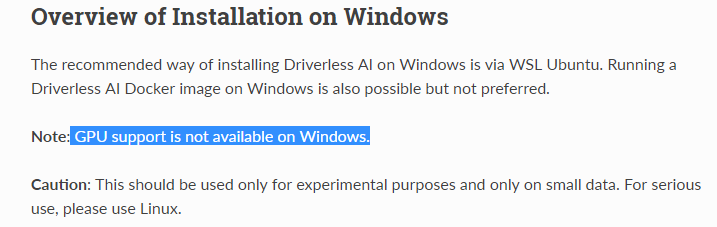
Installation:

<http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/install/windows.html>

Ubuntu WSL installation:

<https://docs.microsoft.com/en-us/windows/wsl/faq#what-is-windows-subsystem-for-linux-wsl>



Account:

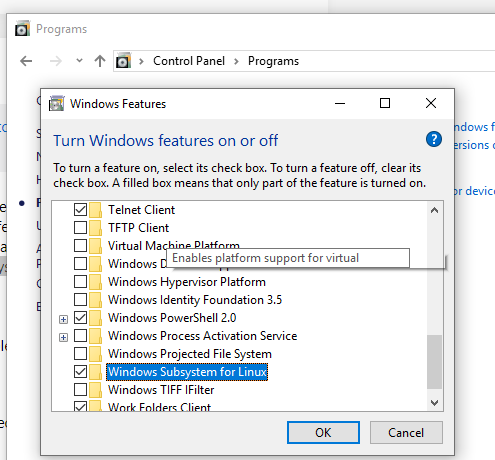
Ubuntu login: ziweifan177/Fzw1985!!!!

Installation:

Step1: download Deb version of driverless:

Step2: download WSL Linux 18

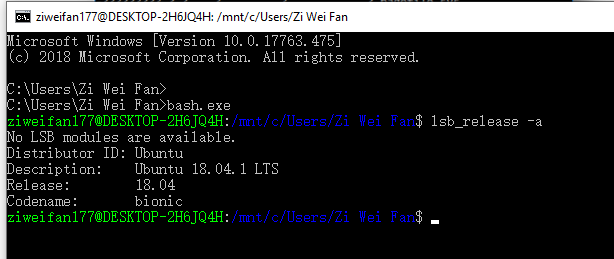
Step3: Enable Windows Subsystem for Linux and restart laptop



Step4: Click Ubuntu icon for auto installing:

Step5: windows cmd and bash.exe for running Ubuntu:

(lsb\_release –a: check the info of current Ubuntu version.)

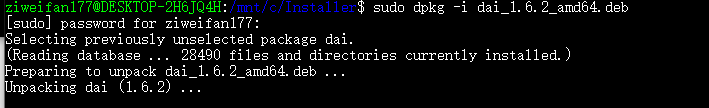


Step6: Direct to different path by:

(/mnt/c is the mount points)

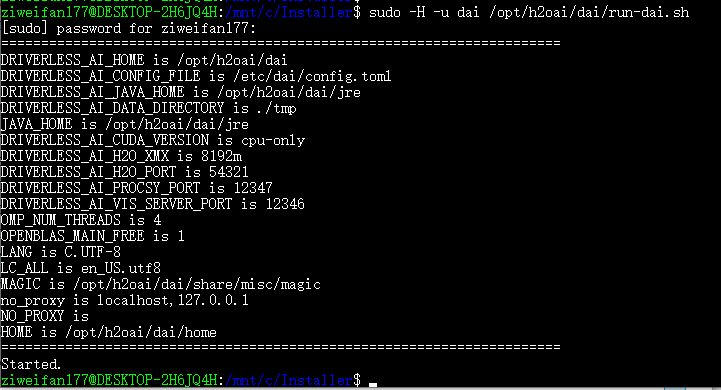


Step7: Install Driverless AI in the corresponding path.



Step8: Run H2O

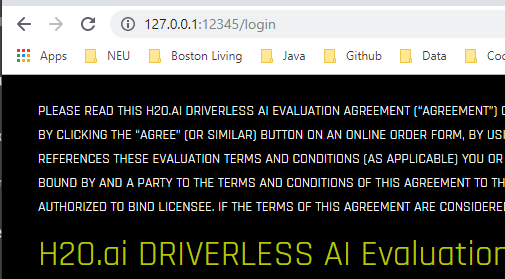
sudo –H –u dai /opt/h2oai/dai/run-dai.sh



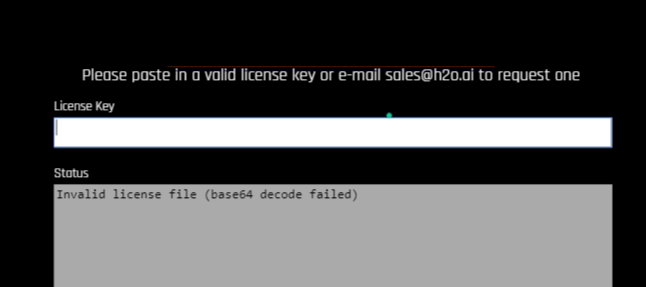
Step9: Run by localhost:

Type 127.0.0.1:12345 (default)

Account h2oai/h2oai



Step10: Input the license code:

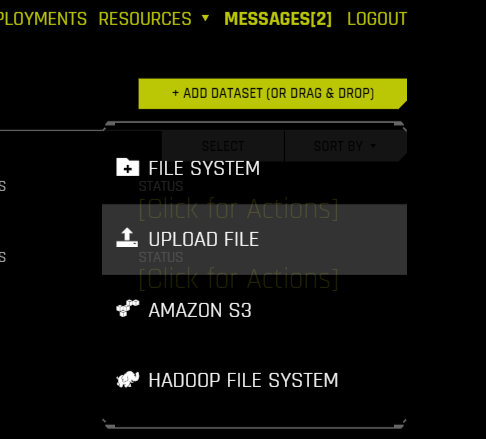


Run the dataset

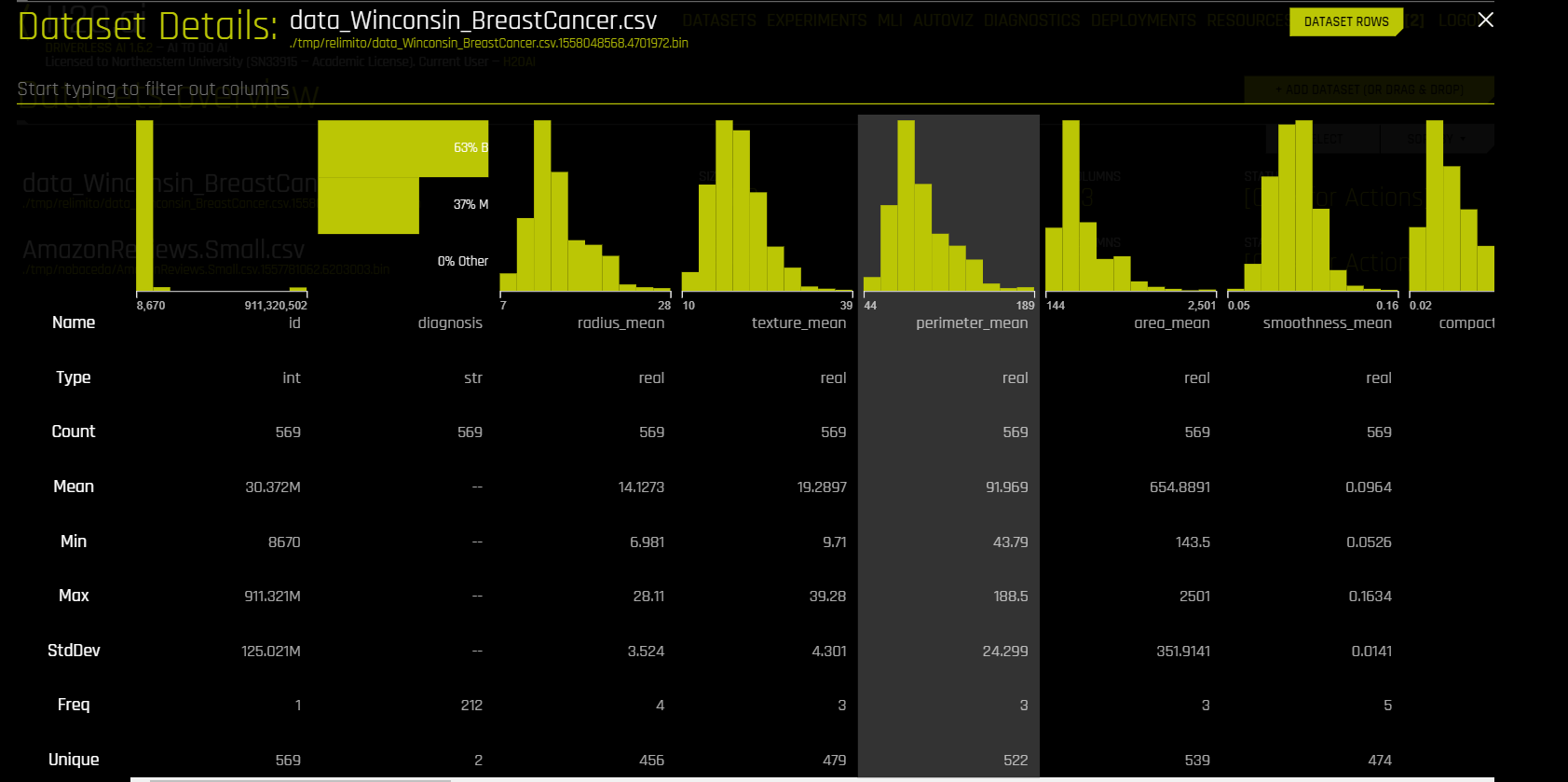
Refer: <https://h2oai.github.io/tutorials/automatic-ml-intro-tutorial/#0>

Dataset: Breast Cancer: <https://www.kaggle.com/buddhiniw/breast-cancer-prediction>

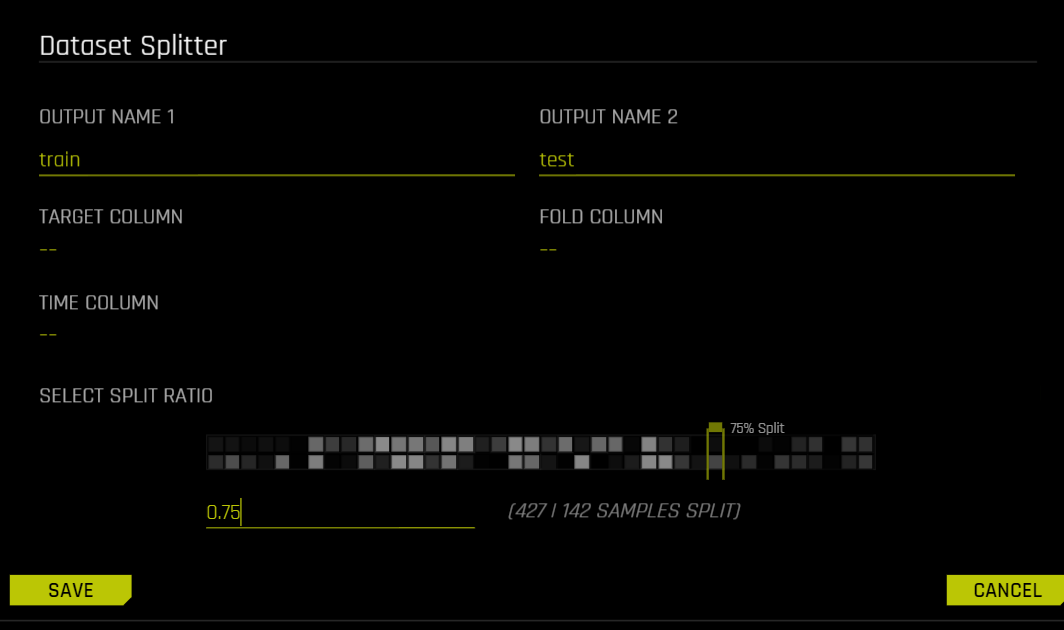
Step1: Upload file:



Step 2(Optional): Check details:



Step 3: Split data:



Step4: Predict on Train data:

**Parameters to notes:**

1. **Time series:**

If the data has a strong dependency on time (such as seasonality or trend), and you would like to treat this problem as a time series problem (with causality in mind during modeling), you can select a time column which provides the time stamp for each row. If it's not a time-series problem, select [OFF]

1. **Dropped Columns: (We dropped ‘id’ for this case.)**

Drop column(s) from your dataset that you don't want to use in the experiment.

1. **Validation Dataset:**

Select the dataset you want to validate. This set will be used to validate parameters like models, features etc.

For advanced users: Optional external holdout dataset for validation (tuning) of the modeling pipeline. This is useful for data that contains rows that are not independent or not identically distributed. In general, it is recommended to provide all labeled data as training data and let Driverless AI do internal validation holdout splits.

1. **Test Dataset:**

The dataset that will be used to test the model generated from the training dataset. It is not used during training of the model and results are available at the end of the experiment.

1. **Target Column:**

What do you want to predict?

1. **Time Column:**

Provides a time order(time stamps for observations)

1. **\*Fold Column:**

For advanced users: Optional column to use to create stratification folds during (cross-) validation.

For example, select a zip code column if you want all observations from the same zip code to be either used for training or for validation. This can prevent data leakage and improve generalization.

1. **Accuracy setting:**

Select higher accuracy to **get more accurate models** and better generalization estimates - click the accuracy knob and drag it up or down.

1. **Time Setting:**

Select larger time values if **you are willing to wait longer** (you can always Finish early)

1. **Interpretability:**

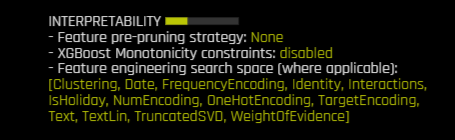
Select lower interpretability values for more complex models and more feature engineering.

1. **Scorer:**

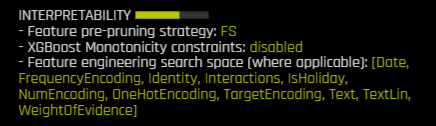
For advanced users: you can select a scorer (metric) to be optimized, otherwise the default values will be fine.

\*Note about interpretability:

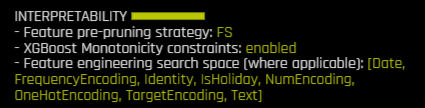
When interpretability = 3:



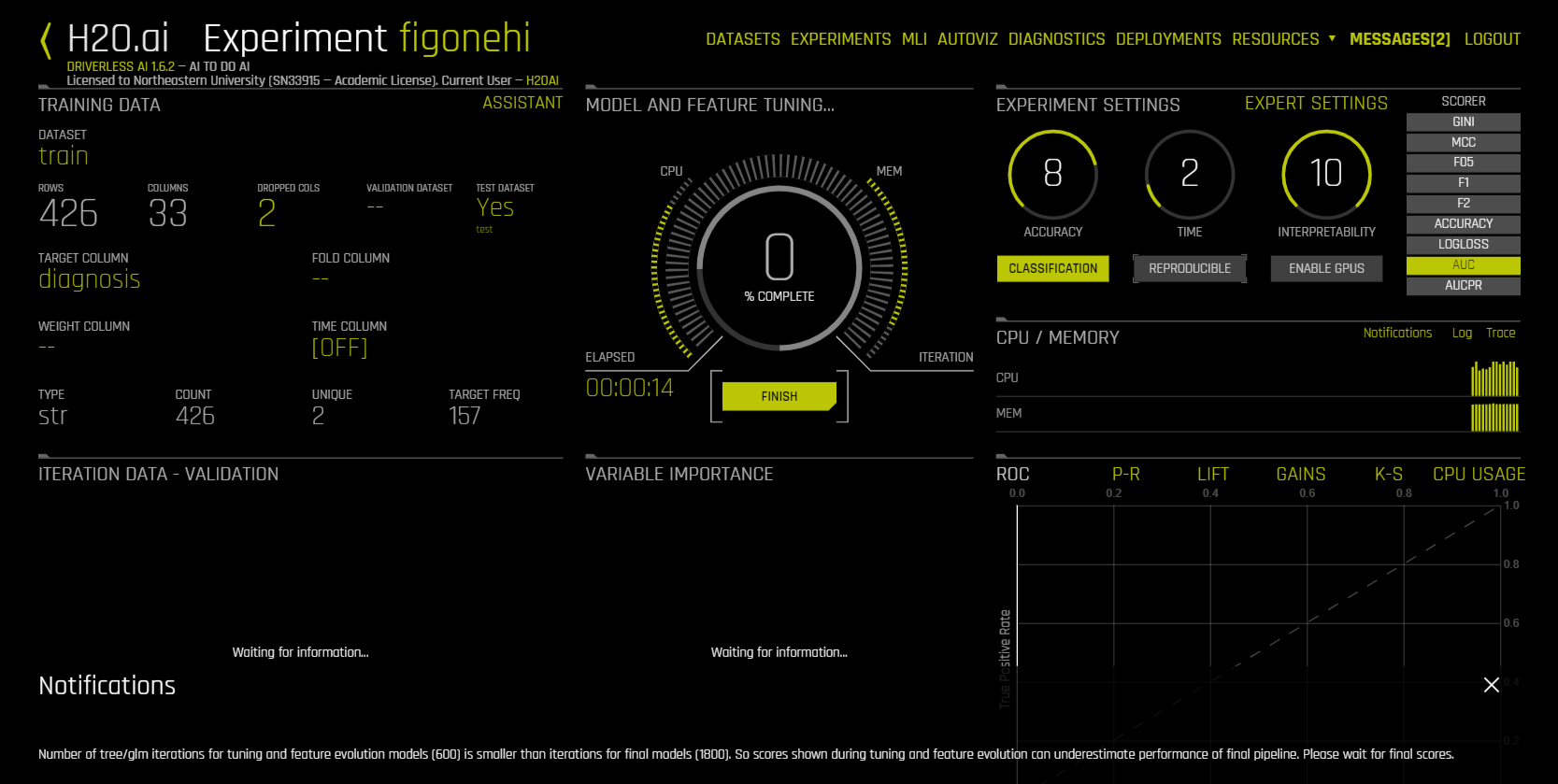
When interpretability = 3-9:



When intrepretability =10:



Step5: In the progress:



*Things to note:*

1. **Experiment Name** - Randomly generated experiment name. Name can be changed at any time
2. **Training Data** - Dataset details
3. **Running Status Display** - Status of parameter tuning followed by feature engineering and scoring pipeline. Experiments can be Stopped by clicking the Finish button.
4. **Experiment Settings -** Overview of experiment settings (unable to adjust the while experiment is running)
5. CPU/Memory information including Notifications, Logs and Trace info
6. **Iteration Data and Variable Importance** - Iteration Data is the internal validation for each cross-validation fold with the specified scorer value. You can hover over any of the iteration points in the Iteration Data graph, and the see the updated variable importance for that iteration on the **Variable Importance**
7. **Classification Problem Graphs -** Toggle between a ROC curve, Precision-Recall graph, Lift chart, Gains chart, and GPU Usage information (if GPUs are available)

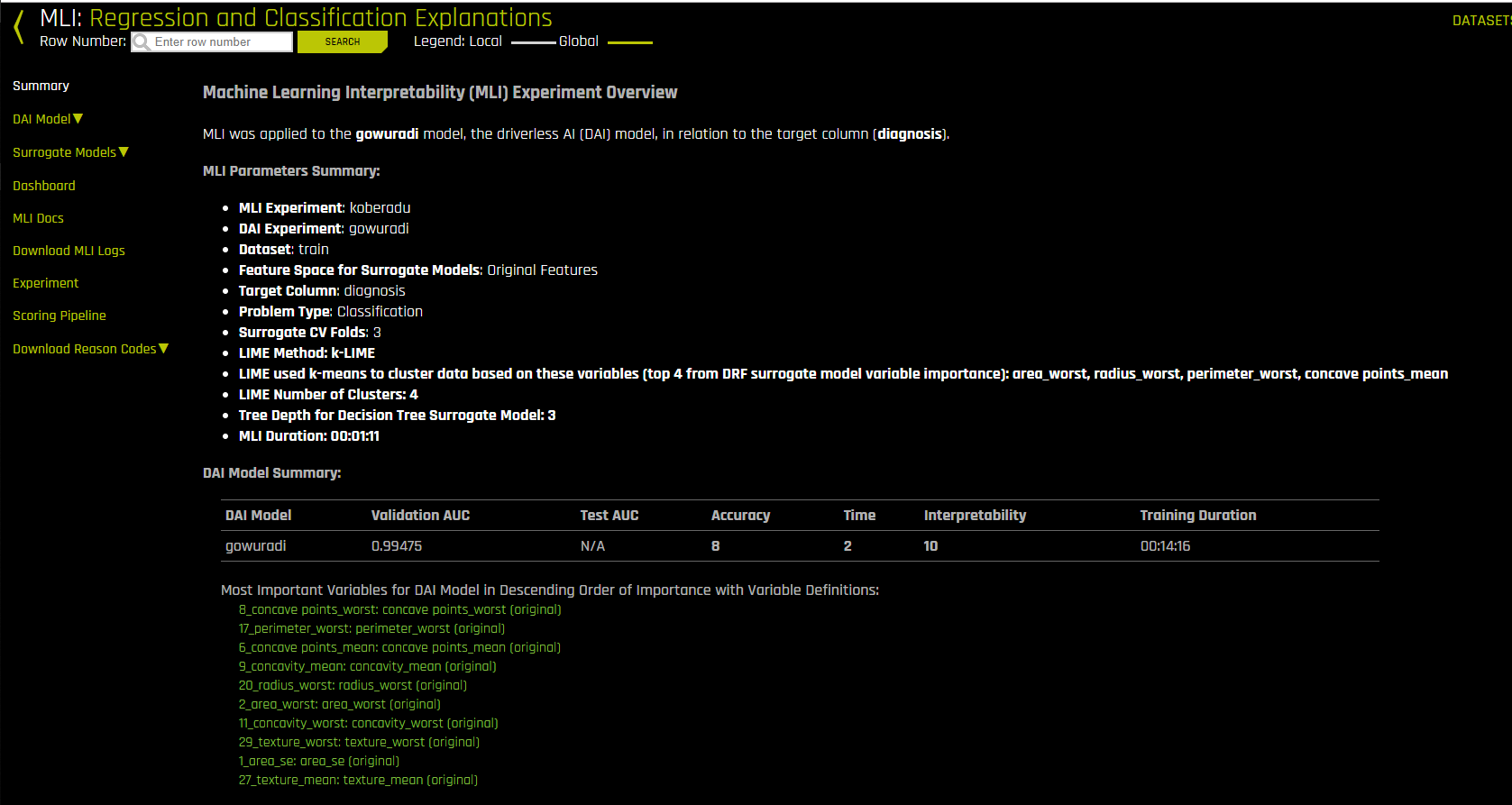
Step6: Interpret the model:

Ref1 : <https://h2oai.github.io/tutorials/machine-learning-interpretability-tutorial/#0>

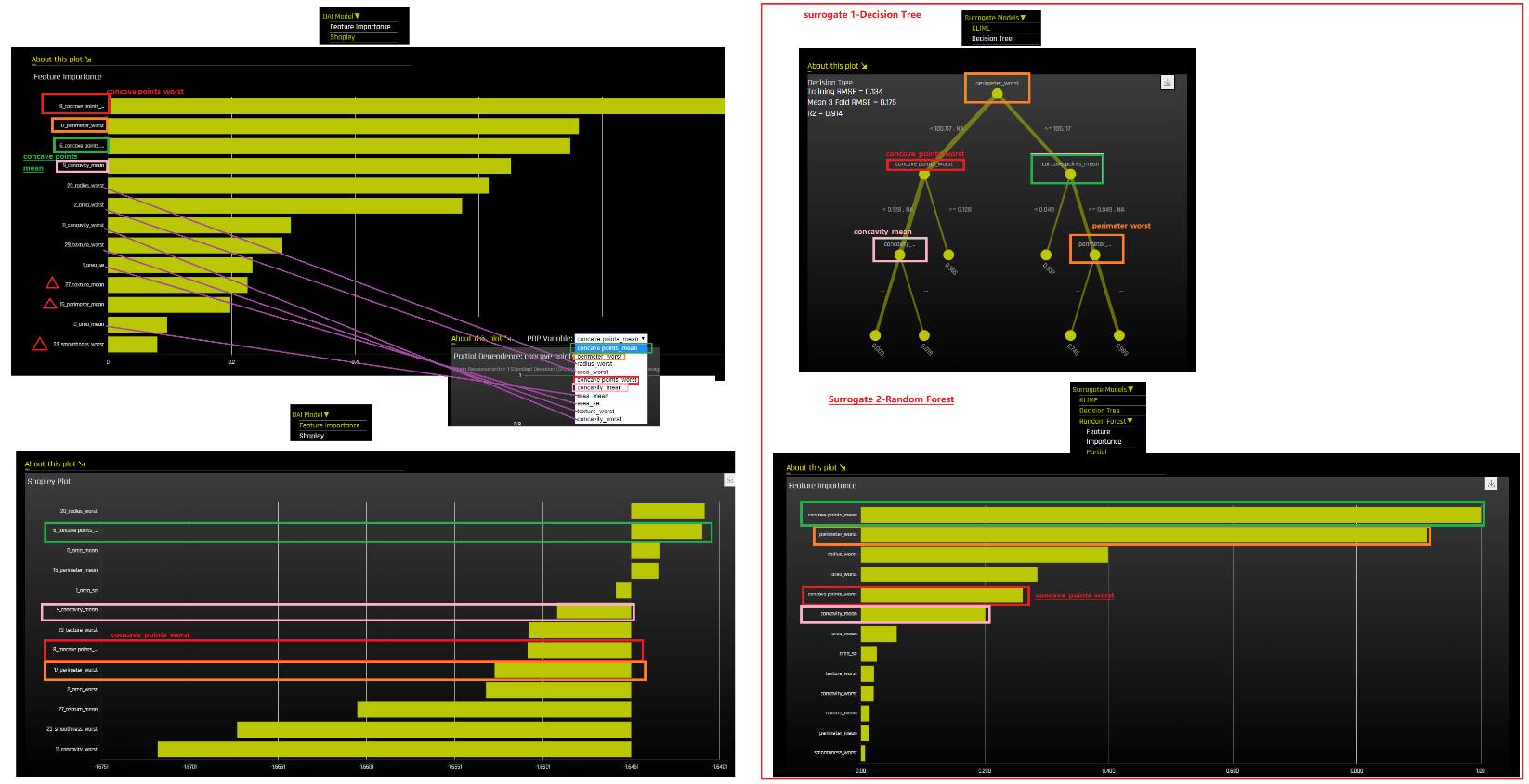
Ref2: <https://h2oai.github.io/tutorials/automatic-ml-intro-tutorial/#9>

After the predictive model is finished, we can explore the interpretability of our model. In other words what are the results and how did those results come to be?

* Which attributes from our Training Set do you think are the most important in relation to target? Make a note of your top 2 attributes to compare it with the model's results.



Global interpretations help us understand the entire relationship modeled by the trained response function, but global interpretations can be approximate or based on averages.



Q:

1. DAI Model-Feature importance & Surrogate 2-RF Feature importance:

? Will the shifting between the important features of these 2 FI be quite different?

? Will we trust the new FI ‘radius\_worst’ & ‘area\_worst’ in Surrogate 2 FI?



