VirnyFlow User Study

In this form, you will be guided through **two tasks** designed to help you explore different features of VirnyFlow. To successfully complete the study, please make sure that you:

- Have a stable internet connection.
- Carefully follow all steps and instructions provided.
- Answer all required questions before proceeding.

In all tasks, you will work with the Diabetes Dataset 2019. Each task appears in a separate section of the form. You may refer to the Tutorial Notes at any time to review information covered in the tutorial, including key concepts, system workflow, and relevant definitions.

* Indicates required question

1. Email *

Task 1: Anon

In this task, you are going to use the **Anon** library—an evaluation module within VirnyFlow—to conduct an in-depth performance profiling for three models trained on the Diabetes Dataset 2019. Using Anon's metric computation interface, you will compute both accuracy and fairness metrics for these models.

Throughout this Google Colab notebook, you will be guided through the whole process from importing necessary libraries to comparing different ML models.

Important: While answering is mandatory, please submit all your text answers **only in this Google Form**. You do not need to fill in the answer sections inside the notebook itself. In this section, each question is under the TODO number from the Google Colab notebook.

TODO 1

After you selected a sensitive attribute and its disadvantaged value in the first step

2. ***Which sensitive attribute and disadvantaged value did you specify?*** *

3. ***Explain why the selected sensitive attribute is appropriate for this prediction task and its context.*** *

TODO 2

After you completed second and third steps

4. ***Based on the accuracy metrics, which model would you choose for this classification task?*** *

Mark only one oval.

- LogisticRegression
- KNeighborsClassifier
- XGBClassifier

5. *[Look at your sensitive attribute and its disadvantaged value from TODO 1]* ***Can you say anything about who may be disadvantaged and in what way when the model makes an error?*** *

TODOs 4 - 5

After you completed 4-th step

6. Consider your chosen model's error rates (FPR, FNR) for the protected classes. * ***Does your chosen model further disadvantage the already disadvantaged group?***

   Hint: *If the FPR or FNR values for the disadvantaged group are far from 0.0, this indicates that the model* is further disadvantaging this group.

7. *[After you selected any additional model and compared the metrics (TODO5)]* * ***Would you change your original choice? Why or why not?***

TODO 6

After you completed 5-th step

8. ***Does this model demonstrate satisfactory fairness? Explain why or why not.*** *

Task 2: VirnyFlow

In **Task 1**, you learned how VirnyFlow—via the Anon library—measures multiple performance dimensions in a flexible and context-sensitive way, enabling evaluation beyond a single accuracy metric. **Task 2** builds on this foundation by focusing on the results of **multi-objective optimization** and how VirnyFlow presents these results through its **visualization interface**.

In this task, you will explore **pre-computed experimental results** generated by VirnyFlow on the same *diabetes* dataset. Using an experiment configuration with four ML models and multiple optimization objectives, VirnyFlow automatically trained and evaluated **800 ML pipelines**, executed efficiently in parallel on **eight workers**. Each pipeline corresponds to a different combination of ML model choice and hyperparameters suggested by multi-objective Bayesian optimization (as described in the tutorial).

Instead of running these experiments yourself, you will use the visualization interface to examine the resulting pipelines, compare trade-offs, and understand how the evaluation logic from Task 1 translates into optimization outcomes and visual insights in VirnyFlow.

Step 1

Open the following HuggingFace space in your browser.

Step 2

Look at the first tab ("Execution Progress") and answer the following questions.

9. *[Refer to the Experiment Configuration YAML file]* ***What optimization objectives were used to tune the ML pipelines? In your answer, describe the metrics used, specify the groups for which each metric was measured, and indicate the weight assigned to each metric in the optimization process (from 0.0 to 1.0).*** *

10. *[Look at the Logical Pipeline Statistics table]* ***What is the best performing ML pipeline based on a cost model score?*** *

Mark only one oval.

- ◯ Pipeline1
- ◯ Pipeline2
- ◯ Pipeline3
- ◯ Pipeline4

11. *[Look at the Pipeline Execution Cost plot]* ***Is the best performing ML pipeline also the one that has the highest pipeline execution cost?*** *

Mark only one oval.

- ◯ Yes
- ◯ No

12. *[Look at the Average Objectives plot]* ***Compare Pipeline1 and Pipeline2. Which pipeline has better F1, and which pipeline has better FNR Difference?*** *

_____

_____

_____

_____

_____

Step 3

1. Look at the second tab ("Pipeline Performance").
2. Select a *Pipeline4* pipeline. Leave all other selectors as default.
3. Choose *FNR_Difference* and *FPR_Difference* as error disparity metrics.
4. Click the "Submit" button (*it may take up to 12 seconds to load pipeline metrics for the plot*).

13. *[Look at the metrics bar chart]* ***How large is the difference between FNR_Difference and FPR_Difference? Provide a single absolute numerical value.*** *

14. *[Look at the Dataset Statistics bar chart]* ***What is the proportion between disadvantaged and privileged groups? Is the dataset balanced?*** *

15. *[Look at the Dataset Statistics bar chart]* ***What would be the accuracy of a majority-class diabetes predictor (i.e., a model that always predicts the most frequent class in the dataset)?*** *

Step 4

1. Look at the third tab ("Pipeline Optimization").
2. Select a *Pipeline4* pipeline (*it may take up to 30 seconds to load all the plots on the page*).

16. *[Refer to the Pareto Frontier plot] The plot shows reversed **F1** values on the x-axis and absolute **FNR Difference** values on the y-axis, where lower values on both axes indicate better performance. Each point represents a pipeline configuration, and the color indicates the optimizer trial / iteration in which it was evaluated.* *

Which point (i.e., from which iteration number) is closest to the origin (0.0, 0.0)? You may hover over the points to see their iteration numbers.

Mark only one oval.

- ◯ 66
- ◯ 74
- ◯ 171
- ◯ 200

17. *[Refer to the Pareto Frontier plot]* ***Compare the point closest to the origin (0.0, 0.0) with the other points on the plot. What does this comparison tell you about the optimization process? In particular, discuss whether the optimizer was able to improve both objectives simultaneously, or whether there is evidence of a trade-off between them.*** *

_____

_____

_____

_____

_____

18. *[Look at the Parameter Importance plot]* ***Which hyperparameter has the highest importance for both F1 and FNR Difference?*** *

Mark only one oval.

- ◯ bootstrap
- ◯ max depth
- ◯ min samples leaf

Step 5

1. Look at the fourth tab ("Pipeline Comparison").
2. Select "*diabetes_w_acc_0_5_w_fair_0_5*" and "*diabetes_w_acc_0_25_w_fair_0_75*" configurations at the top (*it may take up to 15 seconds to reload the page*). Note that the "*diabetes_w_acc_0_5_w_fair_0_5*" will be assigned an "E1" alias, and the "*diabetes_w_acc_0_25_w_fair_0_75*" will be assigned an "E2" alias.
3. On the Metrics Heatmap, select the following ML pipelines:
   "E1: Pipeline1",
   "E2: Pipeline1"
4. Click the "Submit" button (it may take up to 15 seconds to reload the plot).

19. ***[Refer to the Metrics Heatmap]*** *According to the System Configuration table on this page, both* ***E1: Pipeline1*** *and* ***E2: Pipeline1*** *were optimized for the same objectives:* ***F1*** *and* ***FNR Difference****. The only difference between them is the optimization weights:* ***E1: Pipeline1*** *uses equal weights (0.5 for F1 and 0.5 for FNR Difference), while* ***E2: Pipeline1*** *assigns a higher weight to fairness (0.25 for F1 and 0.75 for FNR Difference).* *

    Based on the heatmap, what trade-off do you observe between accuracy and fairness? Specifically, how do the F1 and FNRD values change when moving from E1: Pipeline1 to E2: Pipeline1 (i.e., when the fairness weight increases from 0.5 to 0.75)? Please include absolute numerical differences in your response. Remember that values of FNRD and FPRD closer to zero indicate better performance.