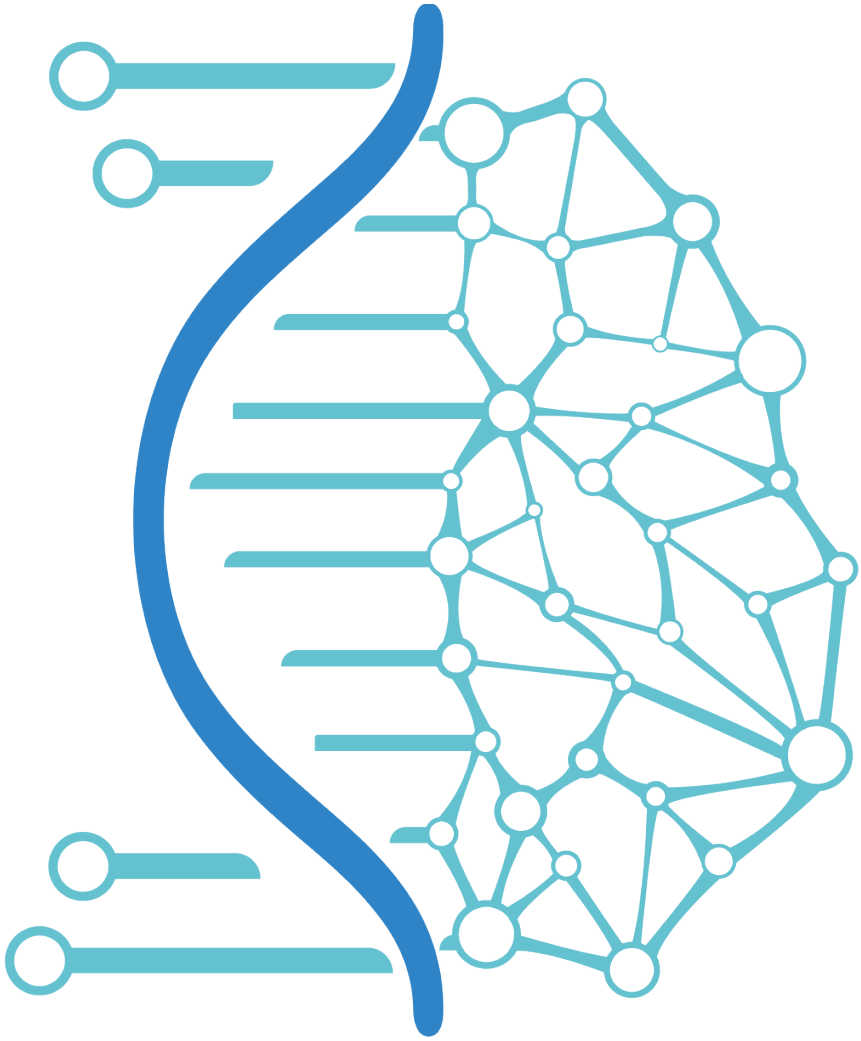

GeneEssenceGUI

User Guide - Desktop Version 1.0

1. Introduction
2. Python Installation
3. Running the software
4. Dataset Model
5. The software
5.1 Main Window
5.2 Analysis Type Selection Window
5.3 Training: How to train the models?
5.3.1 Model Selection Window
5.3.3 Evaluation metrics selection window in training
5.3.4 Results of the Receipt Method Selection Window in Training
5.3.5 Confirmation of Provided Information Window in Training
5.3.6 Project Execution Window in Training
5.3.7 Training Analysis Results
5.4 Prediction: How to perform a prediction?
5.4.1 Prediction Parameters Definition Window
5.4.2 Results of the Receipt Selection Window in Prediction
5.4.3 Information Confirmation Window in Prediction
5.4.4 Prediction Analysis Results
5.5 How to perform an ensemble analysis?
5.5.1 Ensemble Parameters Definition Window
5.5.2 Evaluation metrics selection window in the ensemble
5.5.3 Results of the Receipt Method Selection Window in the Ensemble
5.4.5 Project Execution Window in the Ensemble
5.3.6 Ensemble Analysis Results
5.4 Load information from existing projects
5.6 How to prepare your dataset

1. Introduction

Welcome to the user manual for GeneEssenceGUI, a software developed for the analysis of essential genes in prokaryotes. This tool helps to identify and classify essential genes and provides the user with an intuitive and efficient interface.

GeneEssenceGUI was developed in Python 12 and is compatible with Linux, macOS, and Windows operating systems. Validation tests were performed with Linux, macOS, and Windows distributions.

---

2. Python Installation

To run GeneEssenceGUI, you must install Python 3.12 or higher on your system. The installer can be obtained from the official website: Download Python.

Windows Installation

1. Download the Python installer (python-12.x.x.exe) from the official website.
2. Run the installer and check "Add Python to PATH" before proceeding.
3. Click "Install Now" and wait for the installation to complete.
4. Verify the installation by opening the command Prompt and running:

python --version

Linux Installation (Ubuntu/Debian-based)

1. Update the package list: s**udo apt update && sudo apt upgrade -y**
2. Install Python using apt: **sudo apt install python3**
3. Verify the installation: **python3 --version**

macOS Installation

1. Install Homebrew (if not already installed): **/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"**
2. Install Python using Homebrew: **brew install python**
3. Verify the installation: **python3 --version**

---

3. Running the software

The GeneEssenceGUI executables are available for download at the following link: https://github.com/allanverasce/geneessencegui/releases/tag/v1.0.0

After downloading, follow the instructions according to your operating system:

Linux and macOS

1. Grant execute permissions to the downloaded file:
   chmod 777 GeneEssenceGUI
2. **For Linux:** Double-click the file or run it in the terminal:
   ./GeneEssenceGUI
3. **For macOS:** It is necessary to execute it through the terminal:
   ./GeneEssenceGUI
4. **Windows:** Right-click on the executable and select "***Run as administrator***". This ensures the software has the necessary permissions to access system resources, preventing execution errors and enhancing compatibility with dependencies.

---

4. Dataset Model

For GeneEssenceGUI to function correctly, the dataset shown in Figure 1 must be in CSV (Comma-Separated Values) format. This format organizes data into rows and columns, each representing a data instance and containing a characteristic gene variable.

To simplify data preparation, we provide a processing script that converts your files into the format the software accepts. More details are described in section 5.6. You can access it here: https://github.com/allanverasce/GeneEssenceGUI.

M	F	L	I	V	S	P	T	A	Y	H	Q	N	K	D	E	C	W	R	G	Product Name
3	7	38	16	13	12	17	20	30	8	4	19	1	6	14	16	5	7	15	11	aadA
3	7	37	16	13	13	18	22	27	8	4	19	1	6	13	16	5	7	15	12	aadA
3	7	37	16	15	14	17	20	29	7	5	17	1	7	17	14	3	7	15	11	aadA
3	7	37	16	13	12	18	20	30	8	4	19	1	6	14	16	5	7	15	11	aadA
4	7	37	15	13	12	18	20	30	8	4	19	1	6	14	16	5	7	15	11	aadA
4	7	37	15	13	13	18	22	27	8	4	18	1	8	15	15	5	7	14	11	aadA
4	7	37	16	14	13	18	19	28	8	4	19	1	6	13	17	5	7	15	11	aadA
3	7	37	16	13	14	17	20	29	8	4	19	1	6	13	17	5	7	15	11	aadA
3	7	37	16	15	13	17	20	29	7	5	17	1	7	17	14	4	7	15	11	aadA
4	7	37	15	13	12	18	22	28	8	4	19	1	7	12	17	5	7	14	12	aadA
3	7	37	16	13	13	18	22	27	8	4	19	1	6	14	16	5	7	15	11	aadA
4	7	37	15	13	13	18	22	27	8	4	19	1	6	14	16	5	7	15	11	aadA

Figure 1. Dataset Format.
Source: Created by the author.

Note 1: For the prediction step, the dataset must be submitted without the "Product Name" column.

---

5. The software

5.1 Main Window

When you start running the software, the GeneEssenceGUI main window appears (Figure 2). The window contains a welcome message and a brief description of the tool's main

features, along with information about the partners involved in the project. To start using the software, simply click the **"Start"** button.

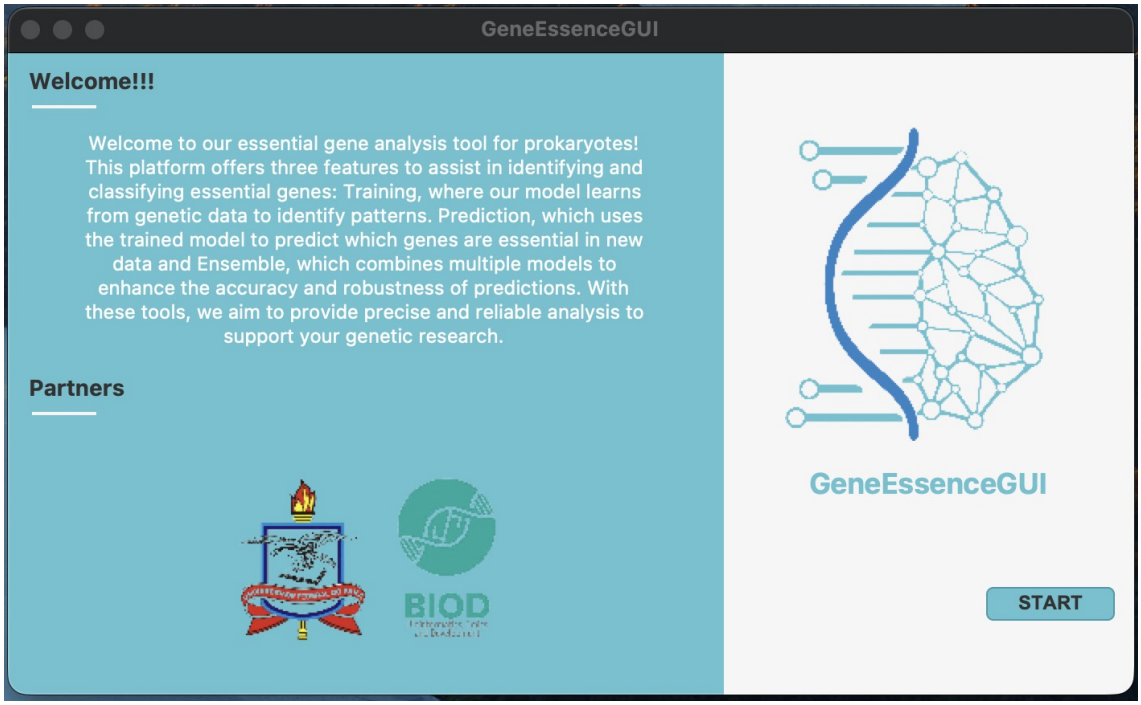

Figure 2. GeneEssenceGUI Main Window.
Source: Created by the author.

5.2 Analysis Type Selection Window

The next window shows the user the types of analysis available in the tool, as shown in Figure 3. At this stage, the user can choose between three analysis options, depending on their needs: training, prediction or ensemble.

- **Training:** During the training process, the machine learning model is trained using a data set. During this step, the model analyzes the data provided to identify patterns and learn how to make predictions based on them.
- **Prediction:** Prediction is the phase in which a previously trained model is used to analyze new data. Based on the knowledge acquired during training, the model concludes, in this case, the classification of essential genes in unknown data.
- **Ensemble:** Ensemble refers to an advanced technique in which multiple machine learning models are combined to improve the performance of the resulting model and thus the evaluation metrics. This approach reduces the risk of overfitting and makes the model more robust. By combining different analysis methods, a more reliable performance is ensured.

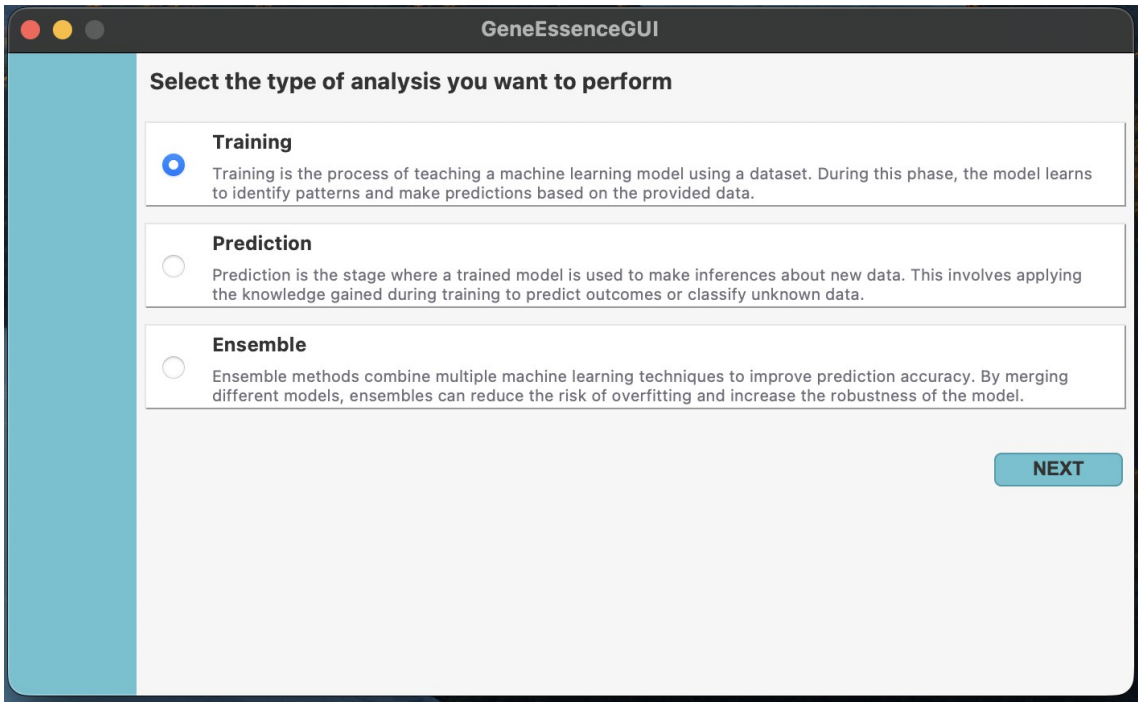

Figure 3. Analysis Type Selection Window.
Source: Created by the author.

5.3 Training: How to train the models?

On your first access, you will be shown a window for creating a new training project (figure 4), in which you will have a field to provide the project's identification name and then select the tabular dataset file for training.

Finally, the user will be able to define the proportion of the dataset to be used for testing the models during this stage, the percentage of this dataset being set directly in the interface.

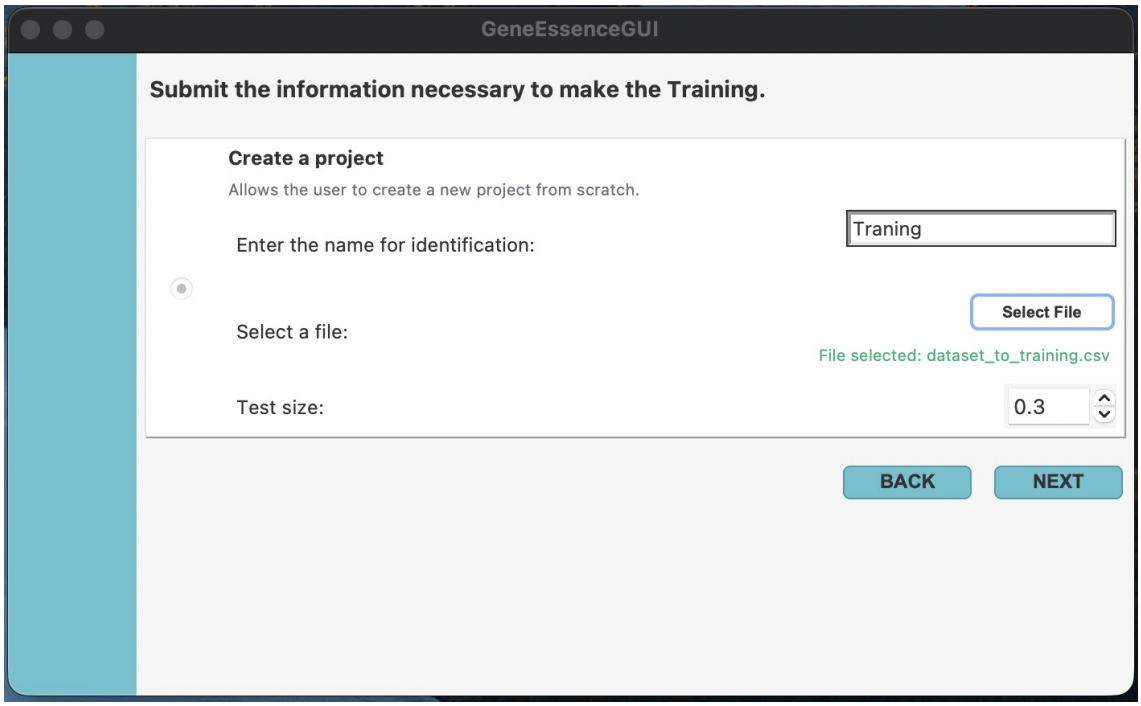

Figure 4. Training Parameters Definition Window.
Source: Created by the author.

5.3.1 Model Selection Window

The next window (Figure 5) allows the user to select the models to be trained. There are 8 types of models available, and the user can choose one or more of them.

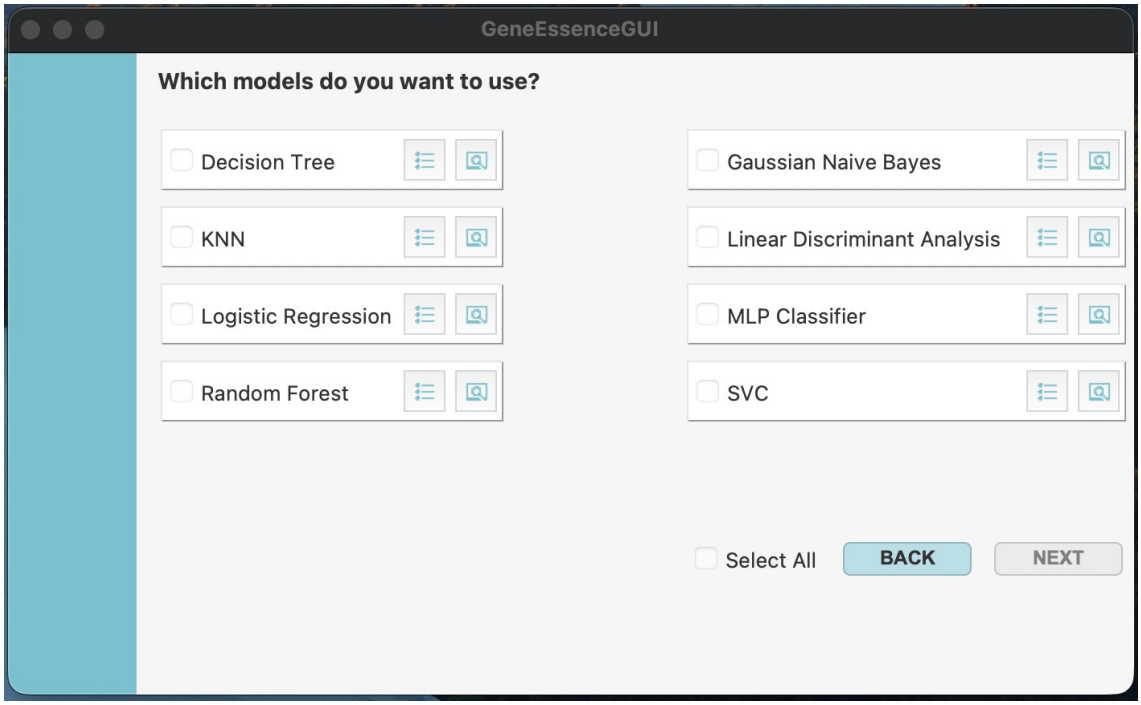

Figure 5. Model Selection Window for Training.
Source: Created by the author.

Next to each model, there are two icons: the first opens the model parameters window, allowing the user to define each parameter or leave the default values (Figure 6); the

second icon redirects the user to the model documentation, providing detailed information about it (Figure 7).

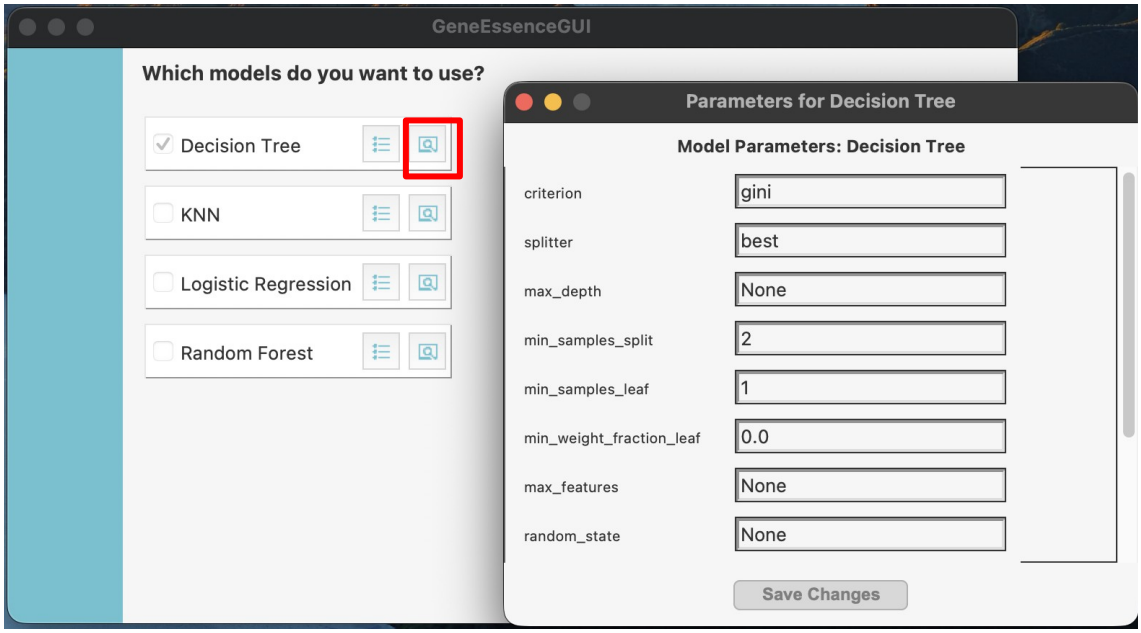

Figure 6. First icon: Parameters window for the DecisionTree model.
Source: Created by the author.

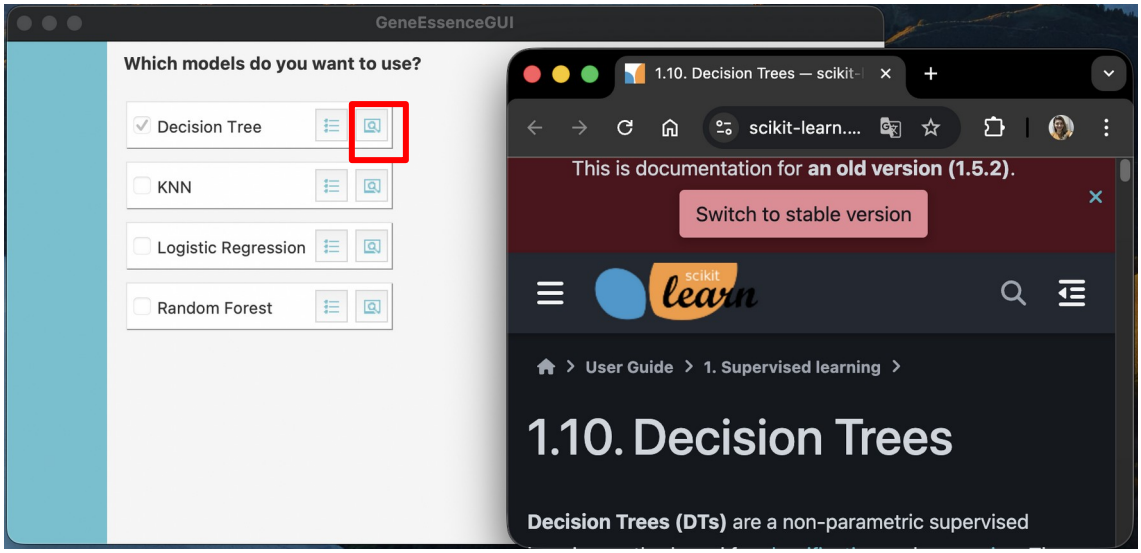

Figure 7. Second icon: Documentation for the DecisionTree model.
Source: Created by the author.

5.3.3 Evaluation metrics selection window in training

The next window (Figure 8) will allow the user to select the evaluation metrics for the models chosen in the previous step. The available metrics for this process include Accuracy, Kappa, Precision, F1 Score, Matthews Corrcoef, and Recall.

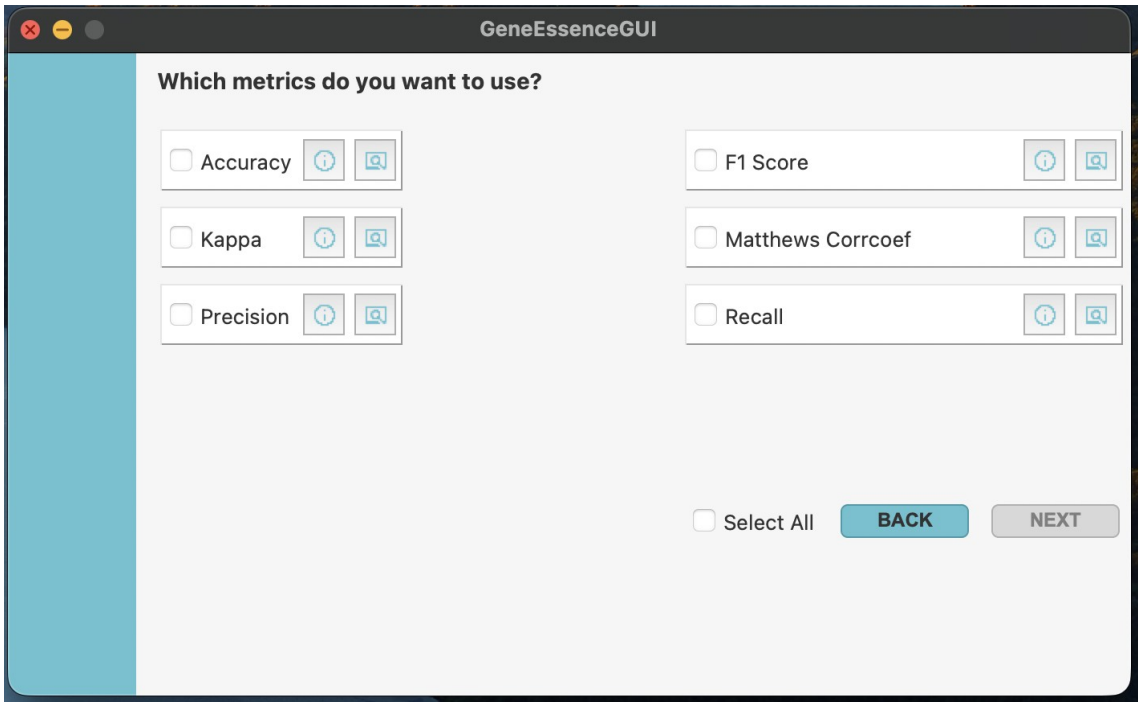

Figure 8. Evaluation Metrics Selection Window for the Selected Models.
Source: Created by the author.

Additionally, next to each metric, there are two icons: the first displays the definition of the metric to assist the user in their choice (Figure 9). The second redirect to the official documentation provides detailed information about how it works (Figure 10).

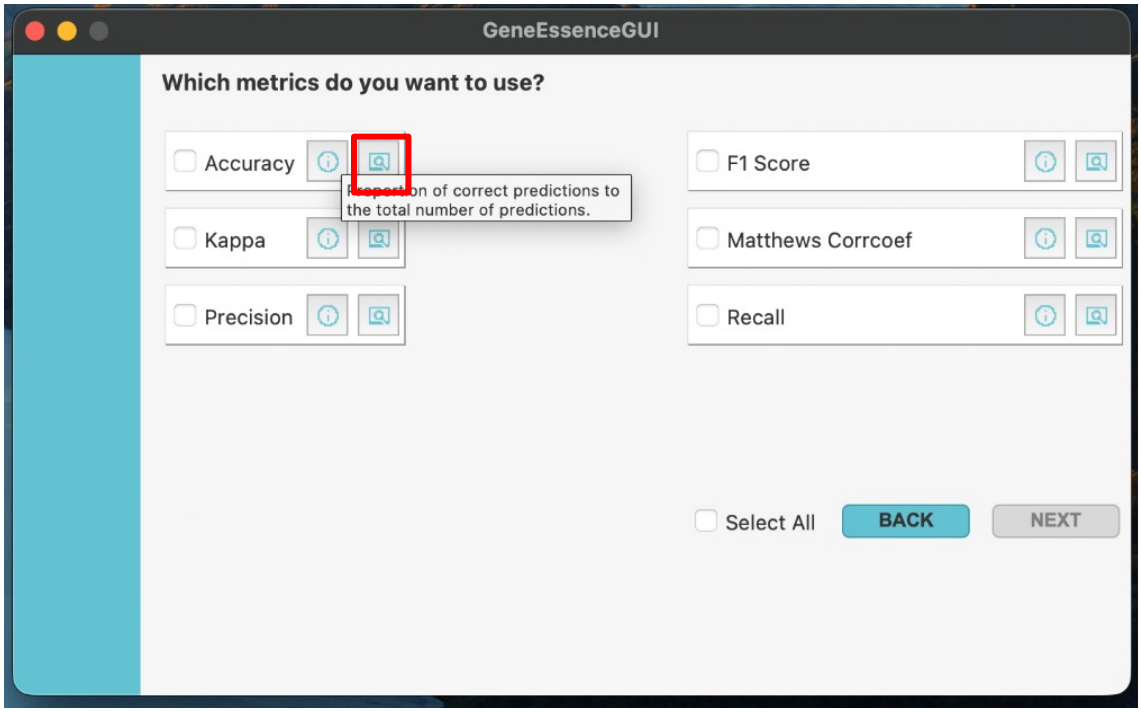

Figure 9. First icon: Information about the Accuracy evaluation metric.
Source: Created by the author.

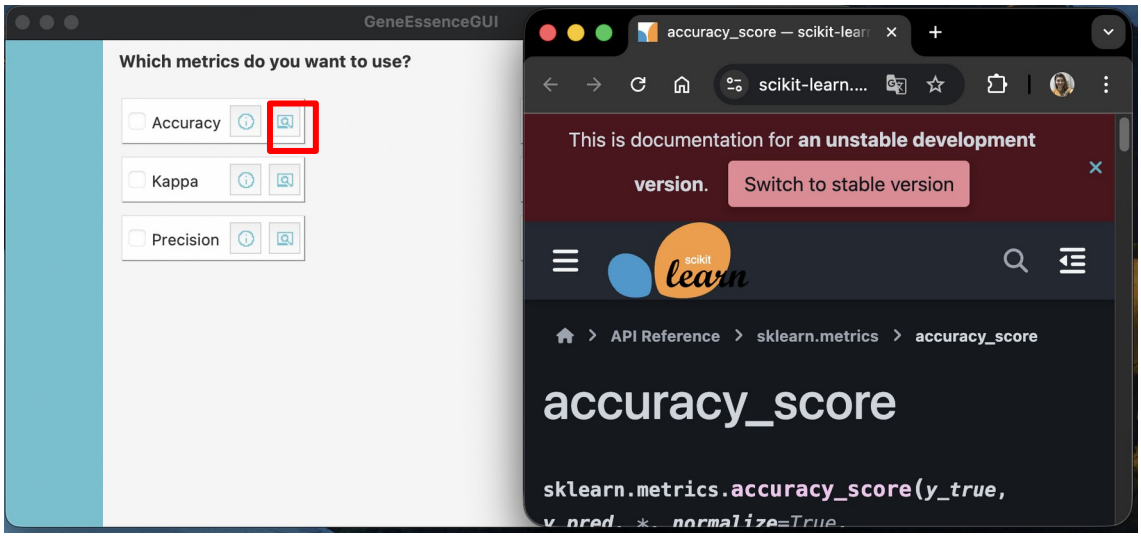

Figure 10. Second icon: Documentation for the Accuracy evaluation metric.
Source: Created by the author.

5.3.4 Results of the Receipt Method Selection Window in Training

After selecting the metrics, the user will need to click on the desired option to select the method for receiving the results (Figure 11). The available options are email delivery or local saving on the computer.

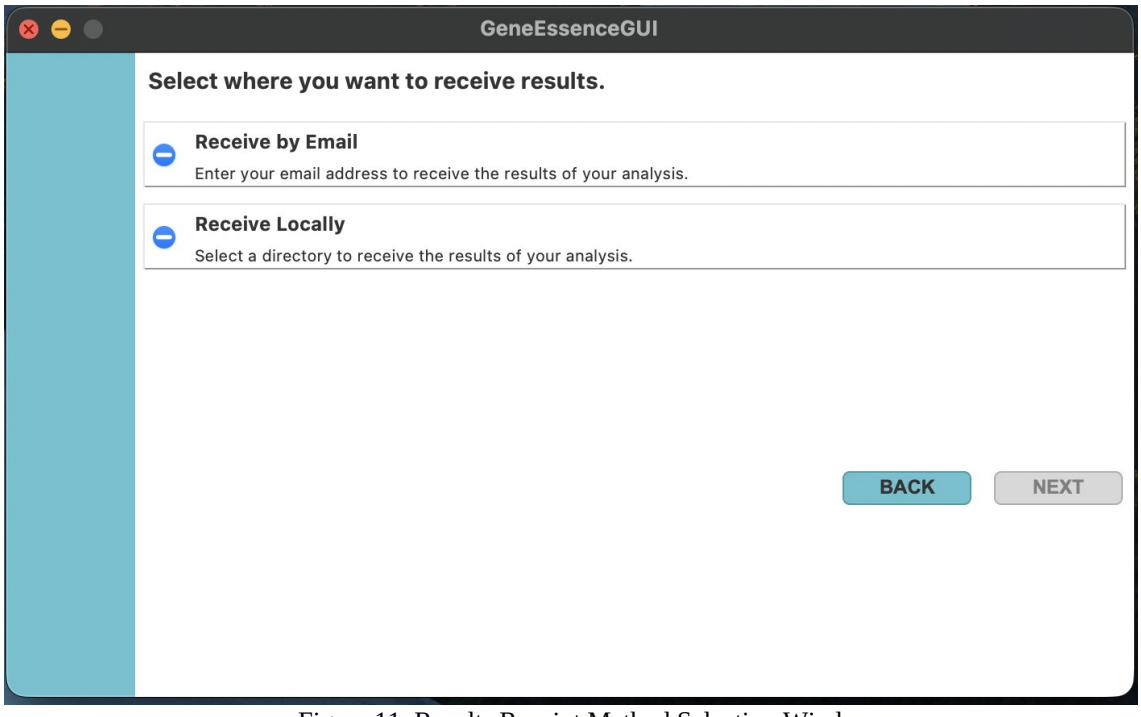

Figure 11. Results Receipt Method Selection Window.
Source: Created by the author.

If the user chooses email delivery, a text field will display to enter the email address (Figure 12). If the option is to save locally, the user will need to select the directory where the results will be stored (Figure 13). After filling in the required field, simply click "Next" to confirm and proceed to the next step.

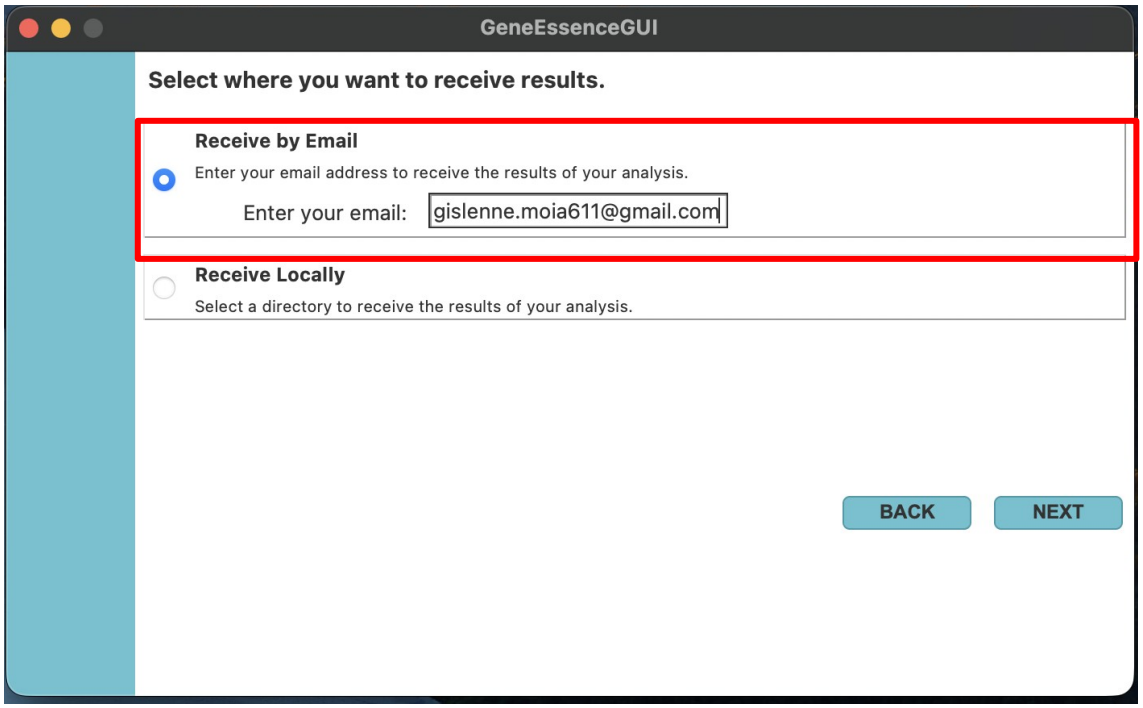

Figure 12. Selection of the email receipt option.
Source: Created by the author.

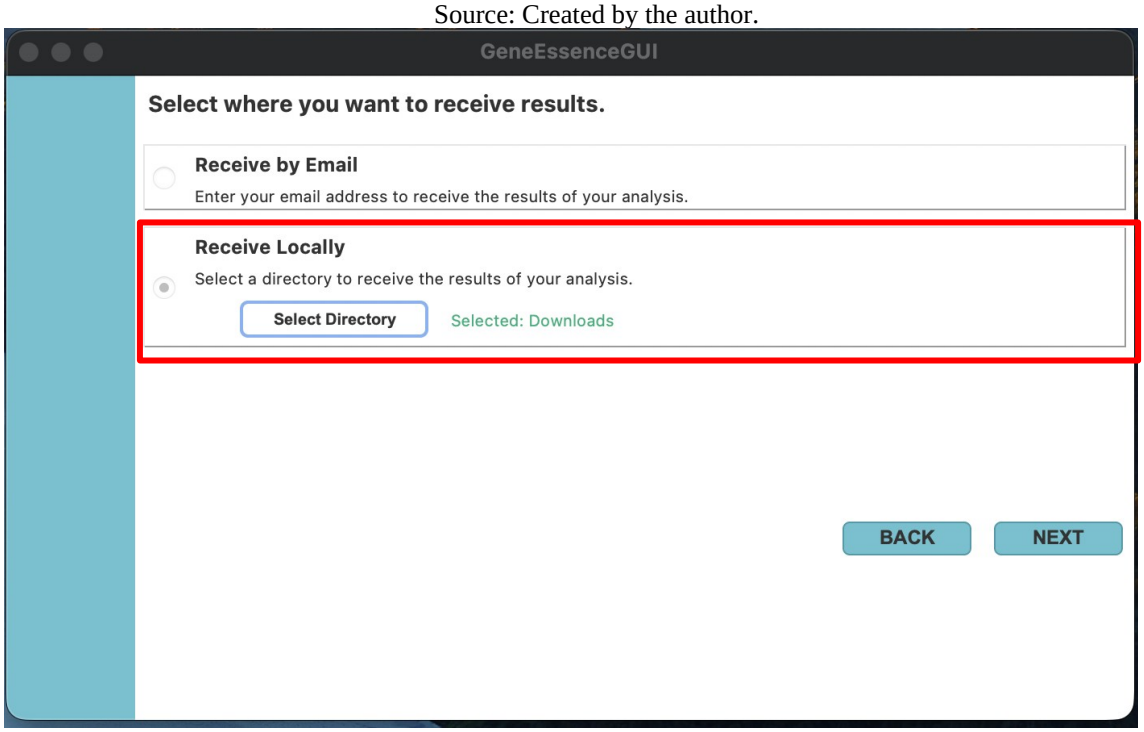

Figure 13. Selection of the local saving option.
Source: Created by the author.

5.3.5 Confirmation of Provided Information Window in Training

Before starting the analysis execution, the user will be asked to confirm the information provided during all previous steps (Figure 14). If everything is correct, simply click the confirmation button to begin the process.

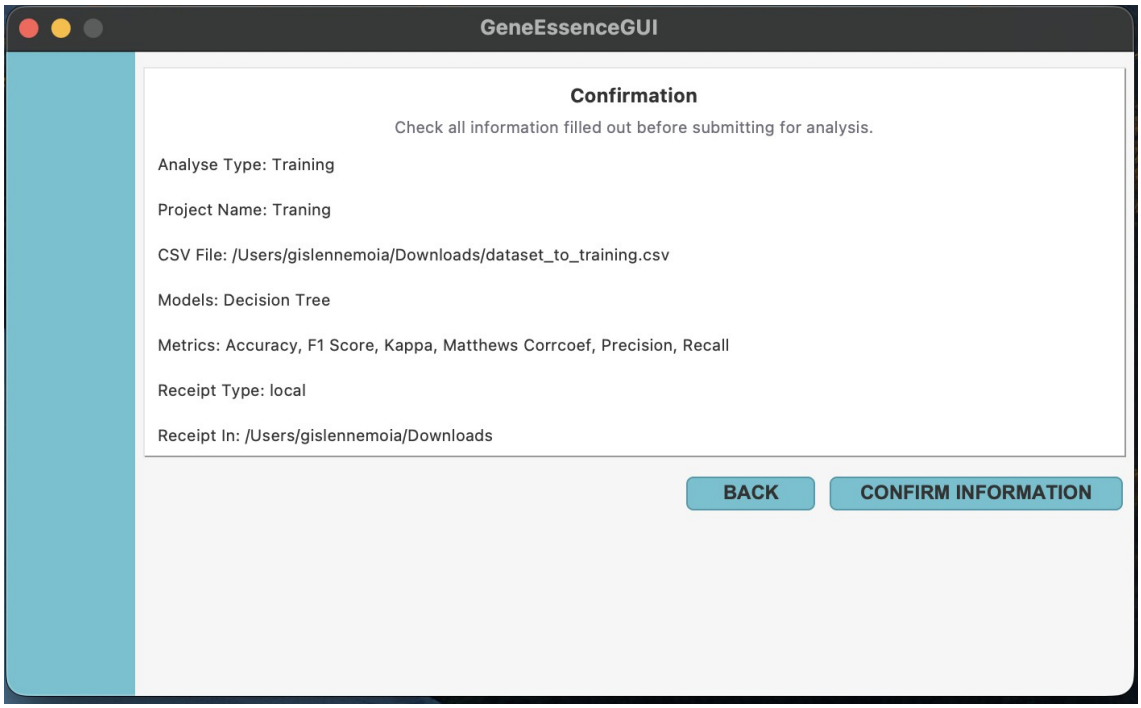

Figure 14. Confirmation of Provided Information Window.
Source: Created by the author.

5.3.6 Project Execution Window in Training

After confirming the project information, the execution window will be displayed, showing details of the steps already completed, as well as the percentage of completion of the training process for the selected models (Figure 15).

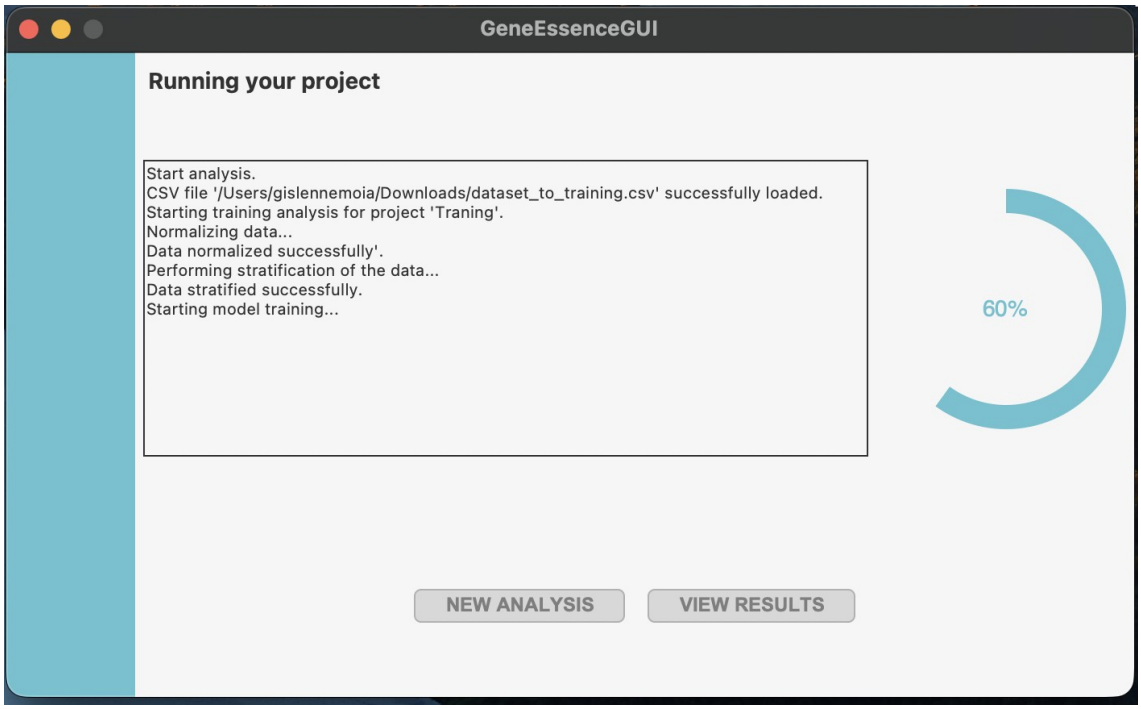

Figure 15. Project Execution Window.
Source: Created by the author.

Once the project execution process is complete, the user can view the full log of the process information. Additionally, the user can create a new project or view the results (Figure 16).

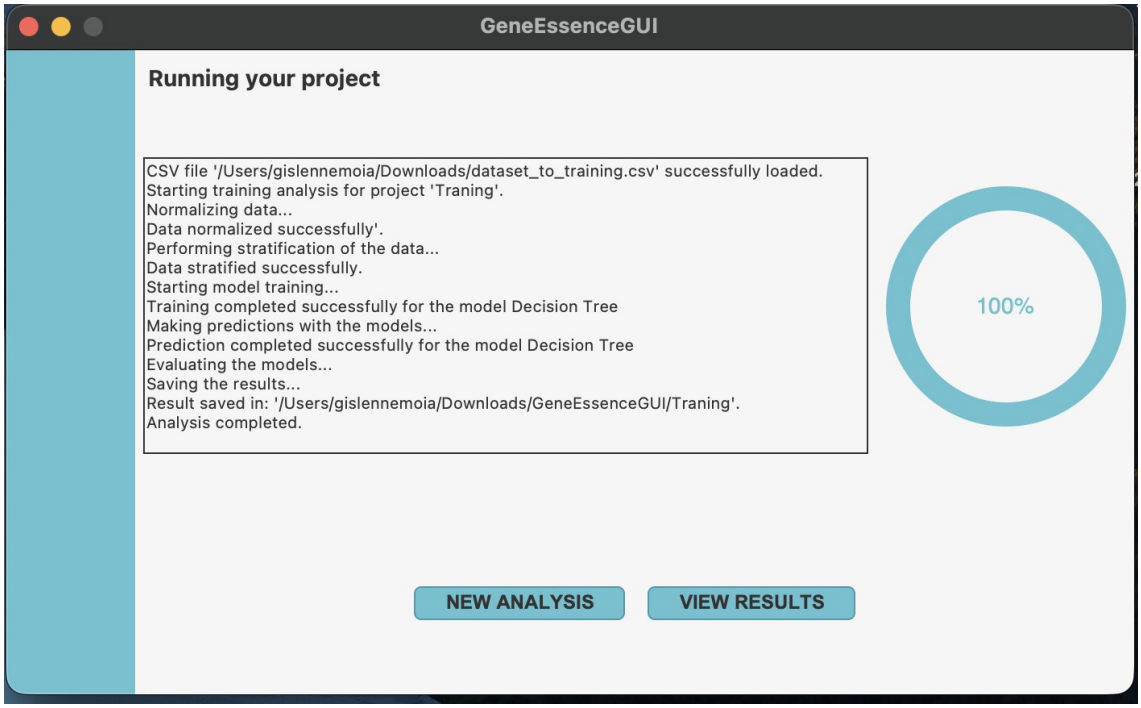

Figure 16. Project Execution Completion Window.
Source: Created by the author.

5.3.7 Training Analysis Results

In the training analysis, the user will receive the trained models based on their selection, along with a graph showing the evaluation of the models' performance. In addition, a CSV file will be provided with the genes predicted for each model using the test dataset.

5.4 Prediction: How to perform a prediction?

5.4.1 Prediction Parameters Definition Window

If it is the user's first access, a window will appear to create a new project. At this stage, the user will have to provide a unique name to identify the project, as well as upload a CSV file containing the tabular data together with a previously trained model in PKL format to carry out the classification of the genetic products.

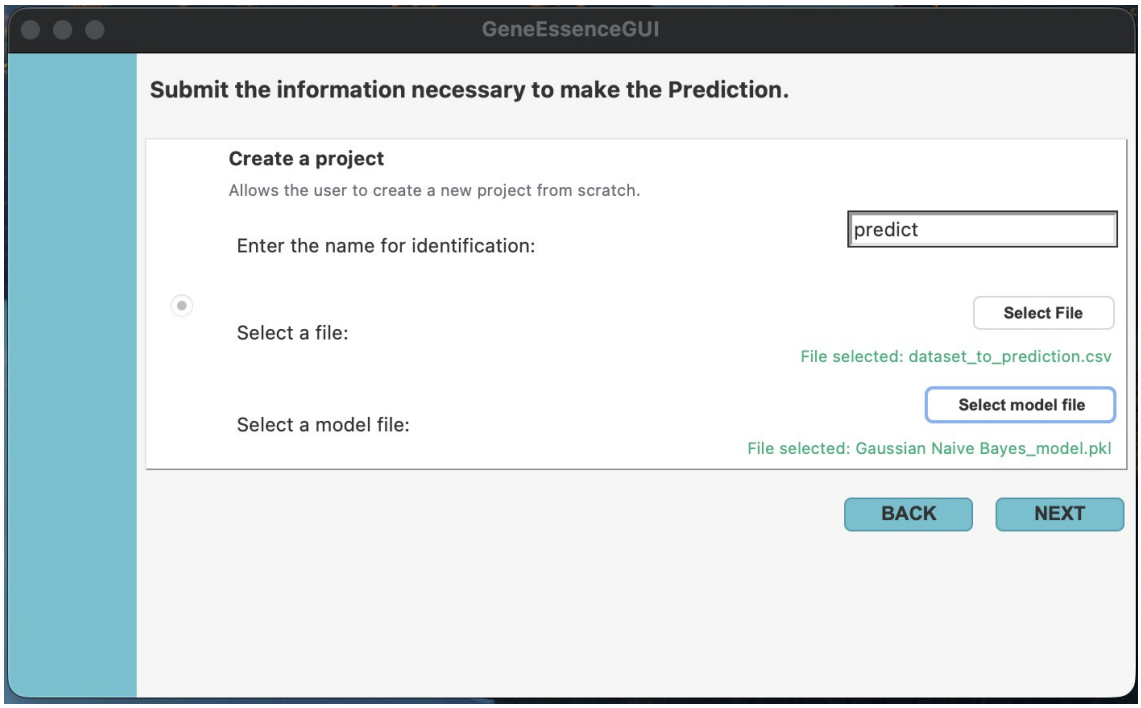

Figure 17. Prediction Parameters Definition Window.
Source: Created by the author.

Note 2: GeneEssenceGUI provides a previously trained model that is available for use in the repository. This model is the result of grouping the eight models provided with the tool. If the user chooses to use this model, a feasibility test will be performed to verify whether the user's hardware meets the minimum requirements to run it, which include at least 32 GB of RAM.

After adding the initial parameters for the analysis, proceed by clicking "*Next*."

5.4.2 Results of the Receipt Selection Window in Prediction

In the next window, the user must select the method for receiving the results. The software offers two options:

- By Email (Figure 18), where the user provides the email address to which the results will be sent.
- In Local Folder (Figure 19), the user selects a directory to save the results.

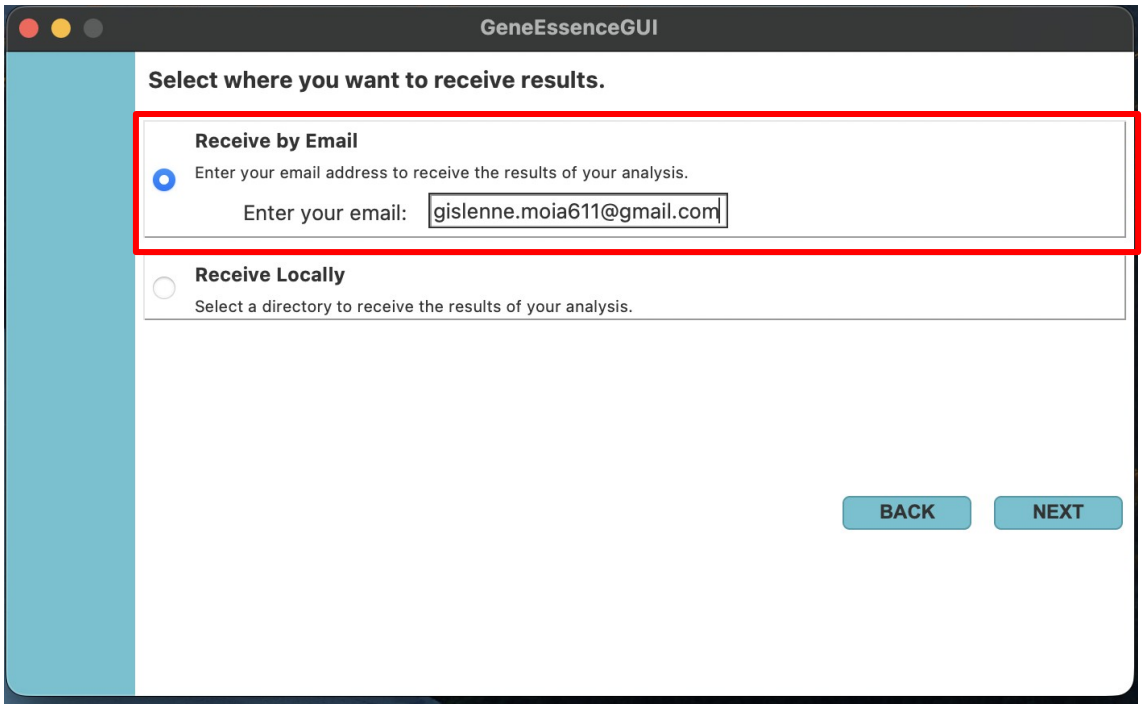

Figure 18. Selection of the email receipt option.
Source: Created by the author.

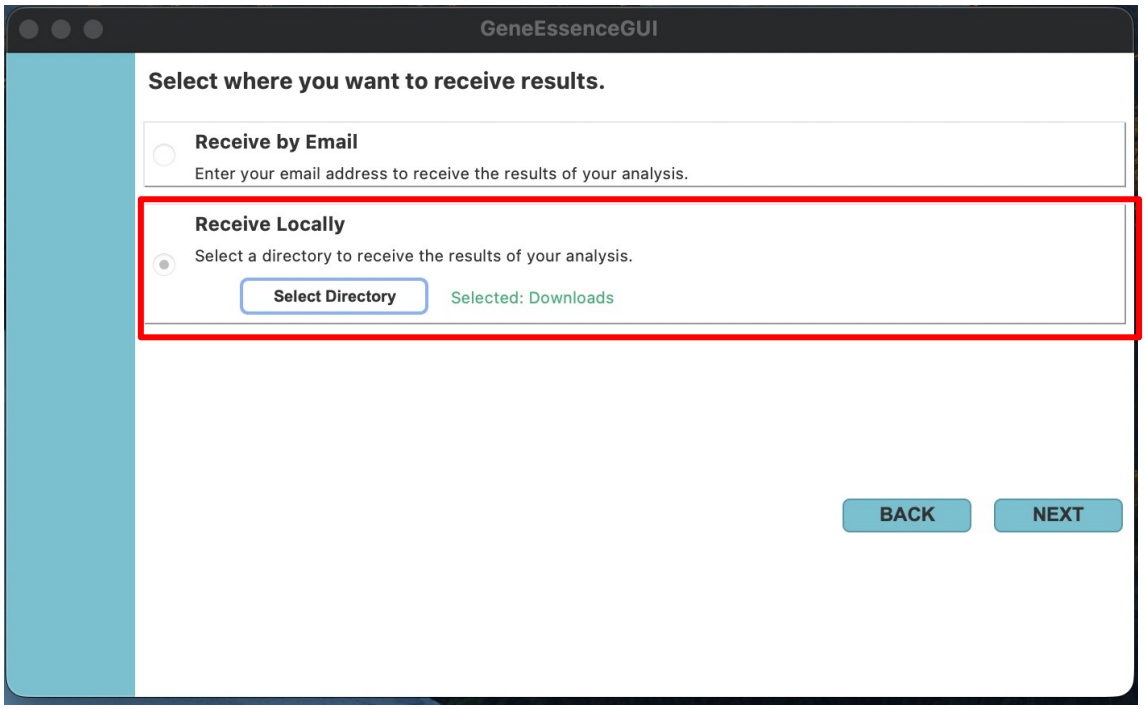

Figure 19. Selection of the option to save locally.
Source: Created by the author.

After choosing how to receive the results, advance to the next screen by clicking on "*next*".

5.4.3 Information Confirmation Window in Prediction

In the next step (Figure 20), all the provided information will be displayed for review. If everything is correct, simply click the Confirmation button to start the analysis.

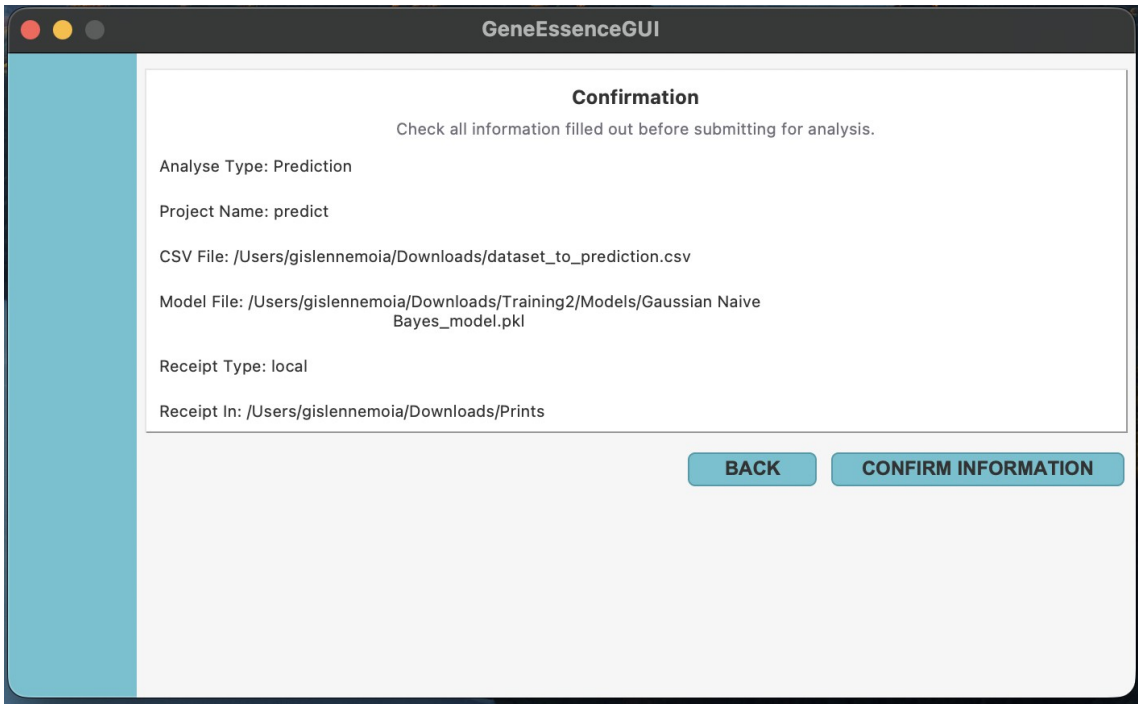

Figure 20. Information confirmation window.
Source: Created by the author.

During execution, it is possible to track the progress of the processing, as shown in the figure below.

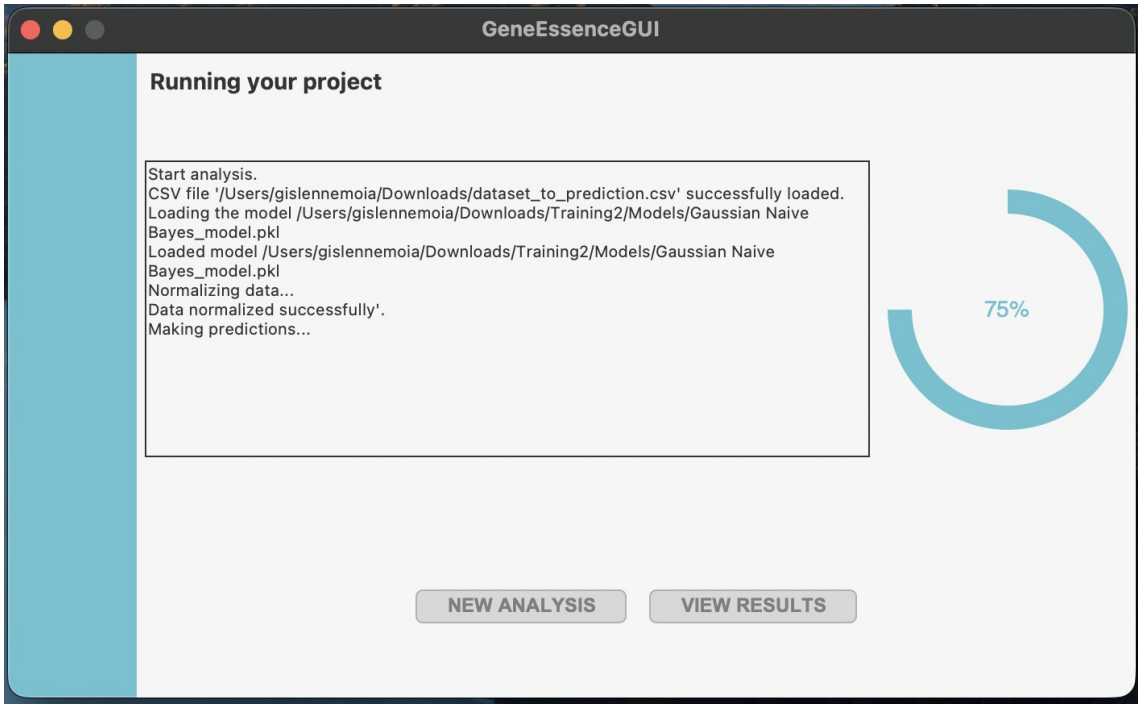

Figure 21. Project execution window.
Source: Created by the author.

At the end of the analysis (Figure 22), the results will be provided according to the option selected. In the same window, the user can view the results and start a new analysis if desired.

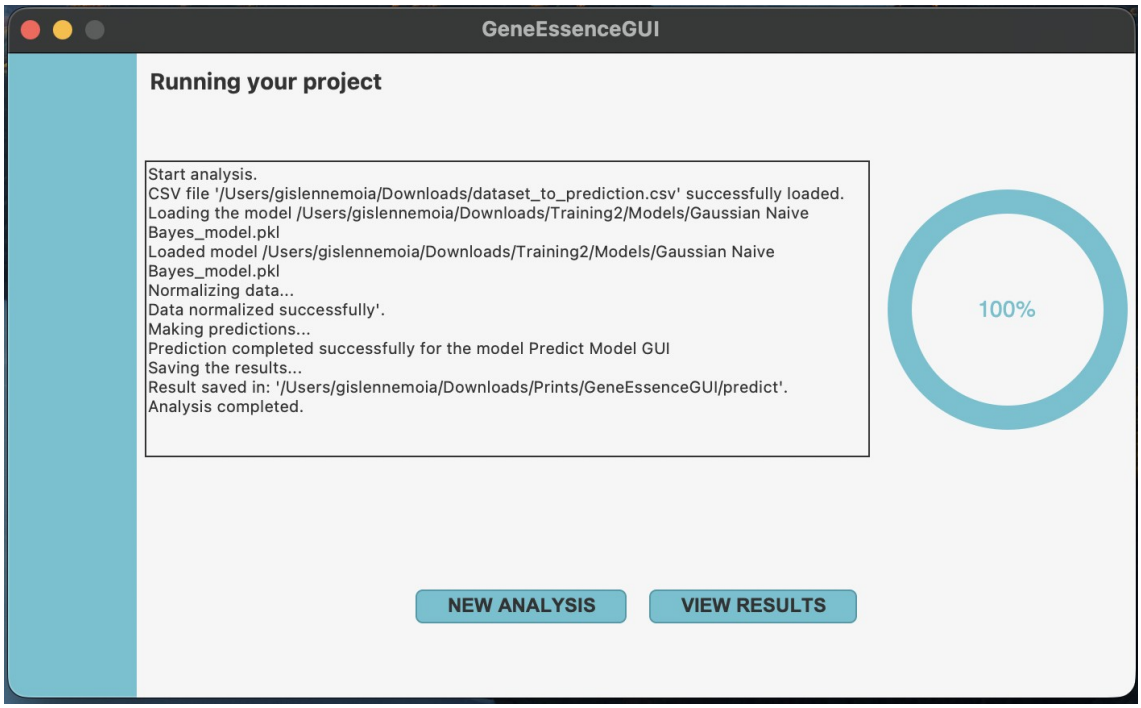

Figure 22. Project Execution Completion Window.
Source: Created by the author.

5.4.4 Prediction Analysis Results

As a result of the prediction analysis, the user will receive a CSV file containing the predicted genes based on the model used.

5.5 How to perform an ensemble analysis?

5.5.1 Ensemble Parameters Definition Window

Upon first access, a window will be displayed to configure a new ensemble (Figure 23). In this window, the user must provide the project identification name and select the dataset file in CSV format for training.

Next, the user can define the proportion of the dataset to be used for testing by specifying the fraction of data reserved for evaluating the model's performance.
Additionally, the user must upload one or more trained models in PKL format, which will be used to compose the ensemble.

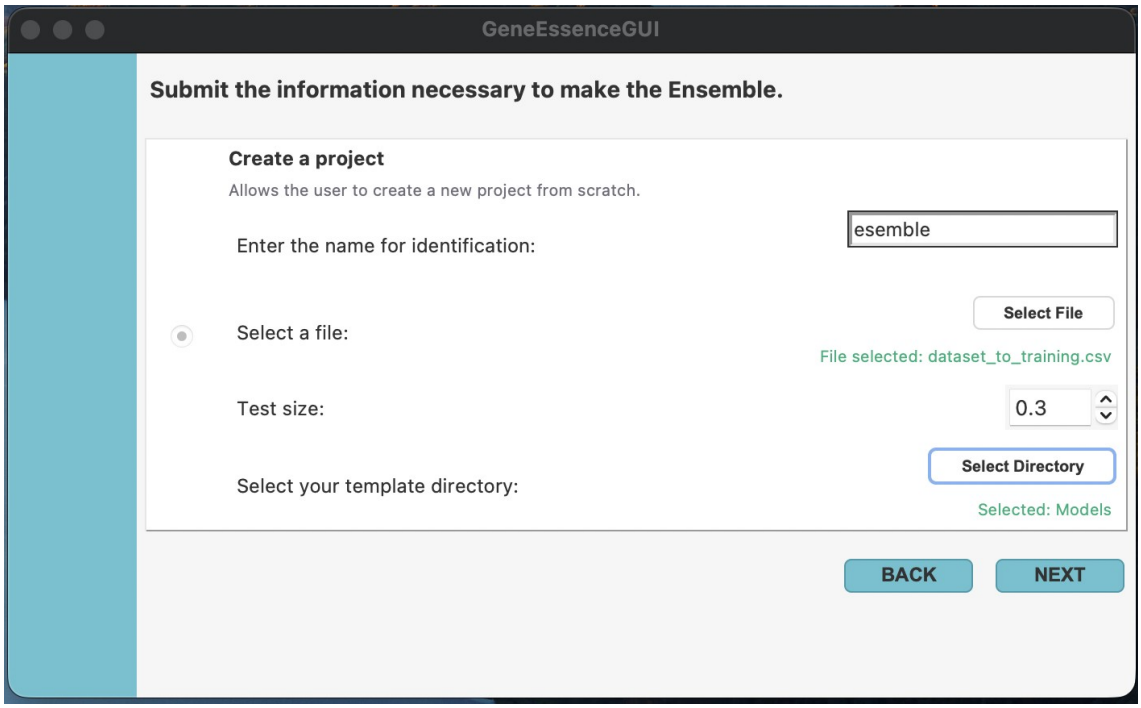

Figure 23. Ensemble Parameters Definition Window.
Source: Created by the author.

Note: If you want to train your model set. Simply enter the name of the project and the dataset in CSV format, then click on the Select Directory button, and you will be told where all the previously trained PKL models are located. The dataset must also be entered so that the ensemble model can be trained. It is important to note that as the user trains their ensemble model, it must be loaded if they want to make predictions with it.

5.5.2 Evaluation metrics selection window in the ensemble

The next screen (Figure 23) will allow the user to choose the evaluation metrics for the models selected in the previous step. The available metrics for this selection are: Accuracy, Kappa, Precision, F1 Score, Matthews Correlation Coefficient, and Recall.

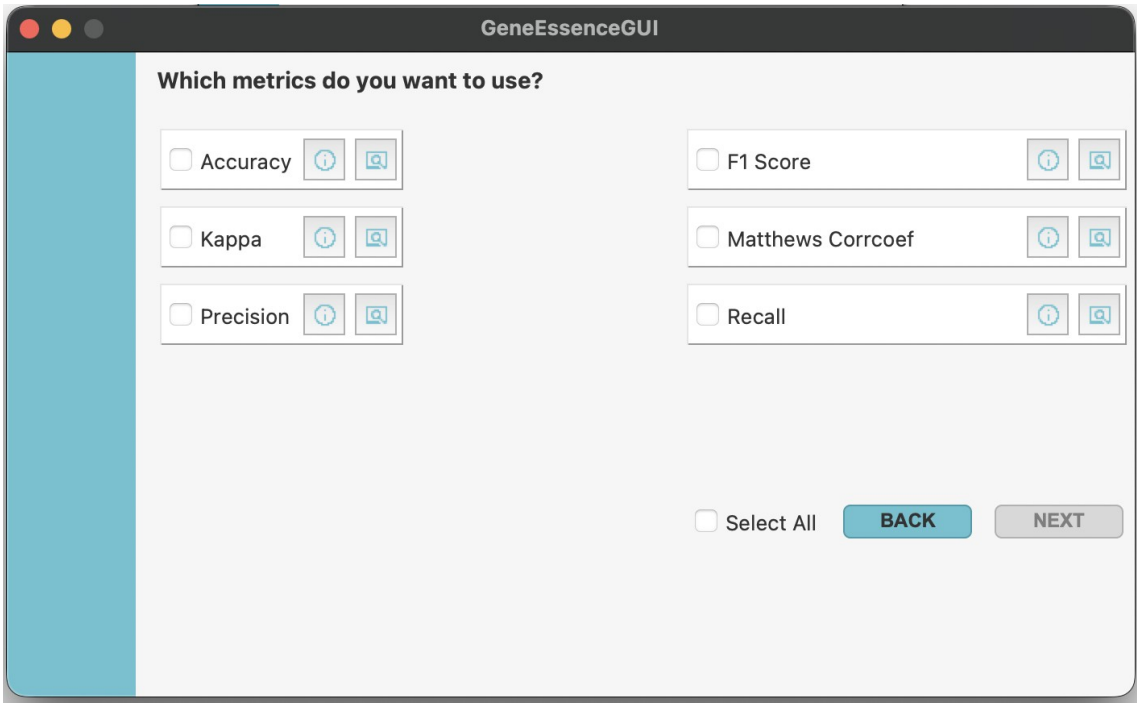

Figure 23. Evaluation Metrics Selection Window for the Selected Models.
Source: Created by the author.

Furthermore, beside each metric, there are two icons: the first shows the definition of the metric to help the user make an informed choice (Figure 24). The second directs to the official documentation, offering comprehensive details on how it functions (Figure 25).

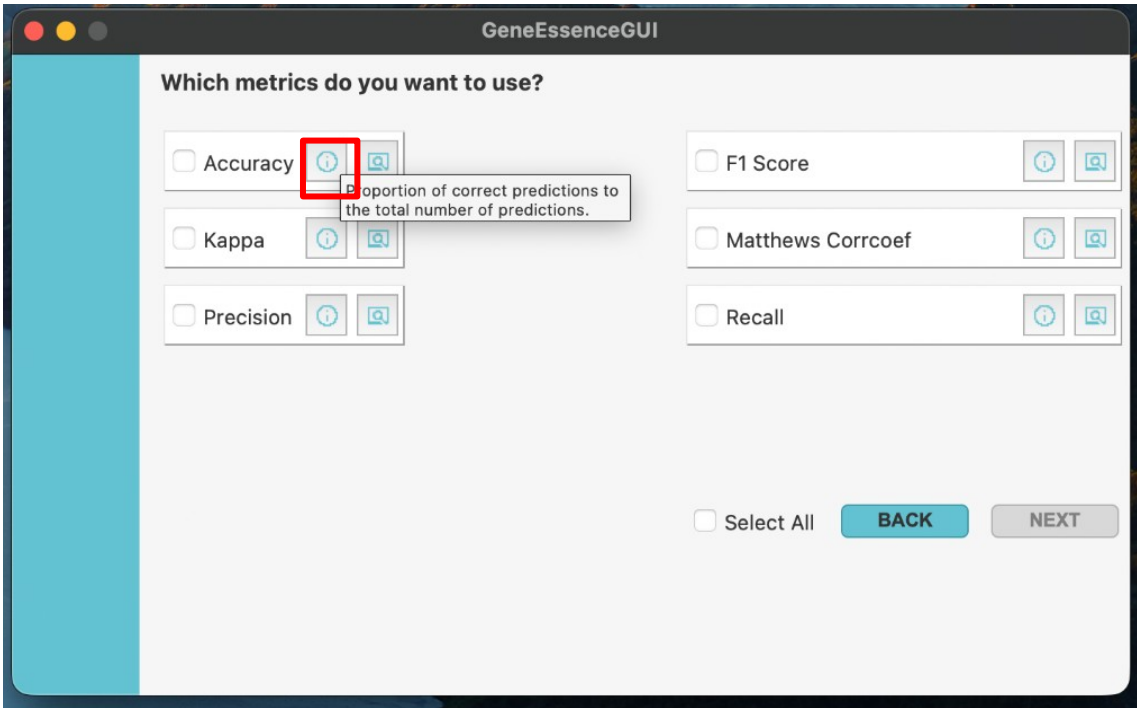

Figure 24. First icon: Information about the Accuracy evaluation metric.

Source: Created by the author.

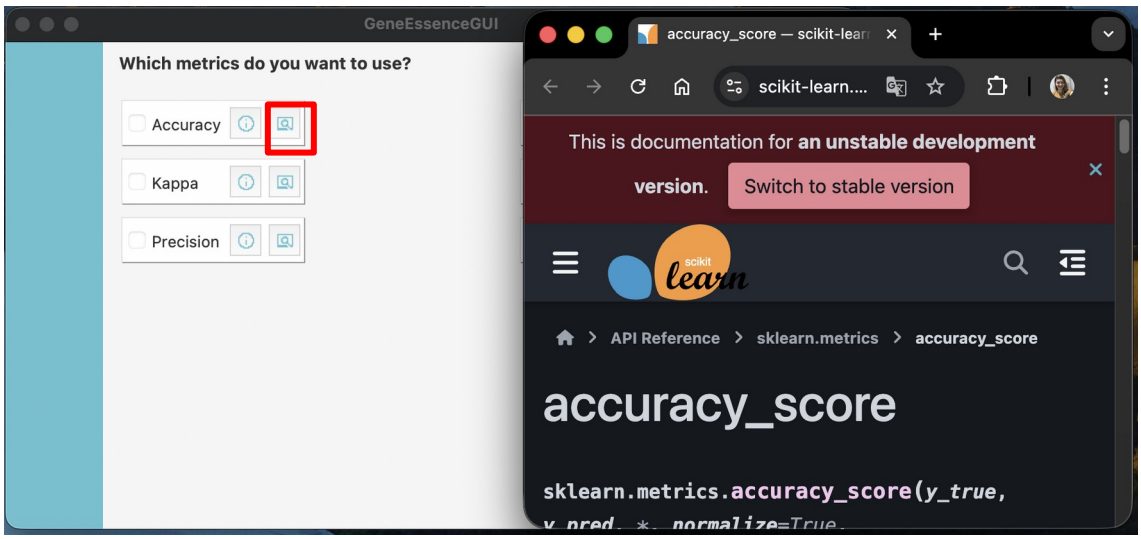

Figure 25. Second icon: Documentation for the Accuracy evaluation metric.
Source: Created by the author.

5.5.3 Results of the Receipt Method Selection Window in the Ensemble

Once the metrics are selected, the user must choose the method for receiving the results (Figure 26) by clicking on the preferred option. The available choices are: email delivery or saving locally on the computer.

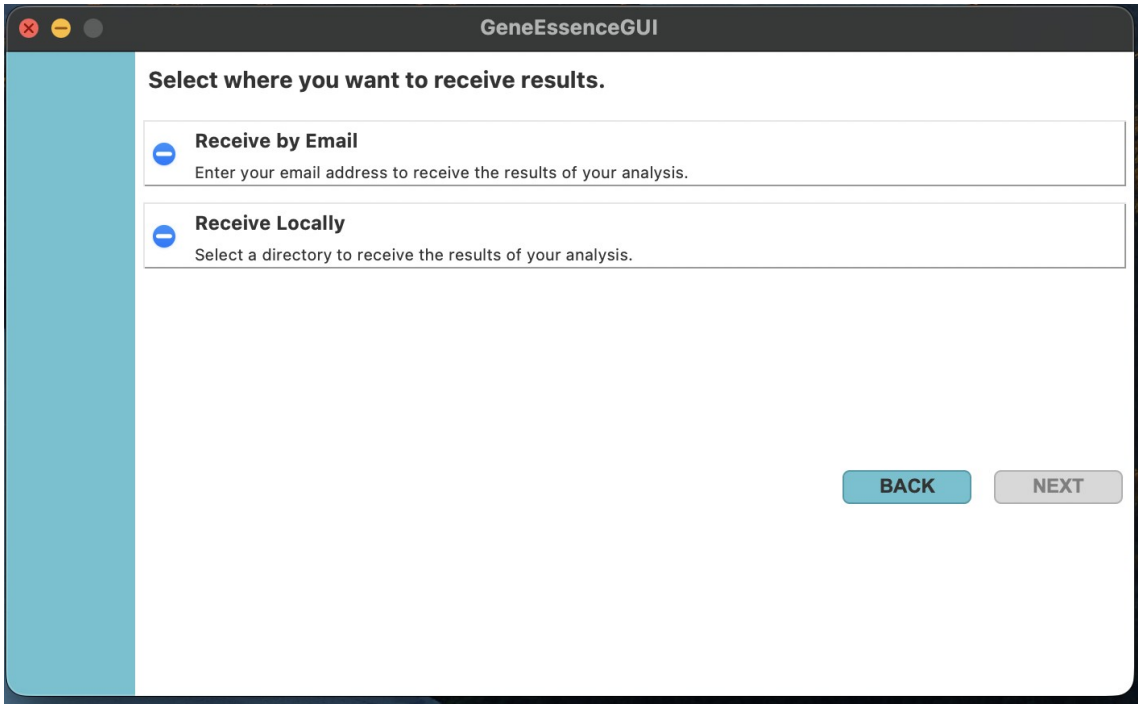

Figure 26. Results Receipt Method Selection Window in the ensemble.
Source: Created by the author.

If the user chooses email delivery, a text field will appear to enter the email address (Figure 27). If the local save option is selected, the user will need to choose the directory

where the results will be stored (Figure 28). After completing the required field, the user can simply click "Next" to confirm and move on to the next step.

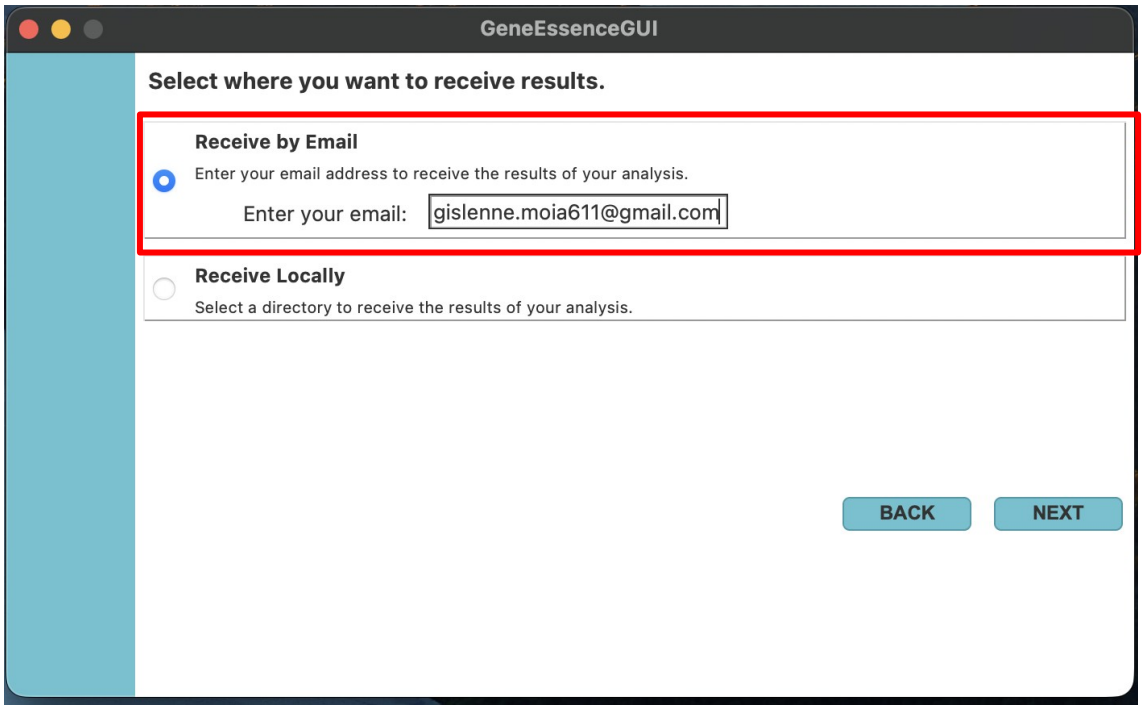

Figure 27. Selection of the email receipt option.
Source: Created by the author.

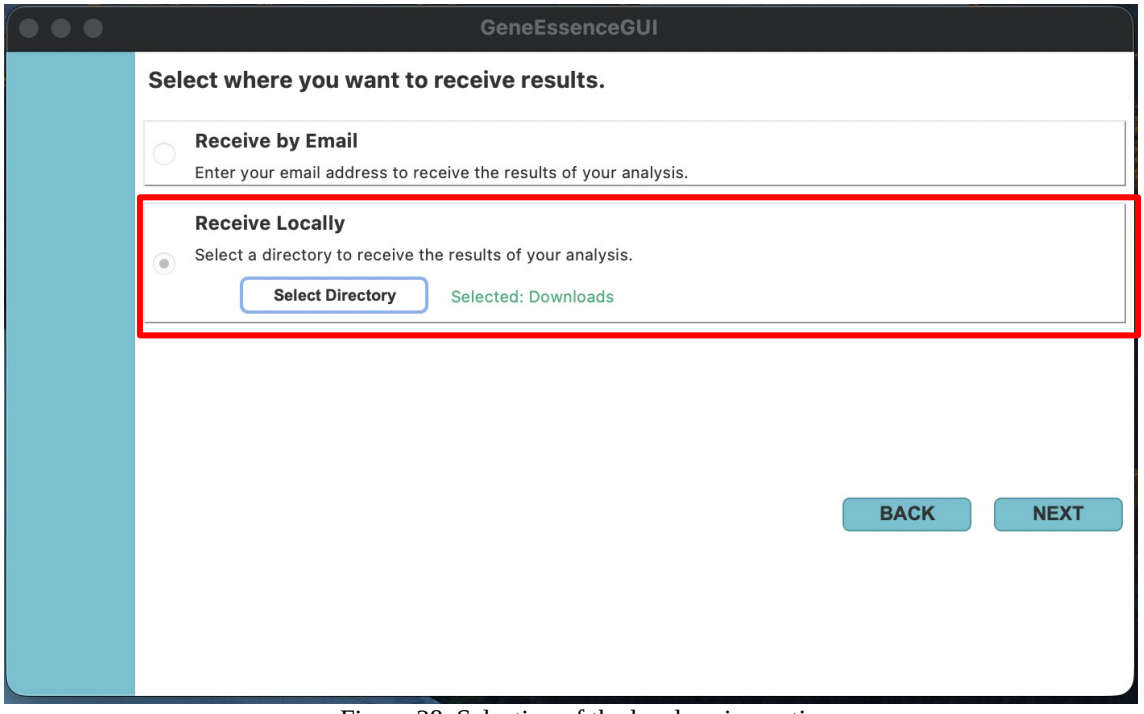

Figure 28. Selection of the local saving option.
Source: Created by the author.

5.5.4 Information Confirmation Window in the Ensemble

In the following step (Figure 29), all the provided information will be shown for review. If everything is accurate, just click the Confirmation button to begin the analysis.

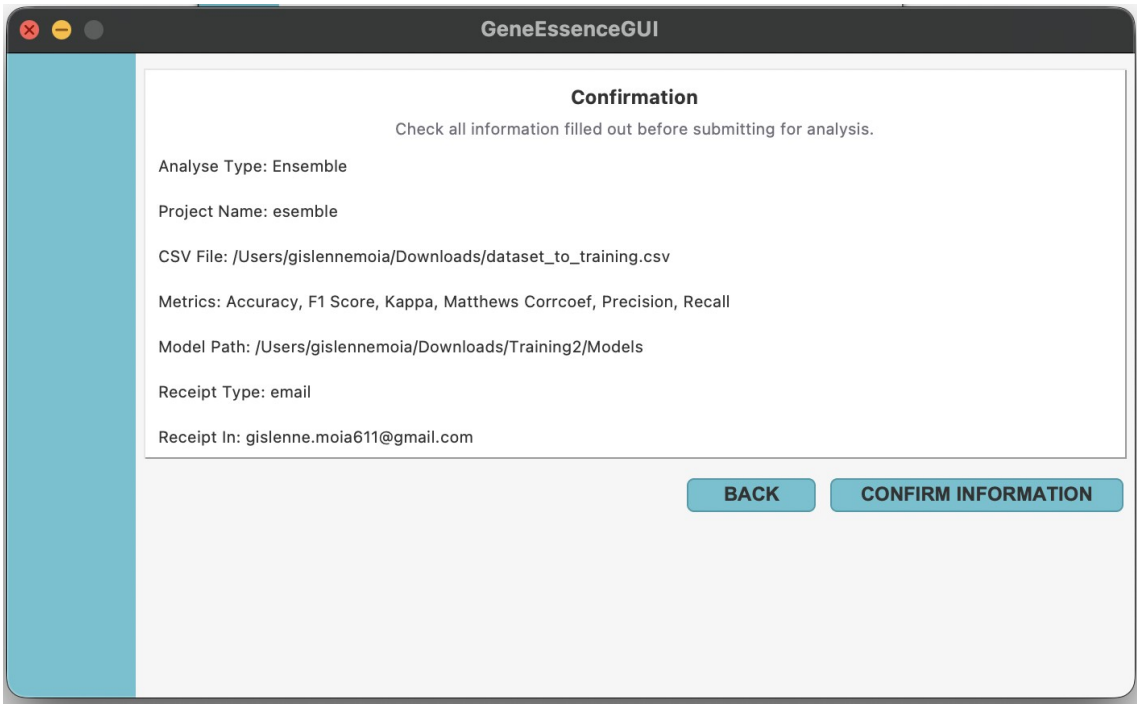

Figure 29. Information confirmation window in the ensemble.
Source: Created by the author.

5.4.5 Project Execution Window in the Ensemble

After confirming the project information, the execution window will appear, displaying details about the steps that have already been completed, along with the percentage of process completion (Figure 30).

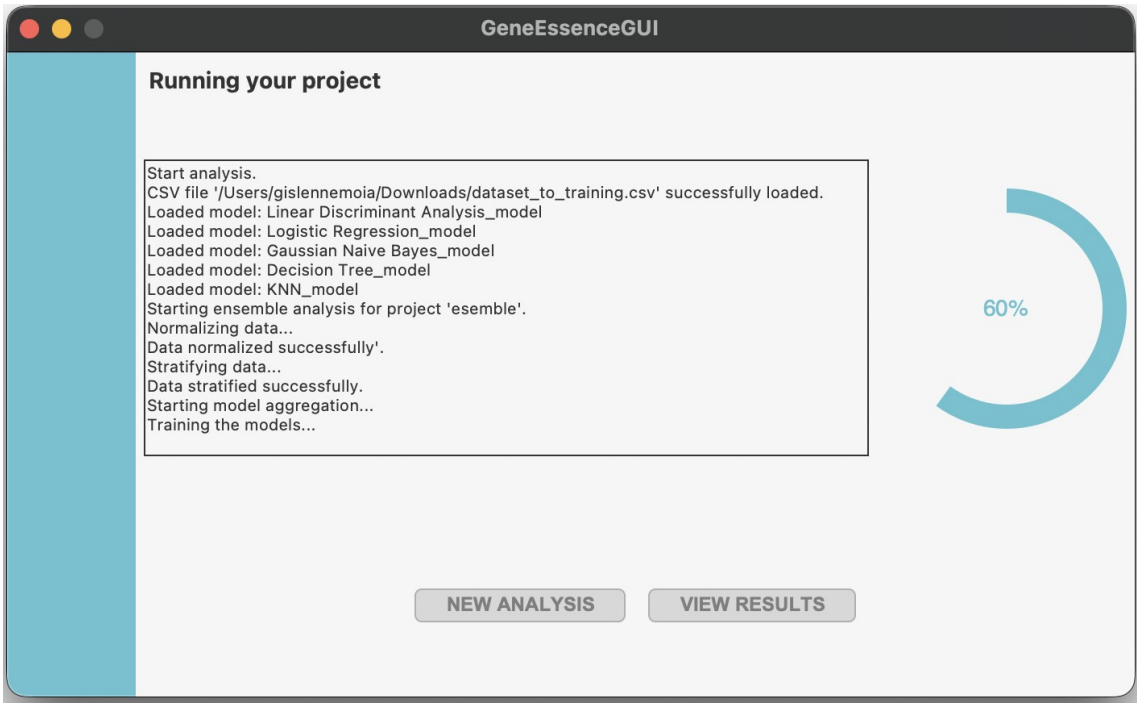

Figure 30. Project Execution Window in the ensemble.

Source: Created by the author.

Once the project execution is finished, the user will be able to view the complete log of the process information. Additionally, the user will have the option to create a new project or view the results (Figure 31).

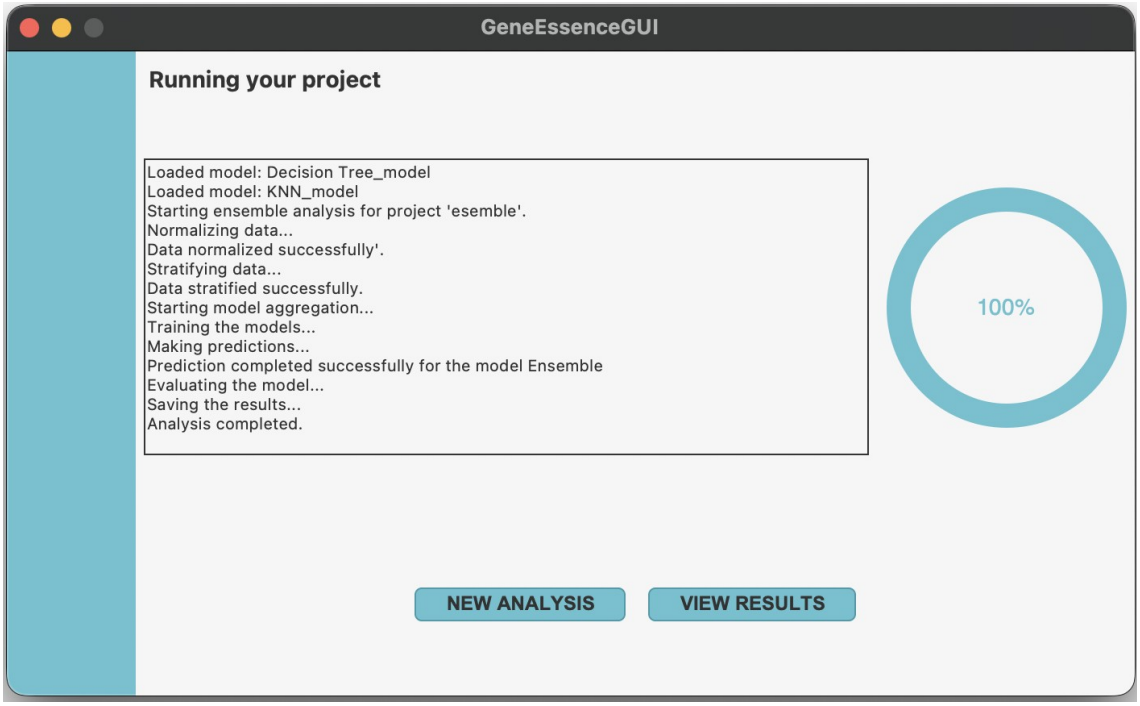

Figure 31. Project Execution Completion Window in the ensemble.
Source: Created by the author.

5.3.6 Ensemble Analysis Results

In ensemble analysis, the user will receive a single model resulting from the grouping of the submitted models, along with a graphical analysis of the model's performance and a CSV file with the predicted genes.

5.4 Load information from existing projects

If the user wants to analyze a project that has already been carried out, the tool offers the option of loading the information from a previous project, as long as it can be found in the database. Figure 32 illustrates the initial screen of the parameter definition stage for one of the analyses, where the data from the selected project is loaded automatically, allowing the user to continue working efficiently and without the need to restart the process.

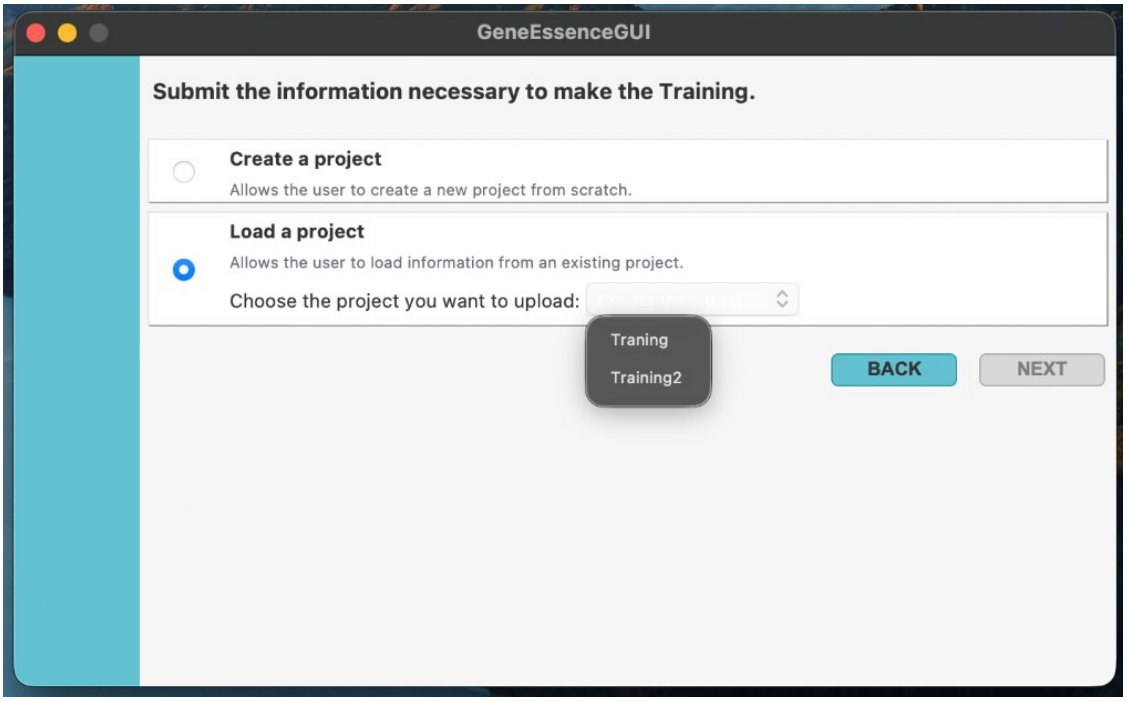

Figure 32. loading a project.
Source: Created by the author.

5.6 How to prepare your dataset

The generation and formatting of the data set, required as input for training and predicting the model, is carried out using the prepareDataset2RNA.jar module, the main parameters of which are listed in the image below.

```
-a,--annotationdeg <arg>   DEG annotation file path
-f,--fastadeg <arg>        Path of the FASTA file of essential genes
-g,--gbncbi <arg>          Path to NCBI GenBank files
-h,--help                  Display this help message
-p,--data4predict <arg>    Path to GenBank files for prediction dataset
```

1. To prepare the dataset for the essential gene prediction step, it is necessary to use the module with the “-p” parameter, see the execution example below:

```
java -jar prepareDataset2RNA.jar -p /home/allan/data

Generating dataset for prediction...
Dataset for prediction generated successfully.
```

2. To prepare the dataset that will be used to generate the models, it is necessary to provide the annotation files in CSV format and the amino acid sequences in

FASTA format, both obtained from DEG, as well as the path to the organism files in GenBank format. See the example run below:

BIOD
Bioinformatics, Omics
and Development