MLighter

Reference manual

Record of revisions

Rev Nº	Issue Date	Inserted	
		Date	By
Initial	01 May 2022		Gorka Sanchez
V0.2	31 March 2024		Aidan Dakhama

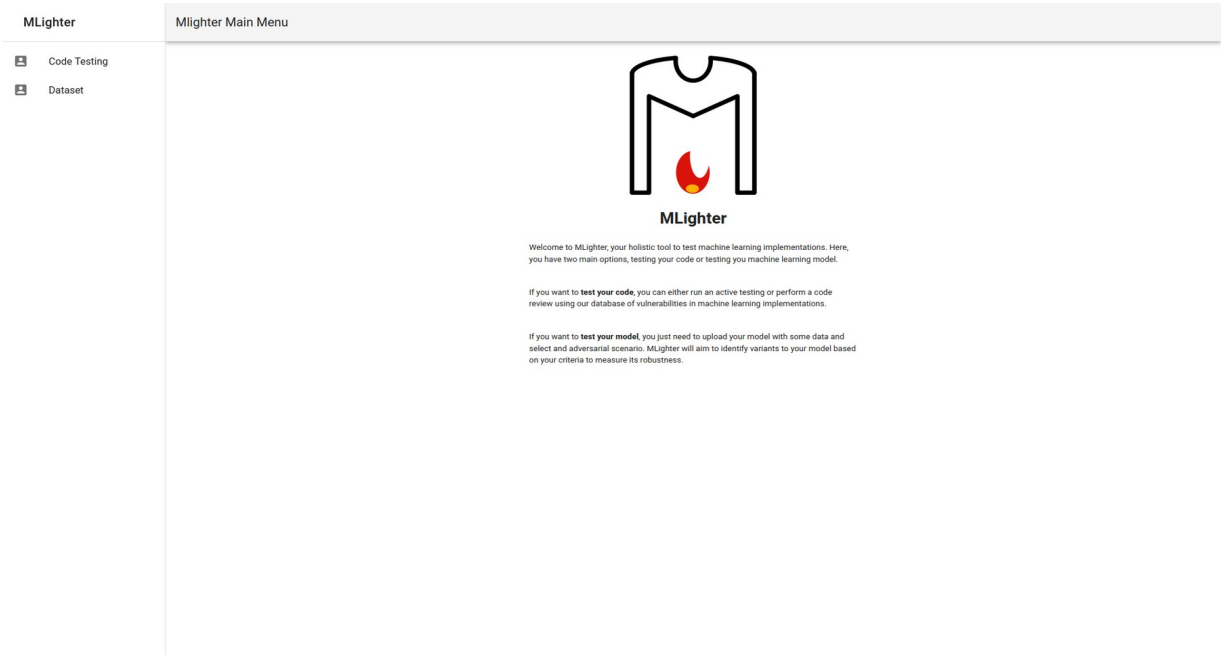
Table of contents

Introduction

Welcome to the manual for MLighter, your holistic tool to test machine learning implementations.

There are two different options for testing:

- **Code Blindspots** for testing your code.
- **Model Blindspots** for testing your model.



MLighter main menu

User Interface

After connecting to the MLighter server, the main menu is shown in your browser. There are four main separated areas:

- The **red area (1)** is the main option title, where the selected tab option is showen, This corresponds to either the Code Blind Spot area, or the Model Blind spot area.
- The **Magenta area (2)** is the tab option title where the current tab selection is shown for each specific area. Here you can also find the help button, this can be found on all pages if you need some guidance on what to do.
- The **blue area** is the side bar where the main navigation options are shown.
- The **green area** is the tab and working area where any sub-navigation is shown, and where you will be able to upload, run, and see the results of your testing.

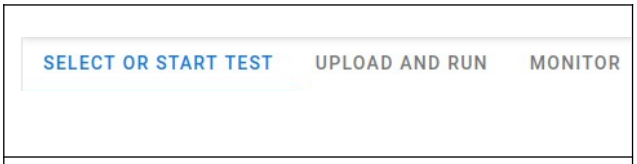
Main user interface

Code Blindspots

Code Blindspots allows you to test your code or review your code using our database of vulnerabilities.

Code Testing

Code Testing has three views: **Select or Start Test**, **Upload and run**, and **Monitor**



Select or Start Test

In the **Test Name** box, you can provide a memorable name for your next test.

In the **Reconnect Box**, you can select a previously run test to view its details.

If you wish to reset your view and begin a new test, click the **Start a New Test** button.

Upload and run

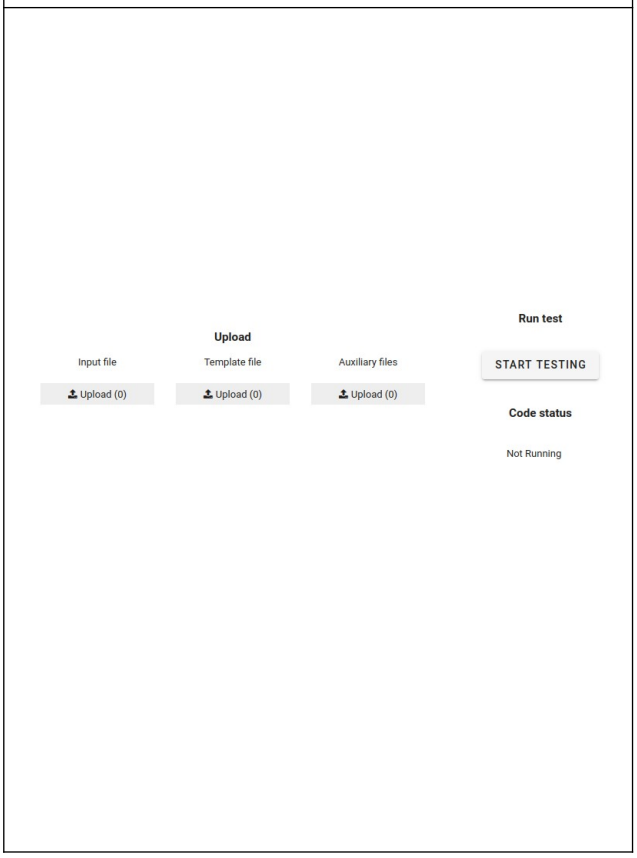
Here you can upload the files for your test and start the testing.

Input File is used to upload the seed data for the test, this should be some valid data for your template.

Template File is used to upload the file to test, this can be either Python or R, and should follow the format of the examples.

Auxiliary Files is used to upload any additional files needed for your test, these will be placed in the same folder as your test.

When you have uploaded everything you need click the **Start Testing** button to begin.



Monitor

In the monitor view, you can view the progress of your tests.

You can view **Crashes**, **Hangs**, **Executions**, and **Paths.**

Crashes are the number of times your program has crashed through the run.

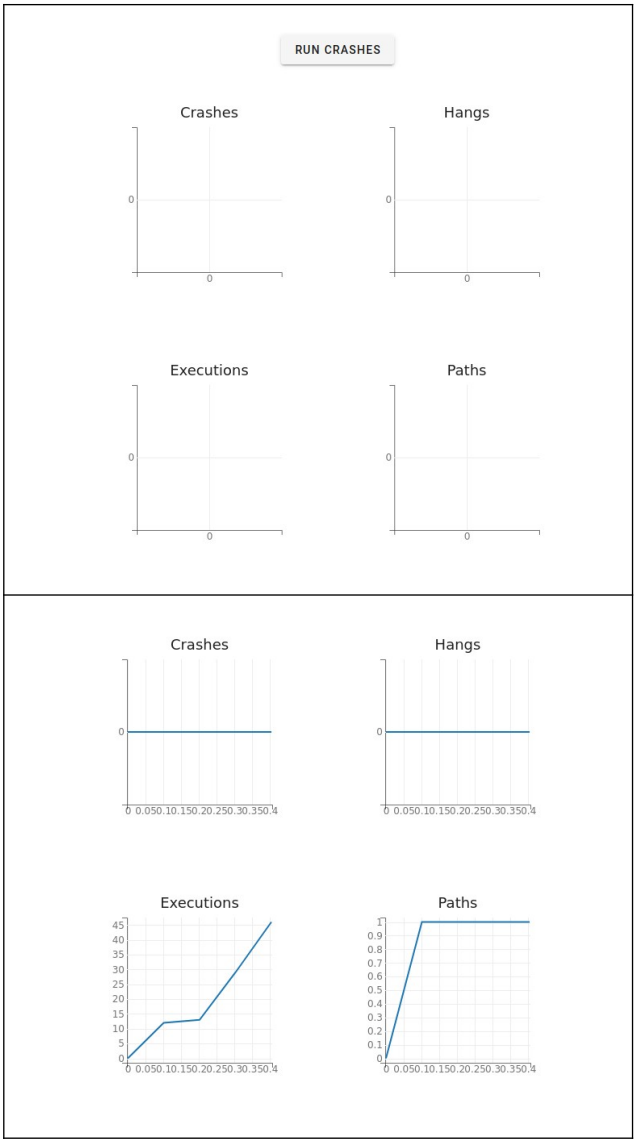
Hangs are the number of times your program has taken longer than the timeout to run.

Executions are the total number of times your program has been run.

Paths are the different paths taken through your program.

For details on what exactly each of these may mean, please refer to AFL++ documentation.

The **Run Crashes** button is used to run the crashes that have occurred so far again to provide more details.

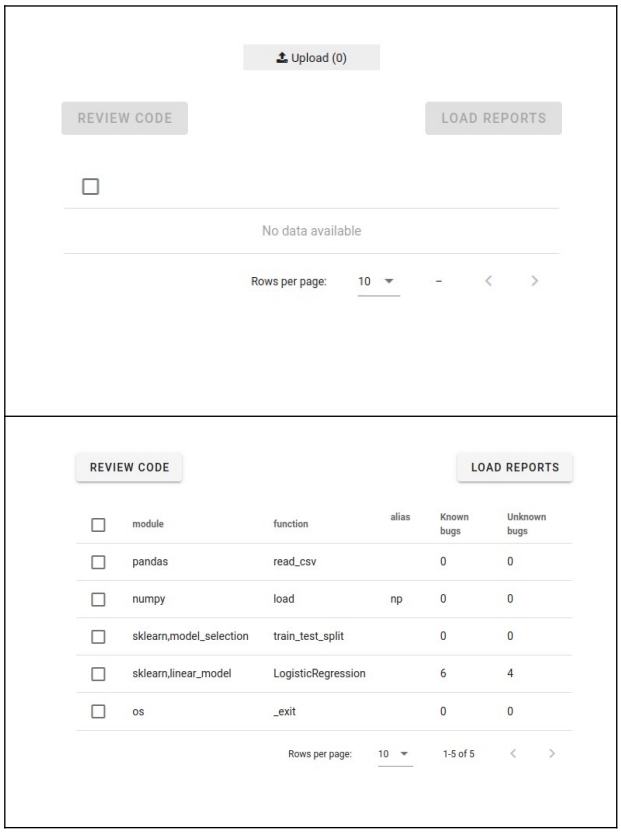


Code Review

Code Review allows you to check your code against our database of potential issues and view details about these.

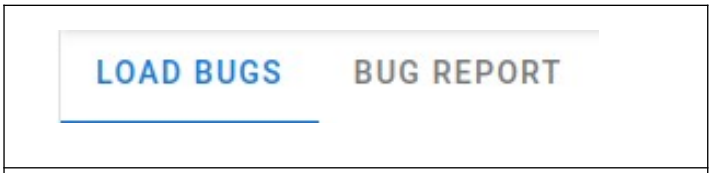
Upload your code to review into the **Upload** box, and when ready select **Review Code**. This will scan your code for any potential issues we know about.

You can select issues in the table, and **Load Reports** to find out more.



Bug Report

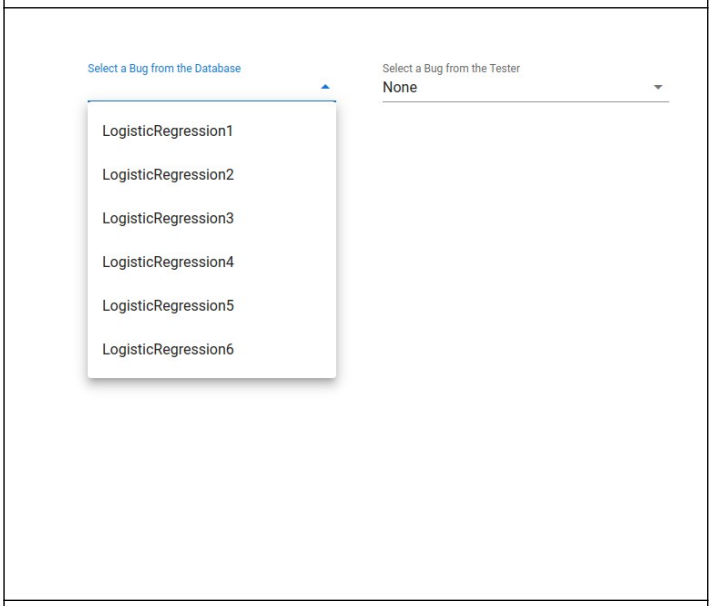
Bug Report has two views: **Load Bugs**, and **Bug Report**



Load Bugs

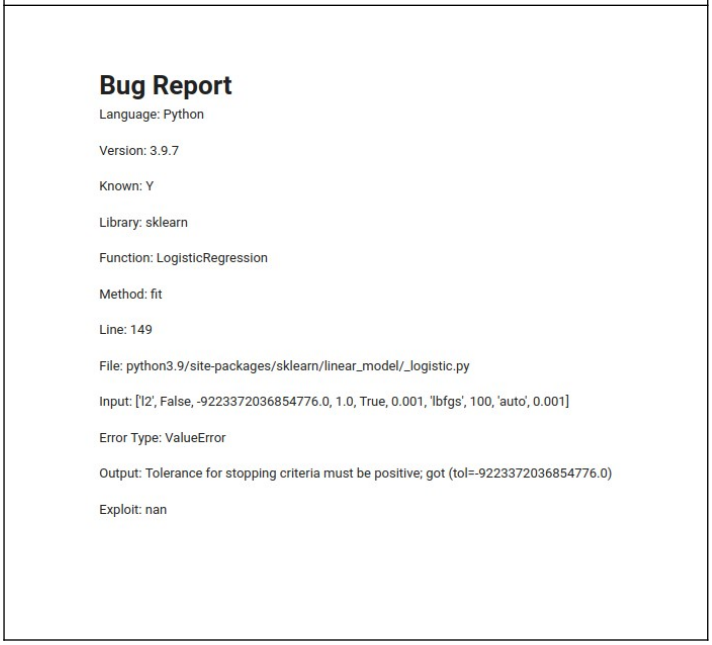
In the **Select a Bug From The Database** box you can select a bug to view from the **Code Review** area. This will then allow you to view it in the **Bug Report** tab.

In the **Select a Bug From The Tester** box you can select a bug to view from the **Code Testing** area. This will then allow you to view it in the **Bug Report** tab.



Bug Report

Here you can view details of the bug, such as the language, version, whether it is a known bug or not, and more specifics about the bug and inputs which created it.



Model Blindspots

Model **Blindspots** enables you to evaluate your model's performance against specific adversarial scenarios. MLighter will seek to pinpoint data variations that challenge your model according to your specified criteria, helping to assess its robustness.

Dataset

Dataset has three views: **Upload Data**, **Visualise Data**, and **Clean and Select Features**

Upload Data

Here there are three options depending on the kind of data you have.

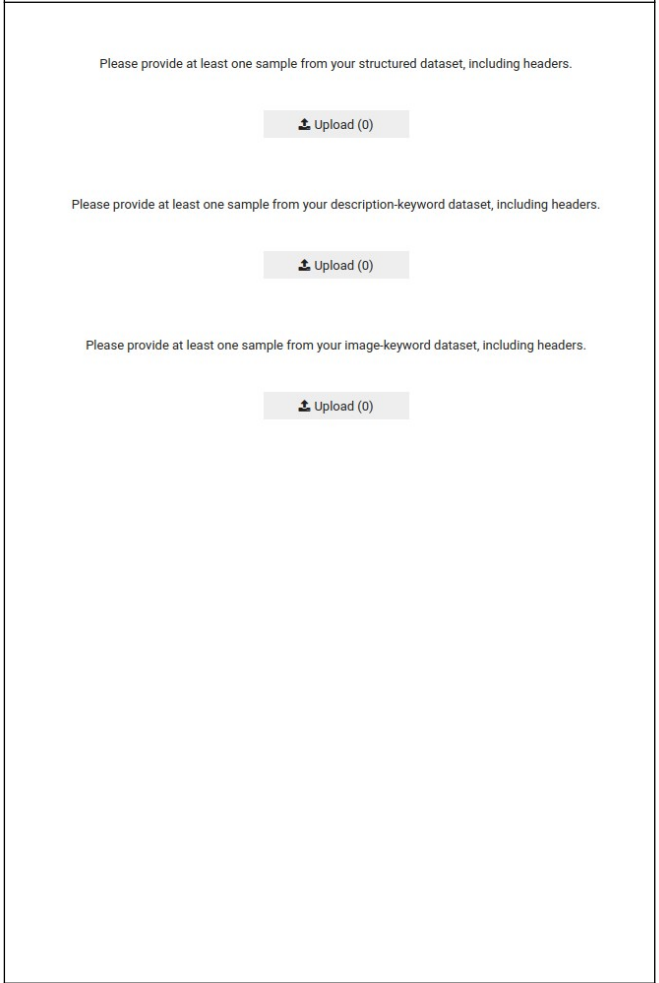
Structured: This is typically used for models which predict based on multiple features.

Description-Keyword: This is for models where you have text descriptions, and a list of possible keywords you expect it to predict.

Image-Keyword*: This is for models where you have Image URLs, and a list of possible keywords you expect it to predict.

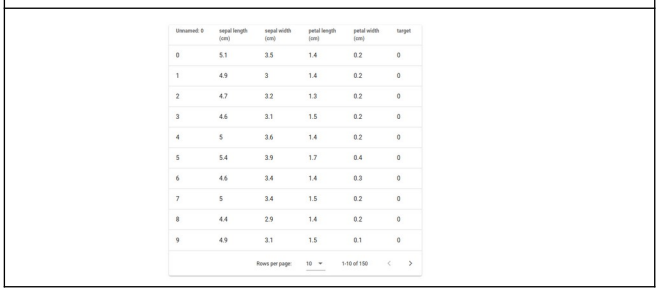
In all cases, please ensure you upload a **CSV** file with **headers** included. Check our examples for more guidance.

**Images are currently unsupported*



Visualise Data

Here you can view the data you have uploaded, and ensure everything is correct.

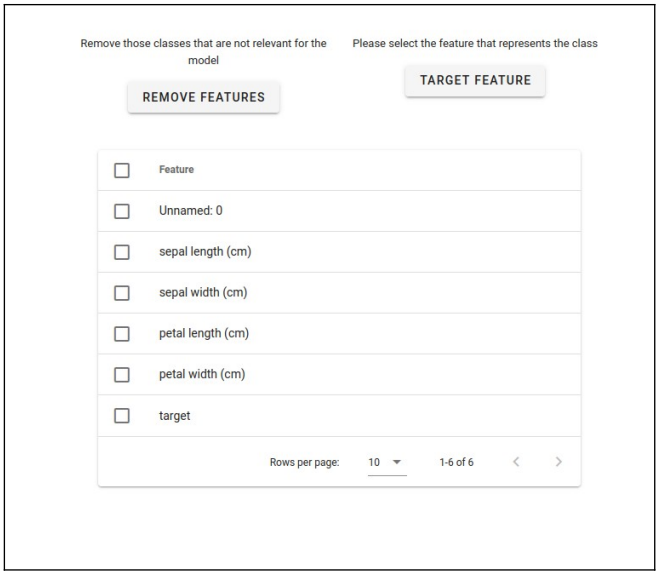


Clean and Select Features

Here you can view the features present in your dataset.

If you need to clean additional features not used by the model for prediction you can select it and click the **Remove Features** button.

You also need to select your target feature, to do this, select the feature and press the **Target Feature** button.



Model

Model is where you will upload the model you would like to test. This can either be an **SkLearn** model, or a **hugging face** model*.

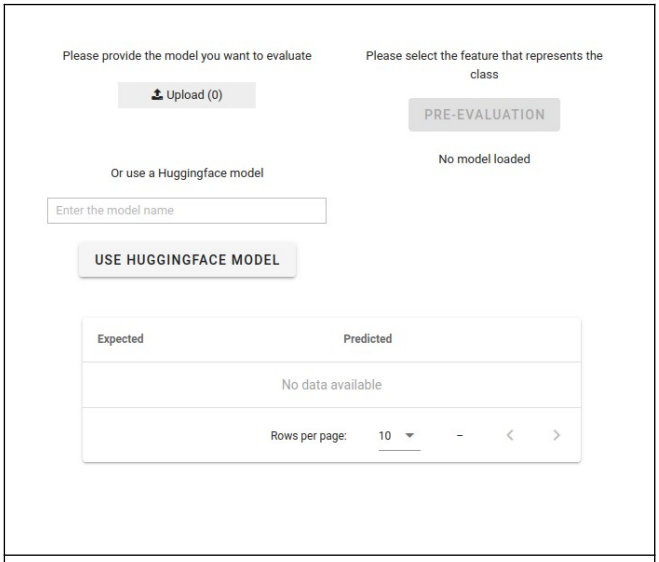
If you would like to use an **SkLearn** model, use the **Upload** box. To use a **hugging face** model copy the model name into the **HuggingFace model** box, and press the **Use HuggingFace Model** button.

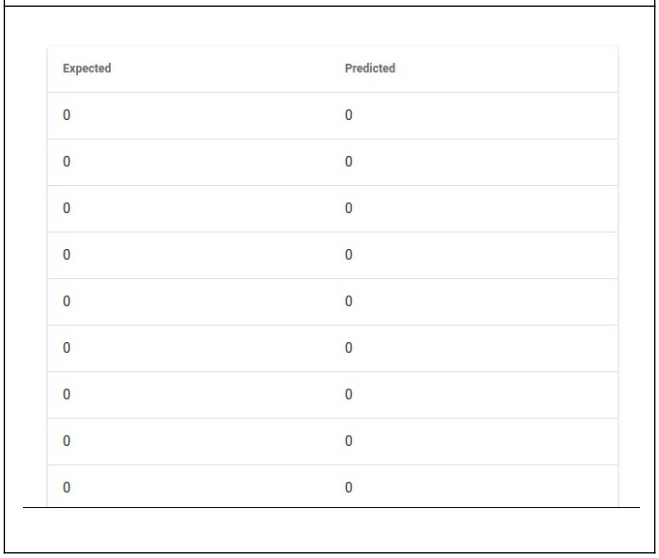
The text below the **Pre-Evaluation** button will update when the model has successfully loaded.

If you would like to ensure everything is working correctly, press the **Pre-Evaluation** button when the model has loaded.

Pre-evaluation will show you information about expected and predicted classes from the model.

**Currently we can only handle models on hugging face for description classification*



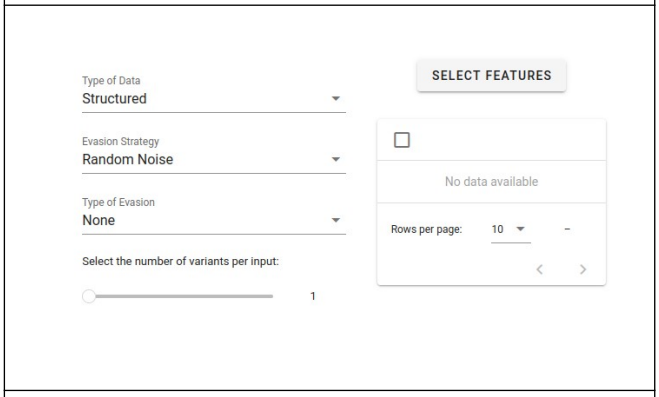


Evasion

Evasions has three views: **Evasion**, **Run**, and **Individual Report**

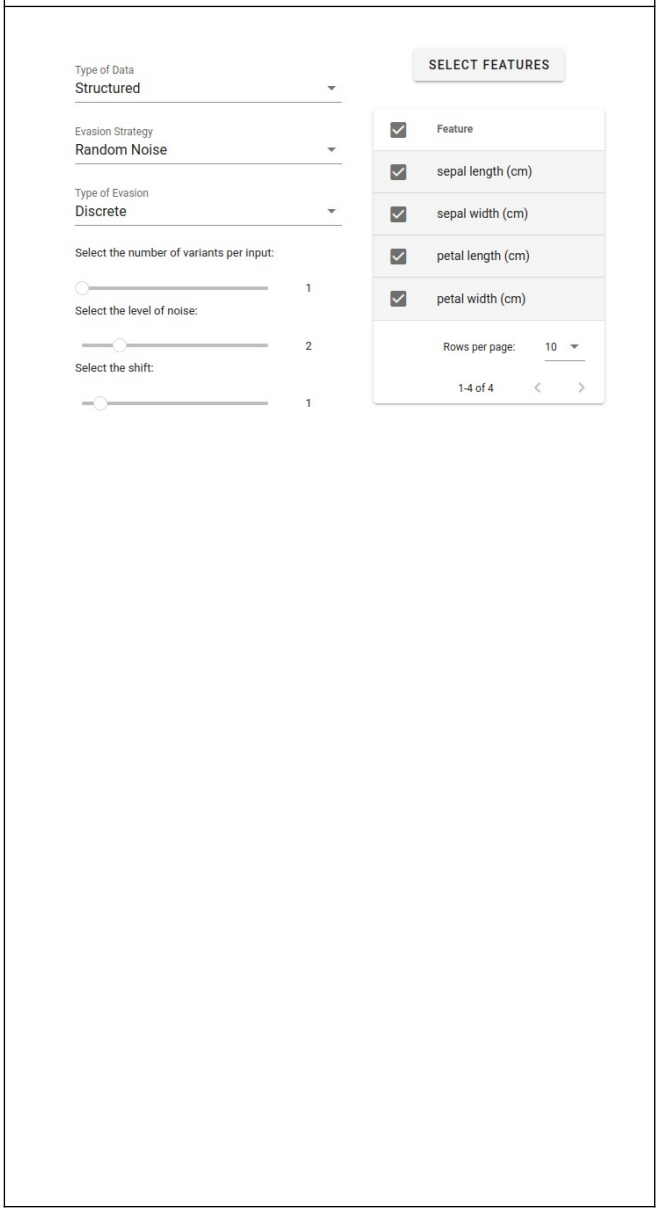
Evasion

Here you can define the evasion strategy, and type you want to perform.



The **Type of Data** drop-down allows you to select whether you have structured data or description data, as different mutations may be applied to each. Please ensure you've selected the correct one for your data.

The **Evasion Strategy** drop-down is to select what strategy you would like to perform, only **Random Noise** works in the Lite version.



The **Type of Evasion** drop-down allows you to select between discrete and continuous transformations on structured data, or the possible mutations available on description data.

Once you've selected your **Evasion type**, and **Strategy**, you can tune it further, such as the **number of variants per input***, and the **level of noise**, or **shift** for the Noise transformations.

You can then select the features you want to apply this transformation to, and click the **Select Features** button.

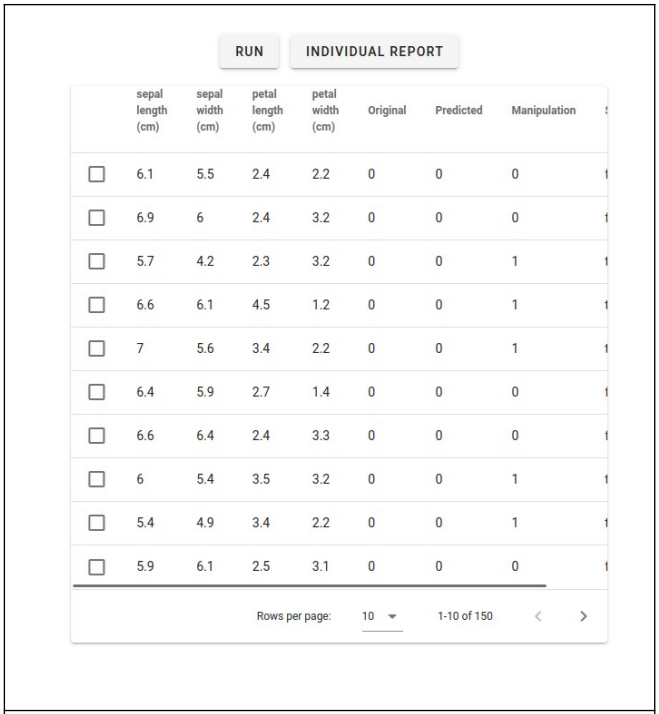
**Currently, we can only handle one variant per input*

Run

In order to run the **Evasion** you just created click the **Run** button.

Once you run, you will see a table populated with data about the evasion, such as the features, original prediction, new prediction, and whether the evasion was successful or not.

If you would like to investigate a specific variant, you can select it in the table, and click the **Individual Report** button to generate a report for that variant.



Individual Report

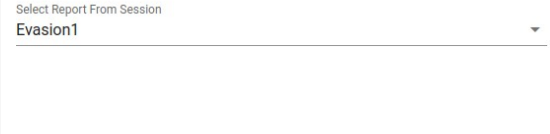
Here you can view data about the transformation you performed for the selected variant in a more focused view.

Report

Report has two views: **Selection**, and **Report**

Selection

Here you can select which report you want to see from all the evasions you have ran by picking it in the **Select Report From Session** drop-down.



Report

Here the report you have selected can be viewed, It will have details about the evasion used, as well as a **heat maps** of the predicted class misclassifications before and after the variants were generated.