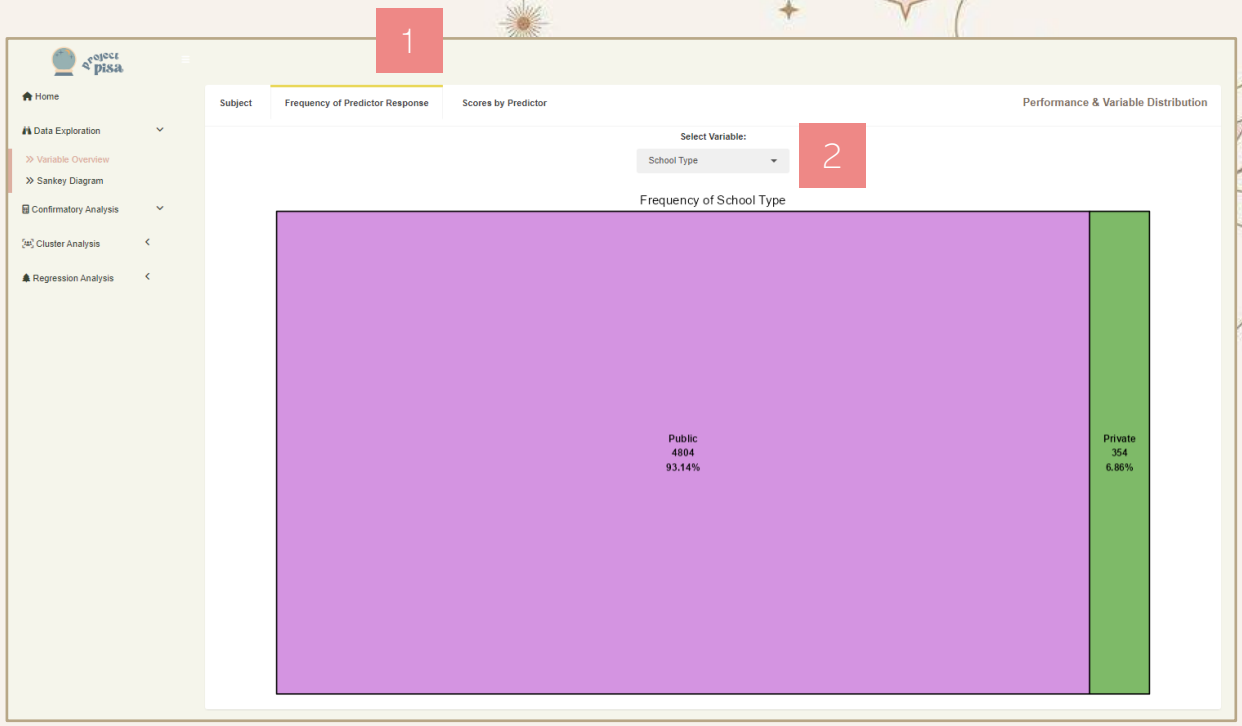


On this page, the following information can be found:

1. **Overview and objectives** of the project, accompanied with a short description of the web application.
2. **Variable List** which consists of the complete set of variables and what each of them represents.
3. **Key statistics** of the dataset which includes the total number of participants and schools involved, the split by gender and school type.
4. **Distribution of target variables** (i.e. Subject Scores).

2. Data Exploration





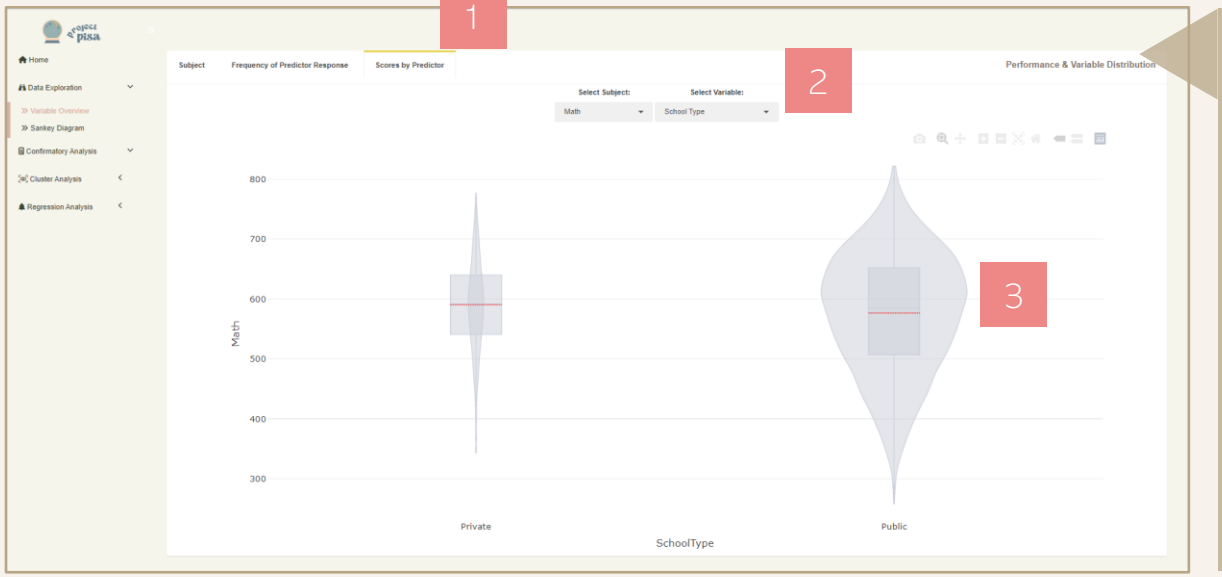
The first feature of the Data Exploration tab is the **Variable Overview** page.

1. First, click on ∇ beside **Data Exploration** on the side menu to show the submenu items. Click on **Variable Overview**.
2. Click on the **Subject tab** on the right panel.
3. Click on the **∇ at the dropdown menu** to select the target variable.
4. Once a selection is made, users can **hover over the interactive histogram** to explore the distribution of data. Hovering over the **interactive boxplot** allows users to see the key summary statistics of the data such as the mean, median, minimum and maximum values.
5. The **table** allows a quick glance of the key summary statistics as well.

The next feature of the Data Exploration tab is the **Frequency of Predictor Responses tab.** The tree-map allows users to visualize the proportion of data represented using area size.

1. Click on the **Frequency of Predictor Responses** tab at the top of the right panel.
2. Click on the **∇ at the dropdown menu** to select the variable of interest
3. Once a selection is made, the tree-map would be refreshed to show users the split of responses for each variable.

2. Data Exploration

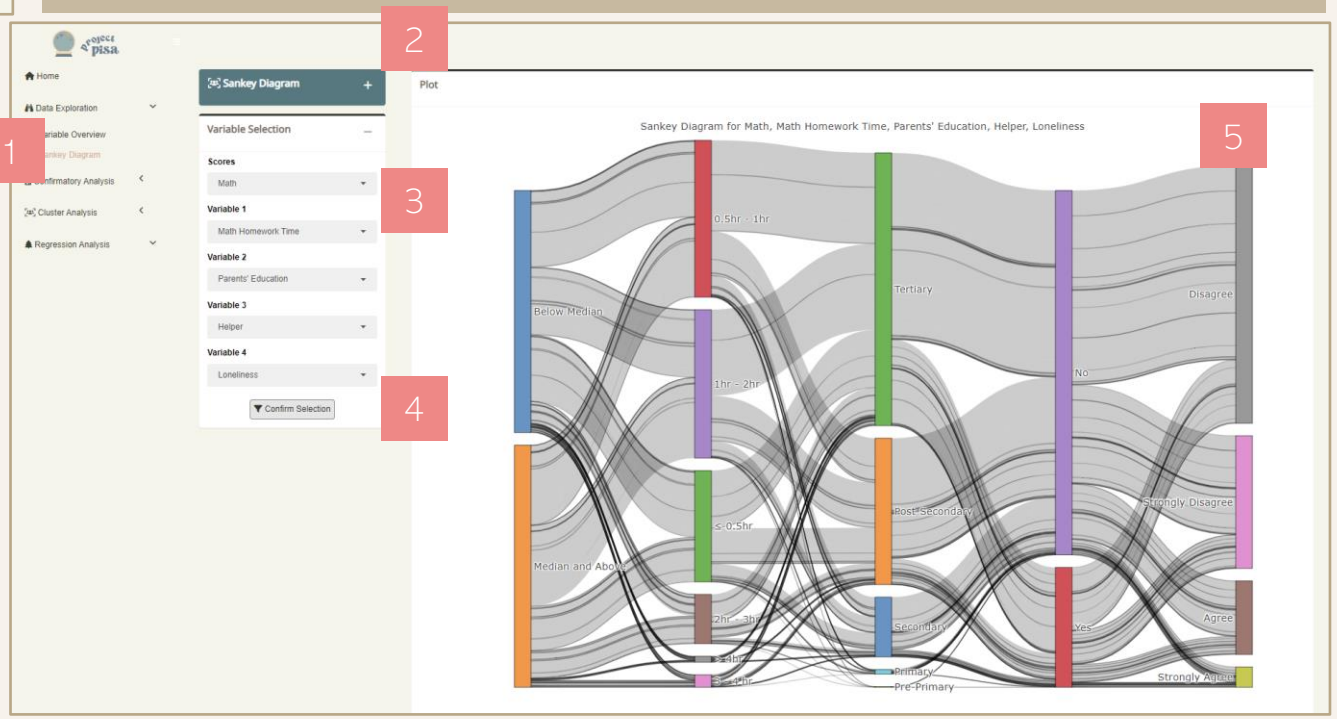


The next feature of the Data Exploration tab is the **Scores by Predictor tab.** The box-violin plot combines a boxplot and a kernel density plot. It is used to visualize the distribution of Subject Scores (i.e. target variable) between the different categorical responses for each variable.

1. Click on the **Scores by Predictor** tab at the top of the right panel.
2. Click on the **∇ of each dropdown menu** to select the target variable and predictor of interest.
3. Once a selection is made, the box-violin plot would be refreshed to show the distribution of Subject Scores across the various responses for the selected variable.
4. Users can hover over the box-violin plot to access key summary statistics of the distribution such as the mean, median, minimum, and maximum values.

A **Sankey Diagram** illustrates the flow of data between different categories or clusters, with the line thickness corresponding to the frequencies of the flows.

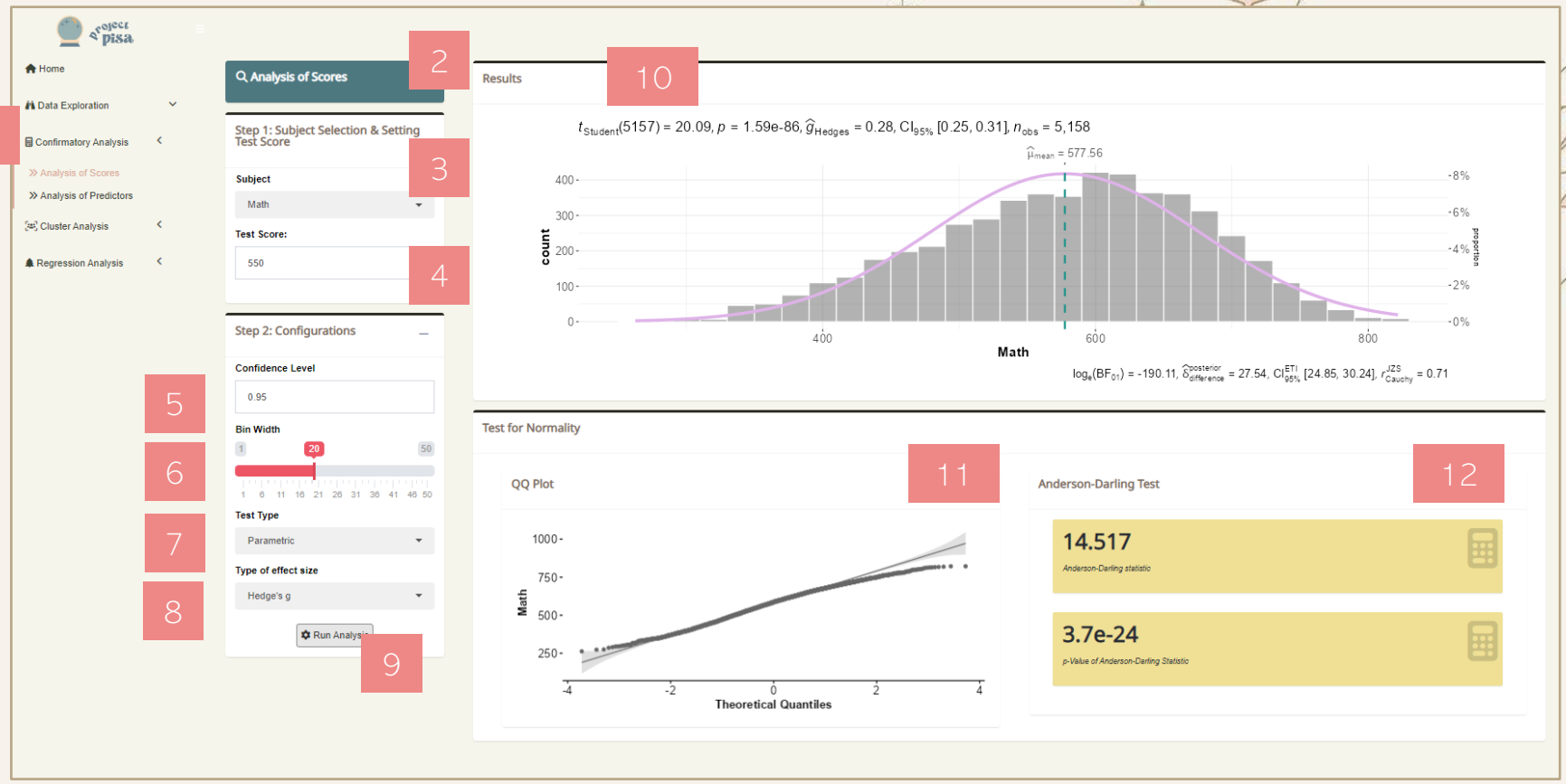
1. Click on **∇** beside **Data Exploration** on the side menu to expose the submenu if the submenu is collapsed. Click on **Sankey Diagram**.
2. The **+** beside the tab header provides a short description for users to understand the purpose of the illustration of the tab.
3. Click on the **∇ of each dropdown menu** to select **1 target variable interest and 4 predictor variables**.
4. Click on the **∇ Confirm Selection button** once selection is made.
5. Hover over the Sankey Diagram to explore how the responses for each variable flows to another.



3. Confirmatory Analysis

The Analysis of Scores tab allow users to evaluate if a user-defined value significantly deviates from the population mean.

1. Click on ∇ beside **Confirmatory Analysis** on the side menu to expose the submenu if the submenu is collapsed. Select **Analysis of Scores**.
2. The **+** beside the header provides a short description for users to understand the purpose of the tab.
3. Click on the **∇ of dropdown menu** to select **a target variable interest.**
4. User can **input a test score** to perform a statistical test on. User can also hover on the box to display the up/down arrows to change the input.
5. Input a **desired confidence level**.
6. Select the **desired bin width**.
7. Select **test type**.
8. Select the **type of effect size**.
9. Click on **Run Analysis.**
10. The top-right panel displays the data distribution with the statistical results. Results in the header includes the type of test performed, the test statistics and the p-value of the test statistic. Table on the right shows type of test performed for each test type >>



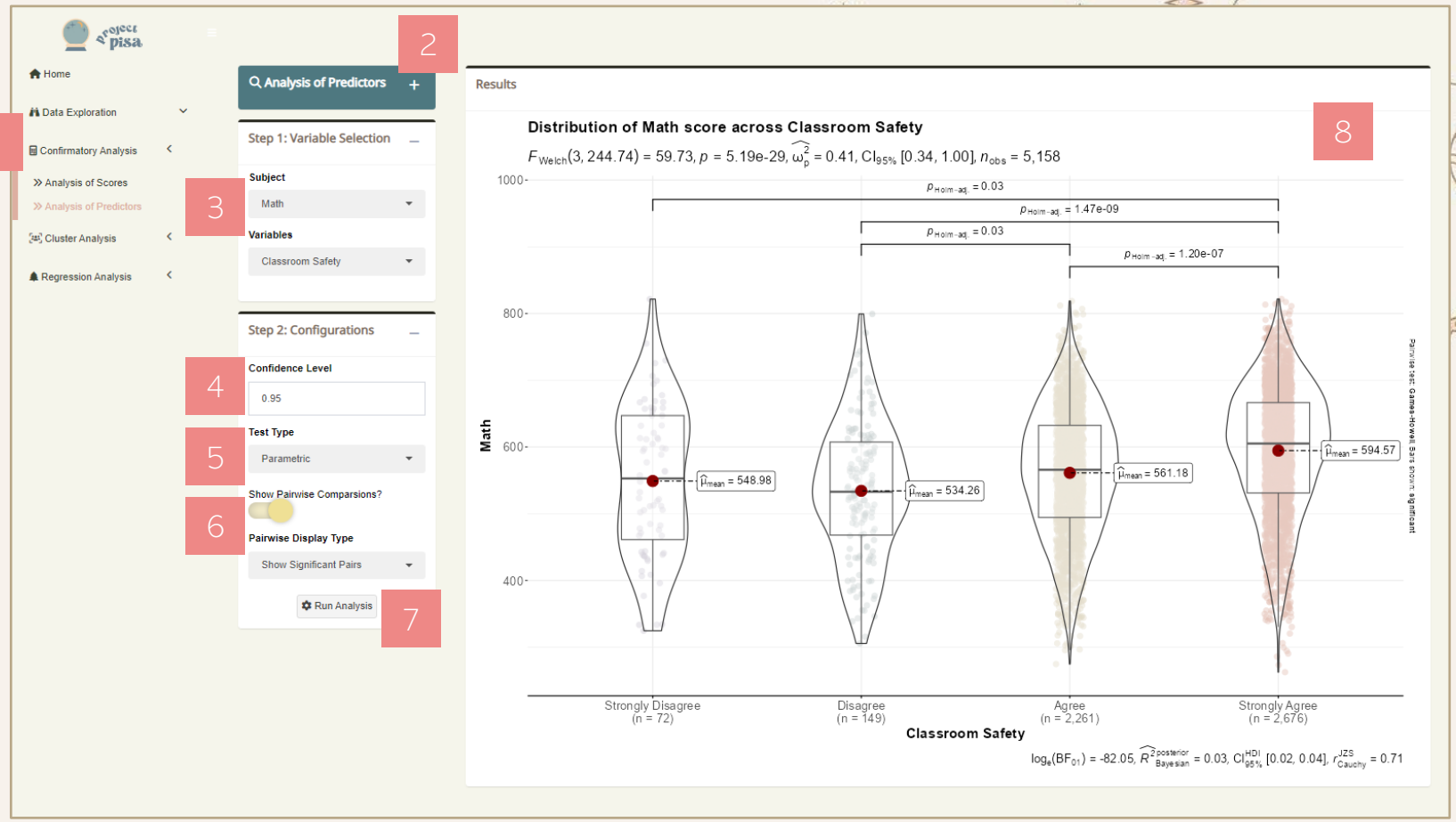
Type	Centrality Type	Test
Parametric	Mean	One-sample Student's *t*-test
Non-parametric	Median	One-sample Wilcoxon test
Robust	Trimmed Mean	Bootstrap-*t* method for one-sample test
Bayesian	MAP Estimator	One-sample Student's *t*-test

11. Points on the **QQ plot** provide an indication of normality of the dataset. If the data is normally distributed, the points will fall on the 45-degree reference line. If the data is not normally distributed, the points will deviate from the reference line.
12. The **Anderson-Darling test** is a statistical test that determines if a data set follows a **normal distribution**, and then determining the p-value for the statistic. **P-value is used to determine whether a null hypothesis should be accepted or rejected.**

3. Confirmatory Analysis

An ANOVA test is a statistical method utilized to ascertain whether there exists a statistically significant distinction among two or more categorical groups by examining differences in means using variance.

1. Click on ∇ beside **Confirmatory Analysis** on the side menu to expose the submenu if the submenu is collapsed. Click on **Analysis of Predictors**.
2. The **+** beside the header provides a short description for users to understand the purpose of the tab.
3. Click on the **∇ of dropdown menu** to select **a target variable and a predictor variable.**
4. Input a **desired confidence level**.
5. Select **test type**.
6. User should toggle the switch to the right should they wish to display the **Pairwise Comparisons** in the plot. If Pairwise Comparisons is turned on, the option for Pairwise Display Type would be displayed. Users can opt to show all pairs, only statistically significantly pairs, or only non-statistically significant pairs.
7. Click on **Run Analysis** once options are selected.



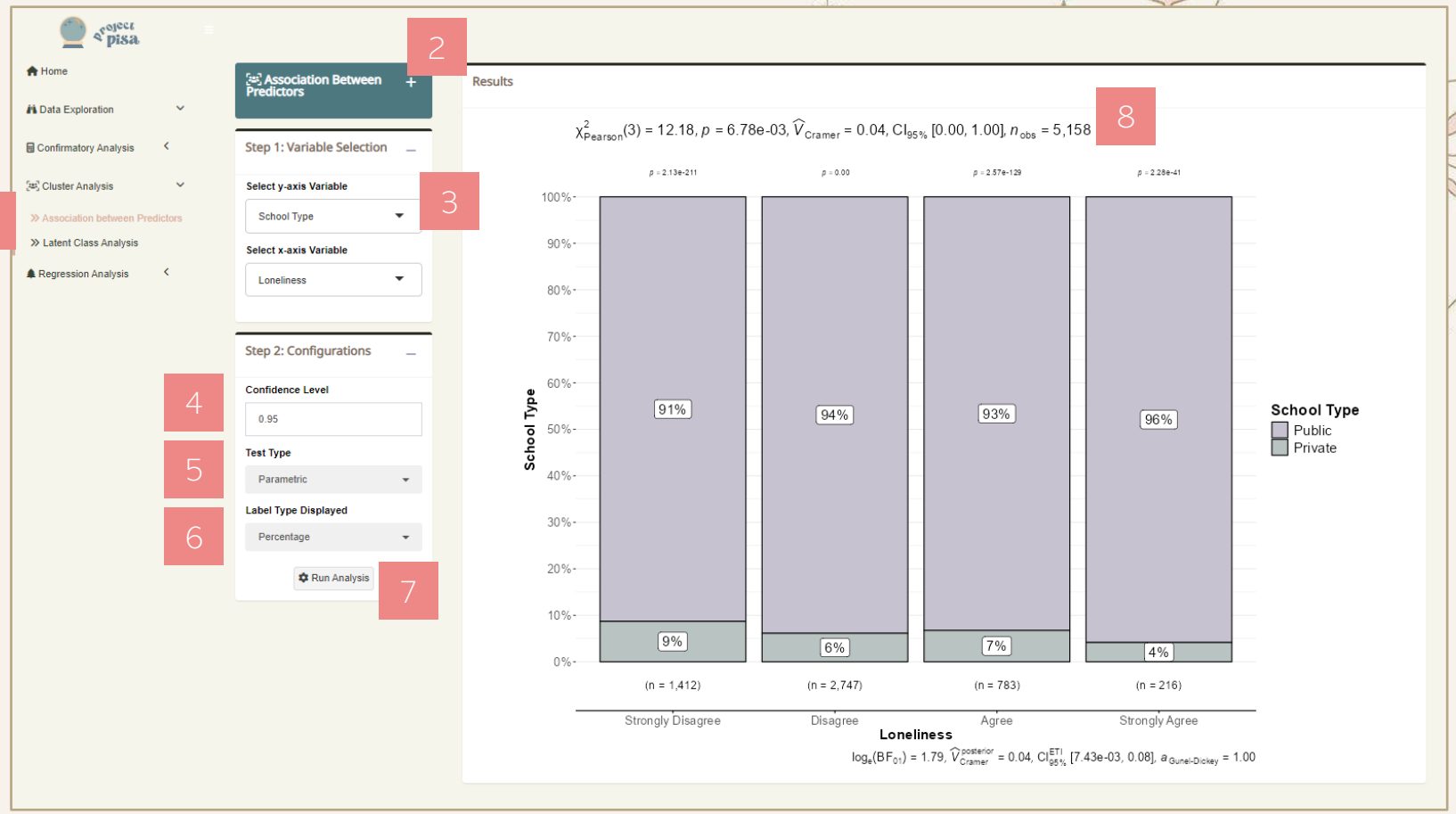
8. The right panel displays the data distribution of each response of the predictor variable with the statistical results. Results in the header includes the type of test performed, the test statistics and the p-value of the test statistic. See table on the right for type of test performed for each test type selection >>

Type	Centrality Type	Test
Parametric	Mean	Student's or Welch's t-test
Non-parametric	Median	Mann-Whitney U test
Robust	Trimmed Mean	Yuen's test for trimmed means
Bayesian	MAP	Student's t-test

4. Cluster Analysis

A test of association is a hypothesis test designed to establish and quantify the relationship between two distinct factors. When dealing with categorical variables, multicollinearity can be identified using the chi-square test. Multicollinearity can result in less reliable statistical inferences and may lead to skewed or misleading results.

1. Click on ∇ beside **Cluster Analysis** on the side menu to expose the submenu if the submenu is collapsed. Click on **Association between Predictor**.
2. The + beside the header provides a short description for users to understand the purpose of the tab.
3. Click on the ∇ **of dropdown menu** to select **a variable to plot on the y-axis and x-axis respectively.**
4. Input a **desired confidence level**.
5. Select **test type**.
6. Users have the option to select the type of labels to be displayed on the plot. This can be in count or percentage or both.
7. Click on **Run Analysis** once options are selected.



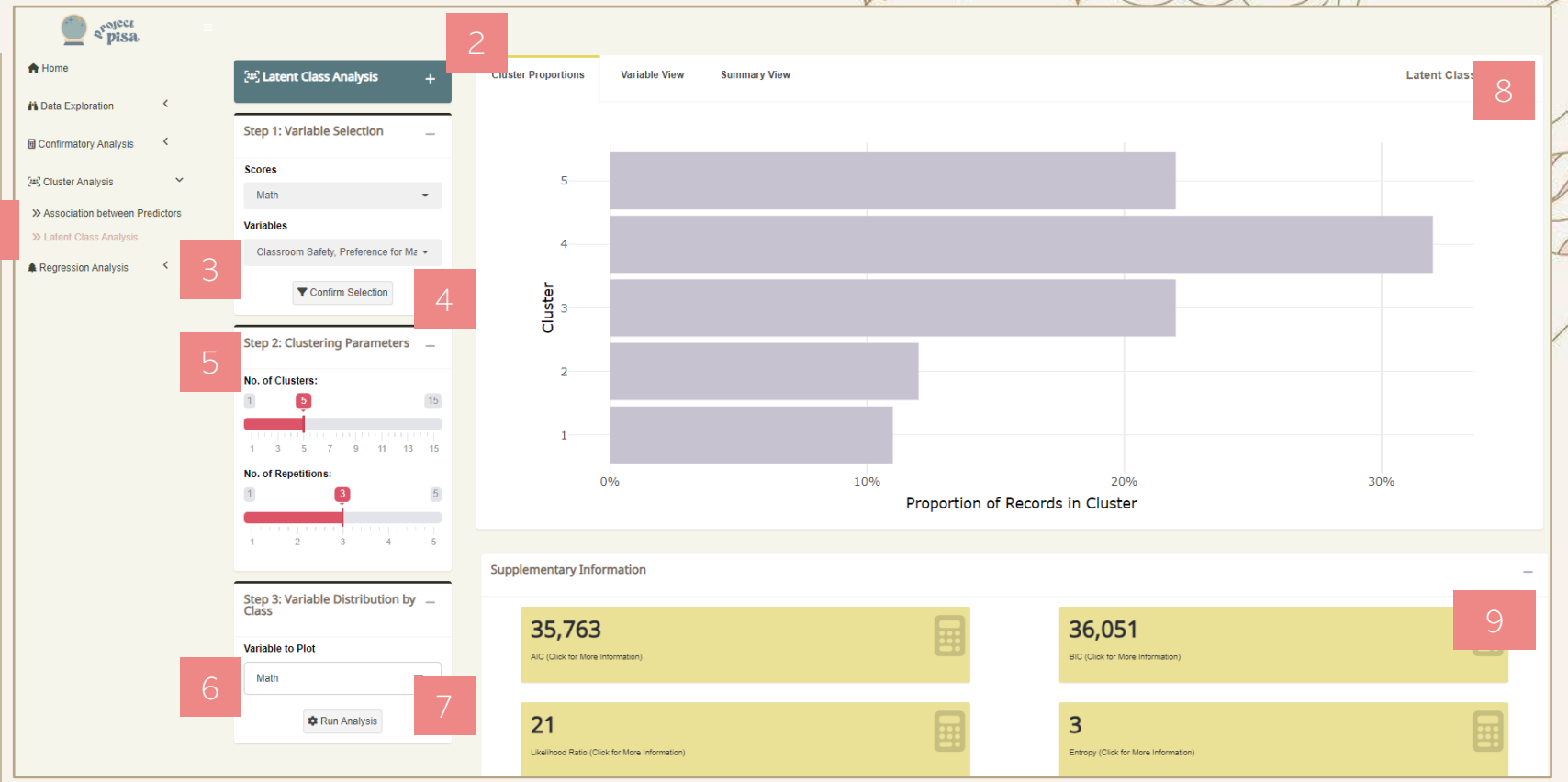
8. The right panel displays the data distribution of each response of the predictor variable with the statistical results. Results in the header includes the type of test performed, the test statistics and the p-value of the test statistic. See table on the right for type of test performed for each test type selection >>

Type	Test
Parametric	Pearson's chi-squared test
Non-parametric	Bayesian Pearson's chi-squared test
Robust	McNemar's chi-squared test
Bayesian	-

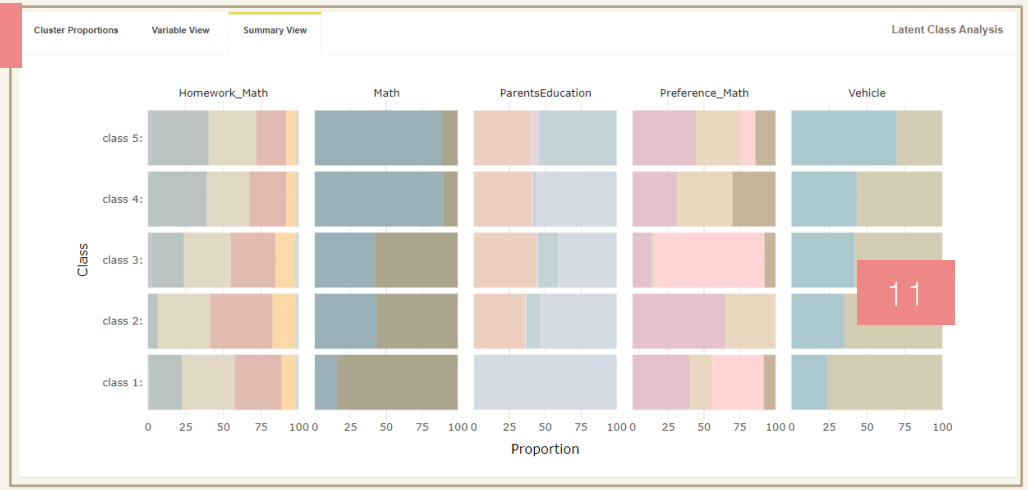
4. Cluster Analysis

A Latent Class Analysis (LCA) Bar Plot is used to show the distribution of categorical variables within each latent class.

1. Click on ∇ beside **Cluster Analysis** on the side menu to expose the submenu if the submenu is collapsed. Click on **Latent Class Analysis**.
2. The **+** beside the header provides a short description for users to understand the purpose of the tab.
3. Click on the ∇ **of dropdown menu** to select **target variable(s) and predictor variable(s).**
4. Click on **Confirm Selection** once variables have been selected.
5. Select the **desired number of clusters** and **number of repetitions**.
6. The selection for **Variable to Plot** is dependent on the variables selected under **Variable Selection**. This will affect the Variable View tab.
7. Click on **Run Analysis** once options are selected.
8. The **Cluster Proportions** tab shows the proportion of data that falls under each cluster. Hover over the plot to show the exact proportion amount.
9. **Supplementary Information** box can be expanded to show the relevant model diagnostic statistic.



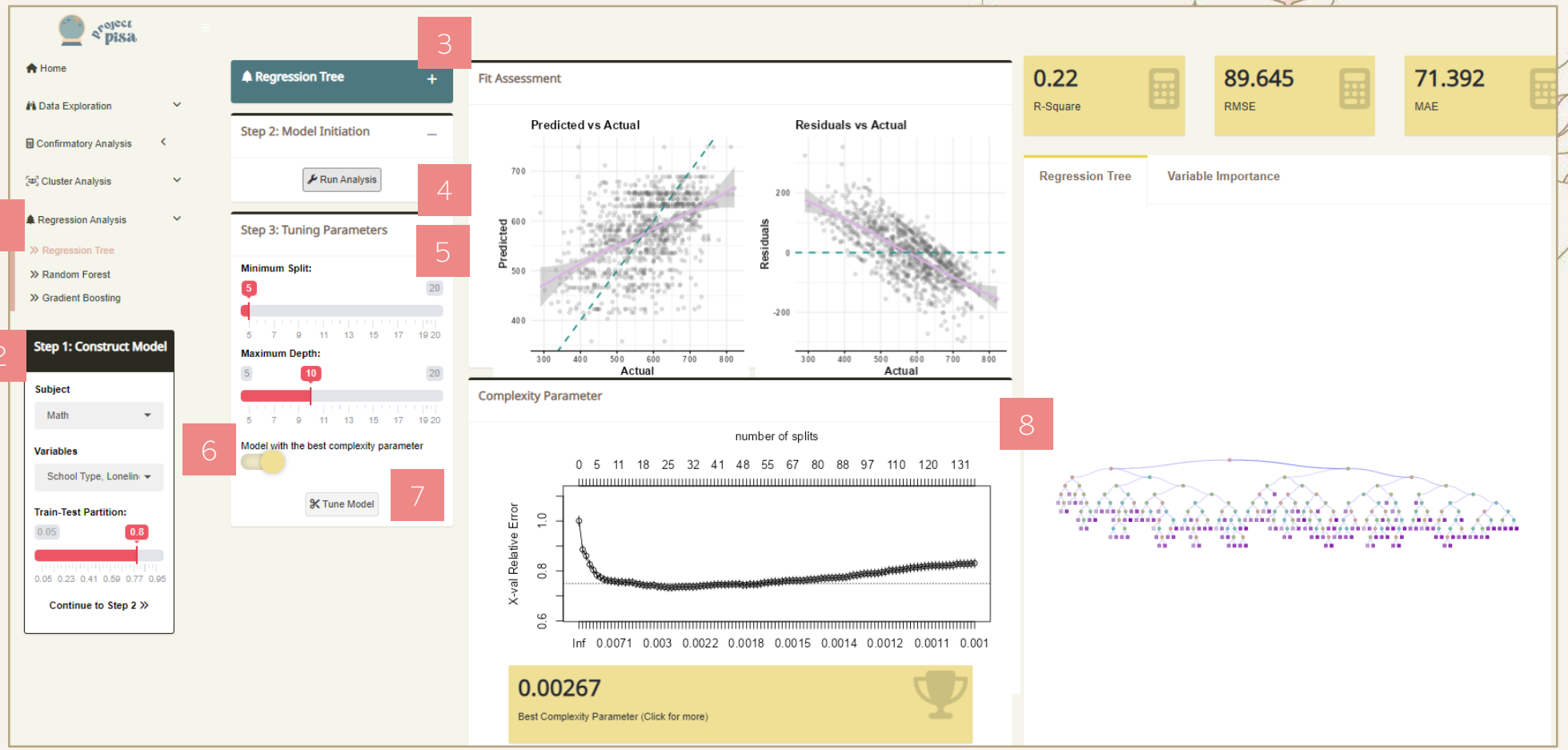
10. The **Variable View** and **Summary View** tabs show the stacked bars of the LCA by class and individual variable and by class and all variables, respectively.
11. Hover over the stacked bars to explore the attributes and proportions of each variable for each class.



5. Regression Analysis

A Regression Tree is a machine learning algorithm that partitions the data into subsets. The goal of a regression tree is to encapsulate the training data in the smallest possible tree, i.e. simplest possible explanation for the variation in scores.

1. Click on ∇ beside **Regression Analysis** on the side menu to expose the submenu if the submenu is collapsed. Click on **Regression Tree.**
2. Click on the ∇ **of dropdown menu** to select the target variable, predictor variable(s), and train-test partition ratio of choice for regression analysis.
3. The **+** beside the header provides a short description for users to understand the purpose of the tab.
4. Click on **Run Analysis** to initiate the model building process.



5. The **Tuning Parameters** panel is only displayed after the initial model has been triggered. The model can subsequently calibrated and refined with by selecting the minimum split and maximum dept.

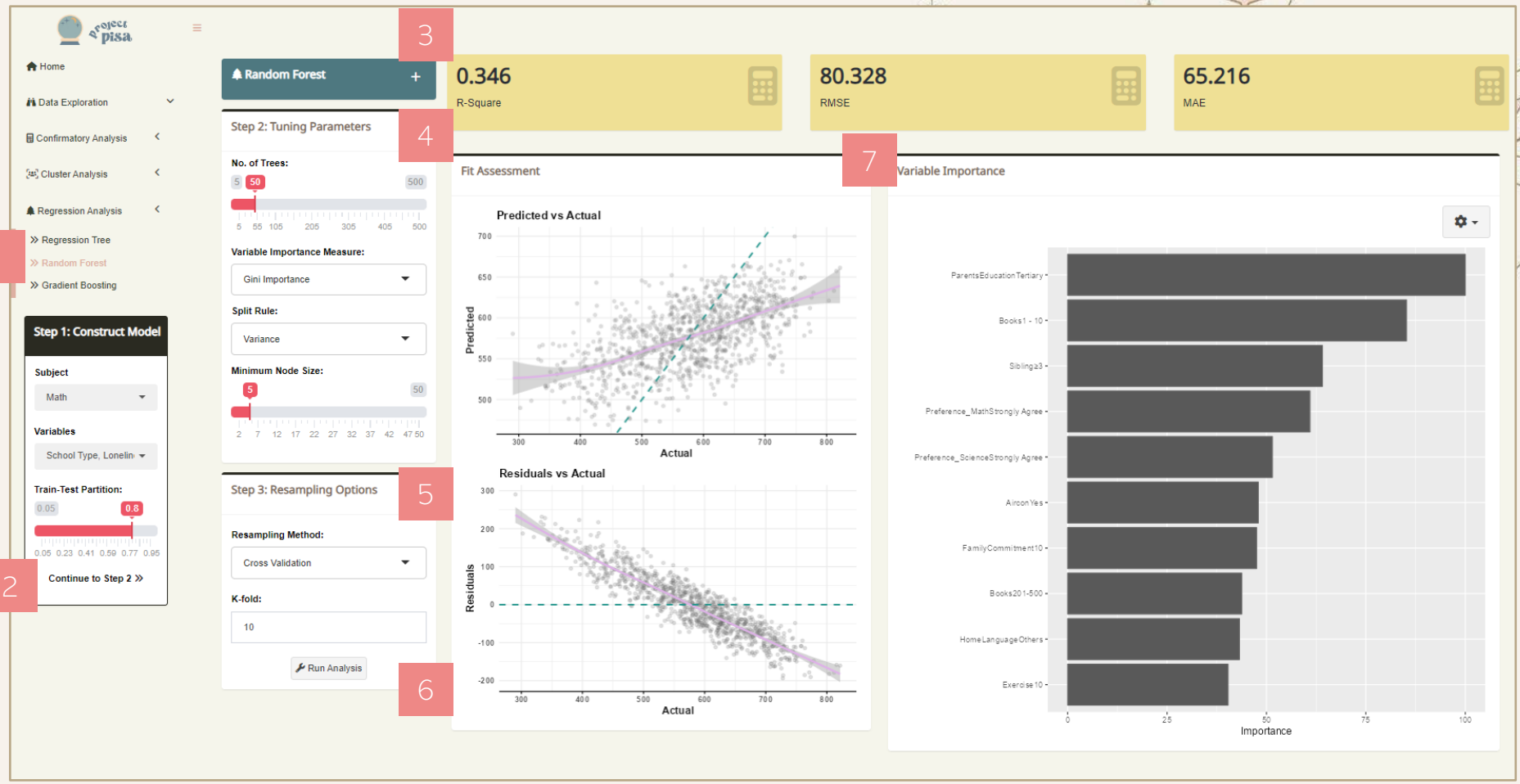
6. If the "**Model with the best complexity parameter**" is switched on, no further action is required. Else, input a complexity parameter value.
7. Click on **Tune Model**. Re-trigger the button to apply any further changes.

8. Users explore the model fit plots, model evaluation statistics, interactive regression tree, variable importance plot. Any changes to the variables or parameters requires retriggering of 'Run Analysis' before 'Tune Model' can be used.

5. Regression Analysis

Random Forest combines the opinions of many "trees" (individual models) to make better predictions, creating a more robust and accurate overall model.

1. Click on ∇ beside **Regression Analysis** on the side menu to expose the submenu if the submenu is collapsed. Click on **Random Forest.**
2. The selection for **Step 1: Construct Model** follows the selection made previously. If not previously selected, click on the dropdown menus to made the selection.
3. The **+** beside the header provides a short description for users to understand the purpose of the tab.
4. To further calibrate the model, users have the option to **change the number of trees and minimum node size by toggling the sliding bar. Variable importance measure and splitting rule** can be changed by selecting the **desired selection from the dropdown menu.**

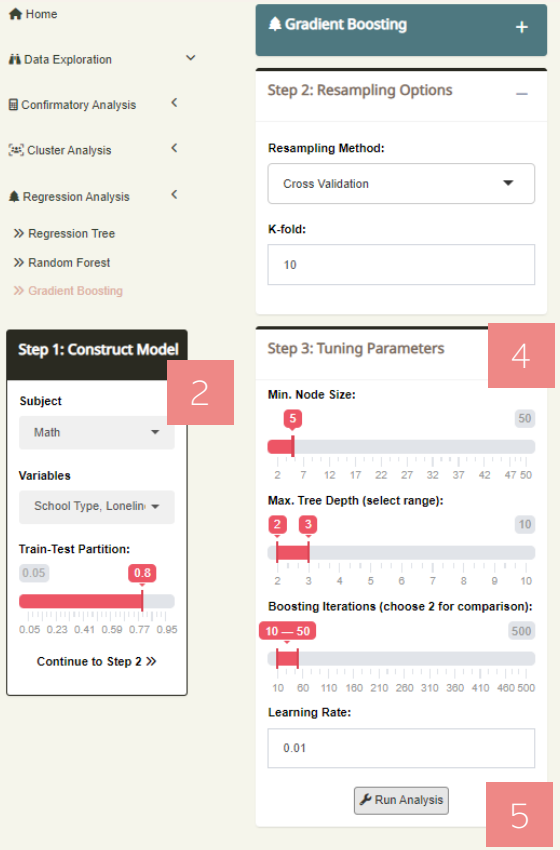


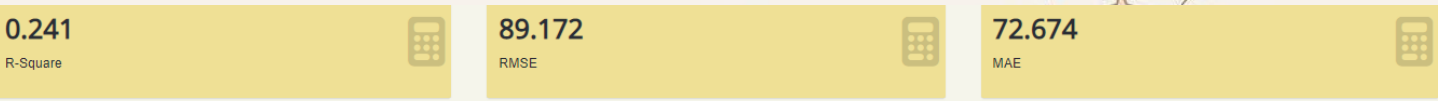
5. **Resampling options** can also be amended using the dropdown menu. **K-Fold and Repeat Count** would only be displayed if **Repeated Cross Validation** is selected. **K-Fold option** would be displayed when **Cross-Validation** is selected. No further options if a user chooses to use **Bootstrap**.
6. Trigger the **Run Analysis** button once selections are made.
7. User can subsequently explore the statistical results, variable importance, and model fit plots. Any further changes to the options will require retriggering of the **Run Analysis** button to refresh the results.

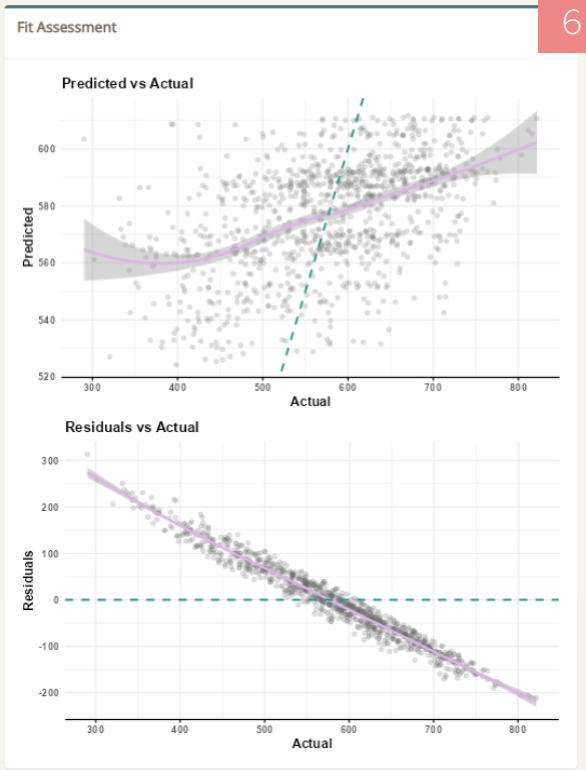
5. Regression Analysis

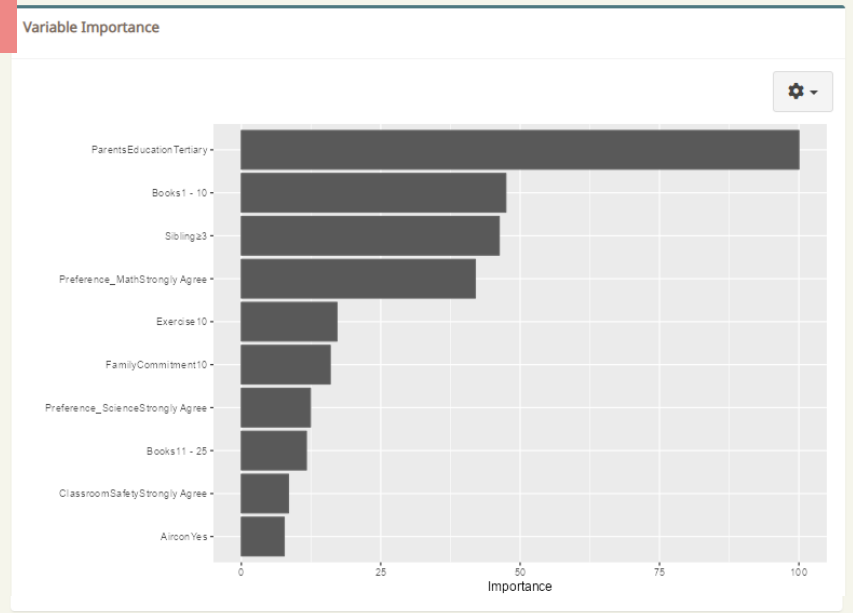
Gradient Boosting is an ensemble machine learning technique that combines the predictions from several models to improve the overall predictive accuracy.

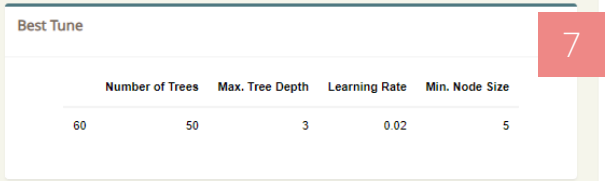
1. Click on ∇ beside **Regression Analysis** on the side menu to expose the submenu if the submenu is collapsed. Click on **Gradient Boosting.**
2. The selection for **Step 1: Construct Model** follows the selection made previously. If not previously selected, click on the dropdown menus to made the selection.
3. The + beside the header provides a short description for users to understand the purpose of the tab.
4. To further calibrate the model, users have the option to **change the min. node size, max. tree depth range, and boosting iteration selections** by toggling the respective sliding bars. **Learning rate** can be changed by inputting or using the up/down arrows.

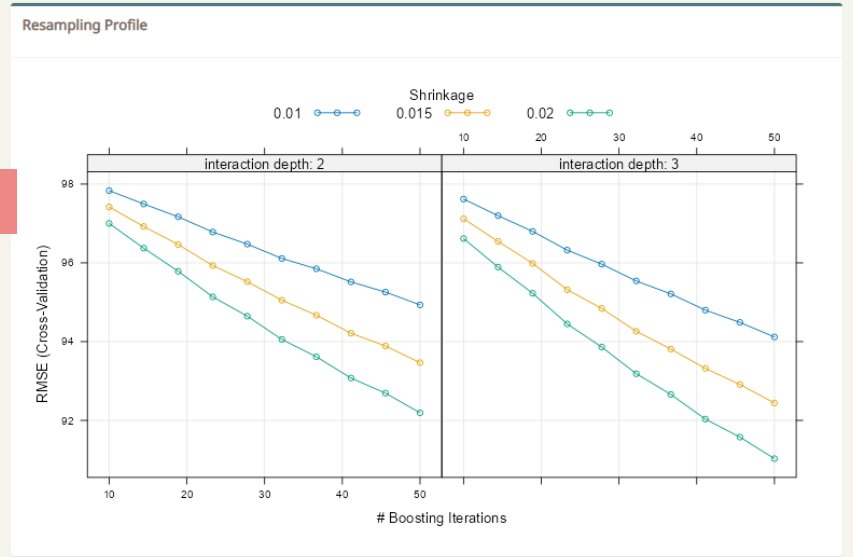






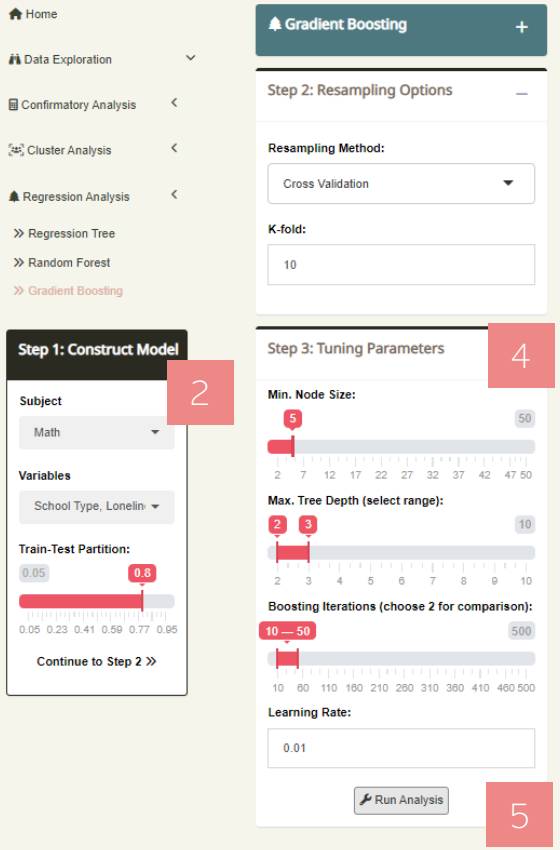


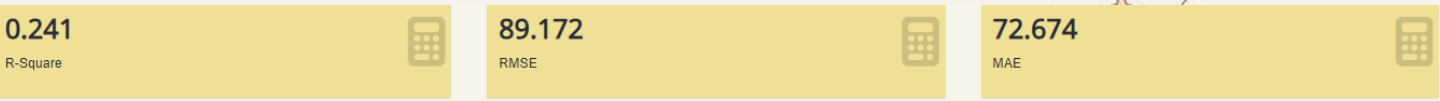


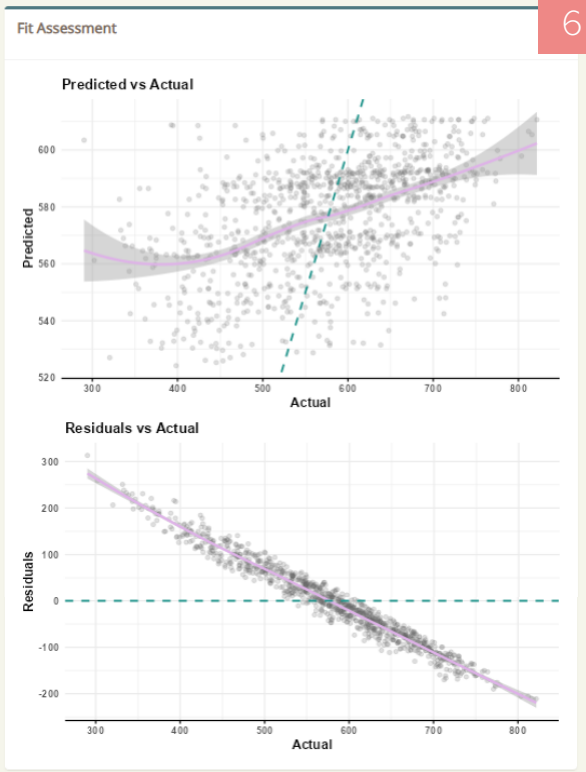


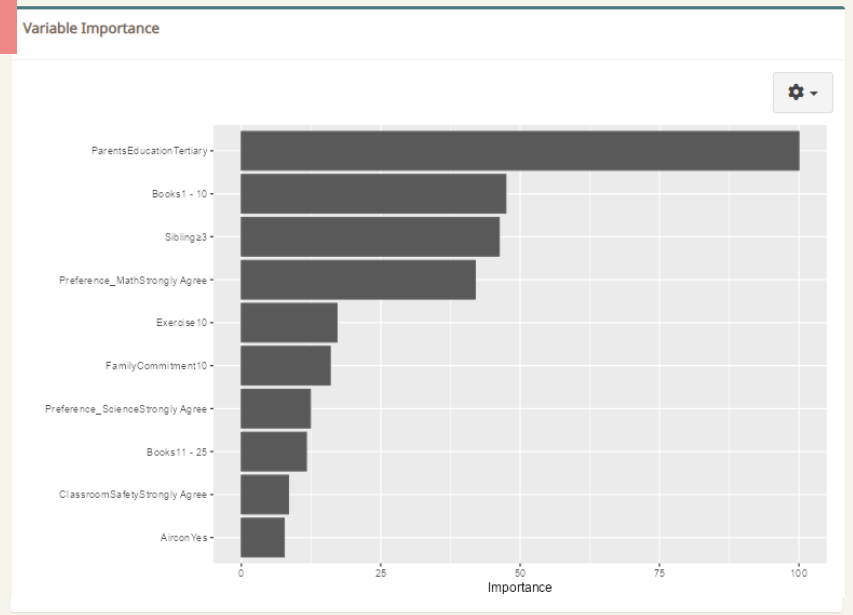
5. Regression Analysis

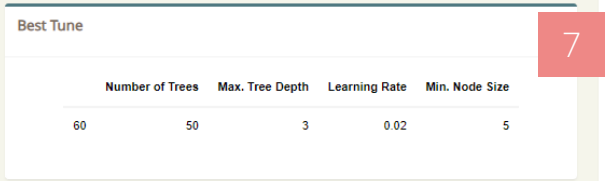
5. Trigger the **Run Analysis** button once selections are made.
6. User can subsequently explore the statistical results, variable importance, and model fit plots. Any further changes to the options will require retriggering of the **Run Analysis** button to refresh the results.
7. **Best Tune and Resampling Profile** panels are displayed only when **Cross Validation or Repeated Cross Validation methods** are selected under **Step 2: Resampling Options > Resampling Method**.

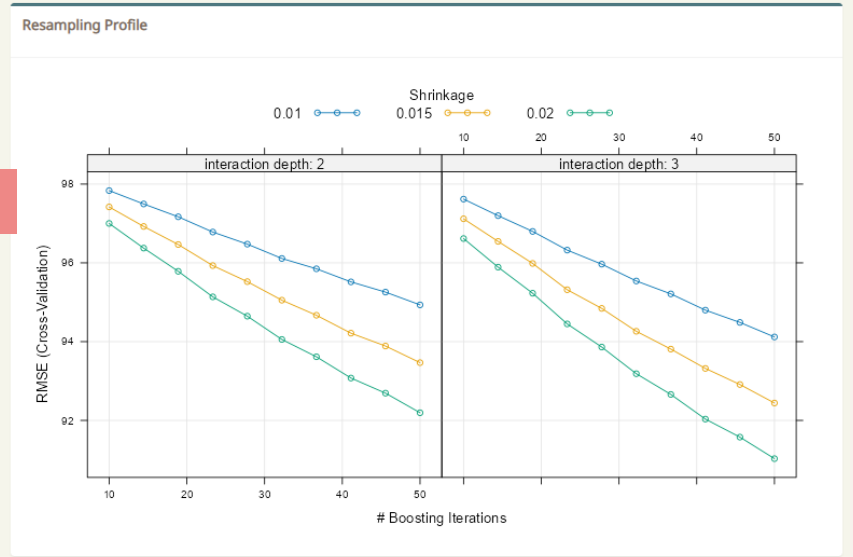












Best Tune and Resampling Profile panels show the best tuning parameters amongst the range selected by the user. This allow users to further finetune the model to obtain the best results.