Created in the cloud with Aspose.Words for Cloud. http://www.aspose.com/products/words/cloud

# Azure ML Part 10, Create Web service from Models

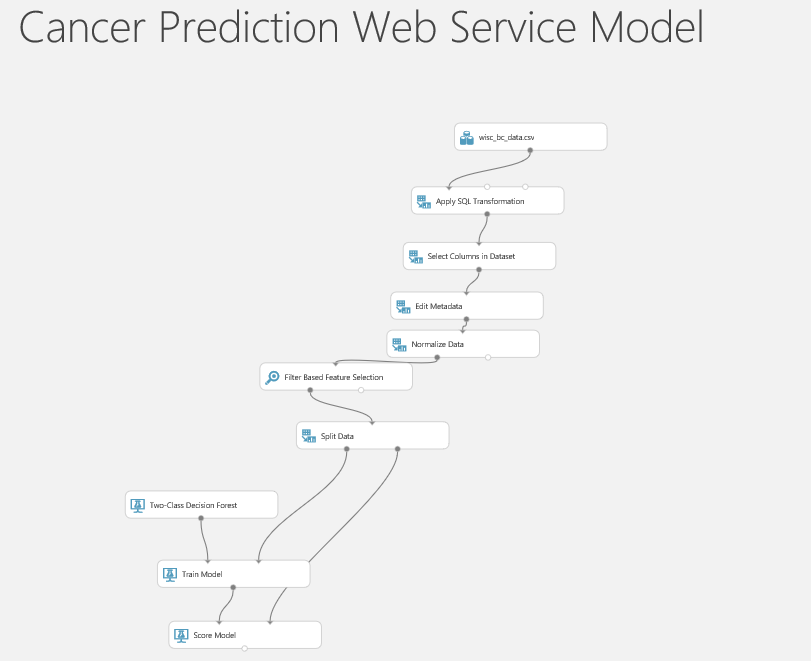
Published Date : June 25, 2017

In previous posts, I have explained how to create a machine learning scenario using Azure ML components.

One the main advantages of using Azure ML is the ability to create a wed service from it. That means the created model can be used in other services via  a API URL and Password.

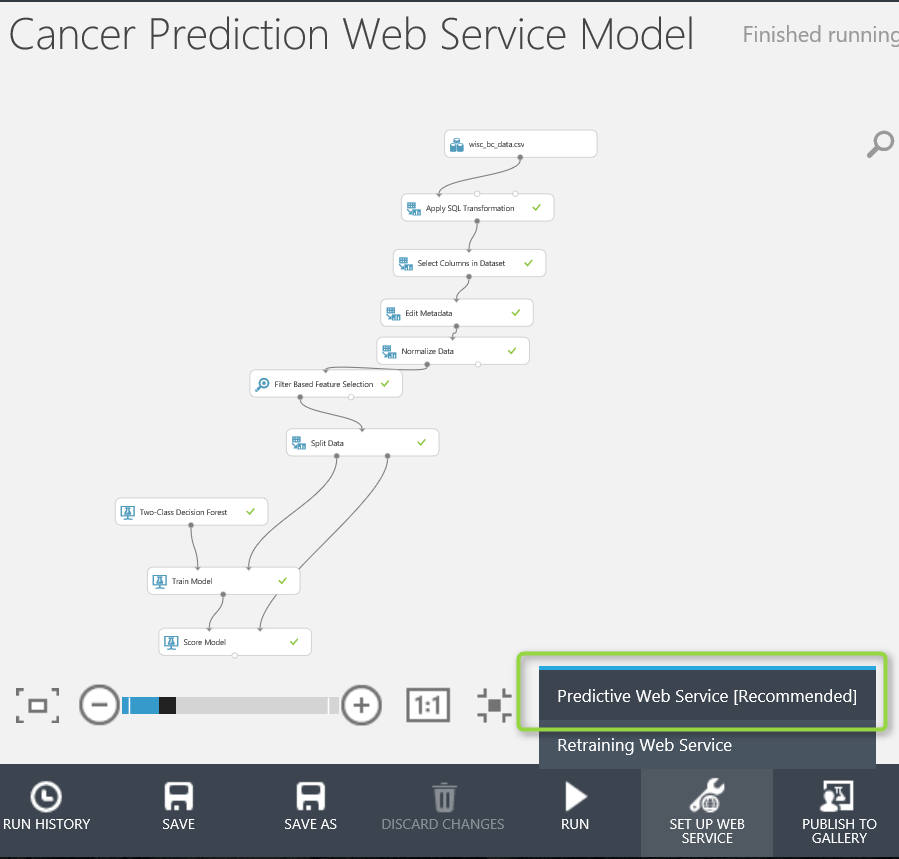
In this post I will show how to create an API for a predictive model.

Back to my previous scenario on prediction of the patient cancer situation (see below picture).

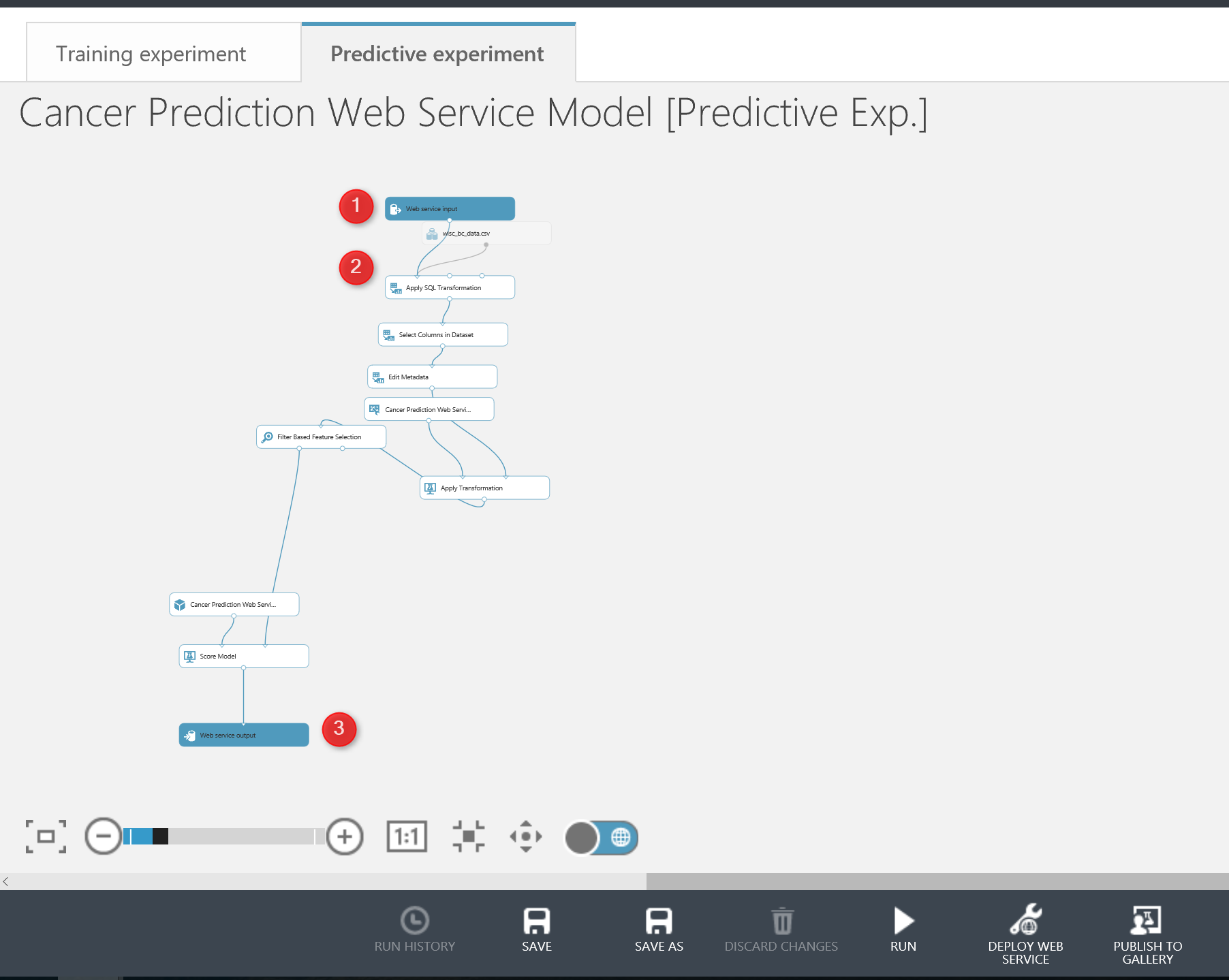
[](http://radacad.com/wp-content/uploads/2017/06/112.png)

I am going to create a web service out of it.

to create a web service, the best and easiest way is to run the experiment first then click on the “**Set Up Web Service**” icon on the bottom

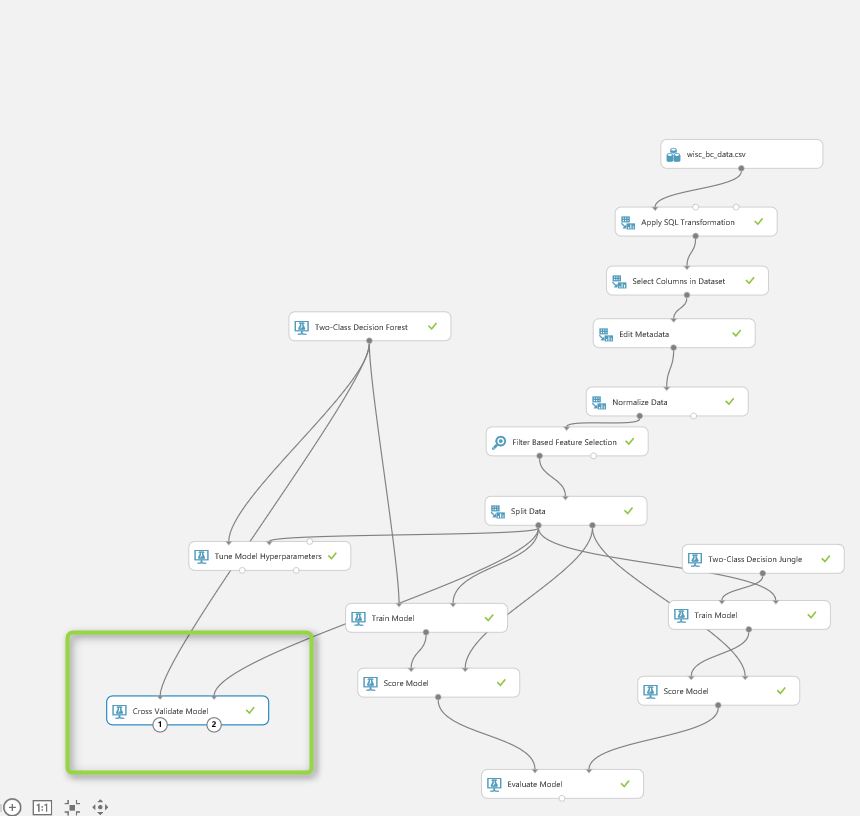
[](http://radacad.com/wp-content/uploads/2017/06/310.png)

After run the code, we will see the below image. Azure Ml created input and output for web service (number 1 and 3 in the below picture). So for creating a web service, we already have the input and output port, just need to run the experiment.

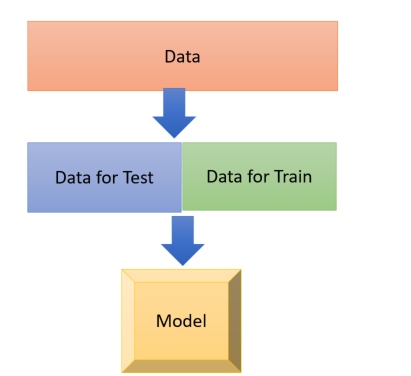
[](http://radacad.com/wp-content/uploads/2017/06/410.png)

# Azure ML Part 9- Cross Validation: Machine Learning Prediction (6)

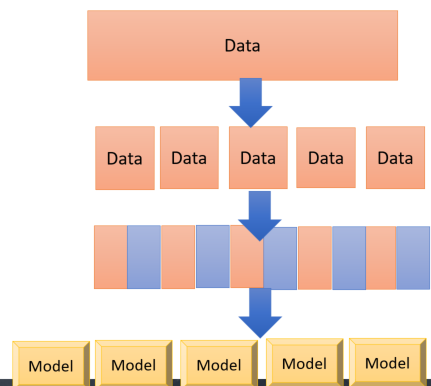
Published Date : June 20, 2017

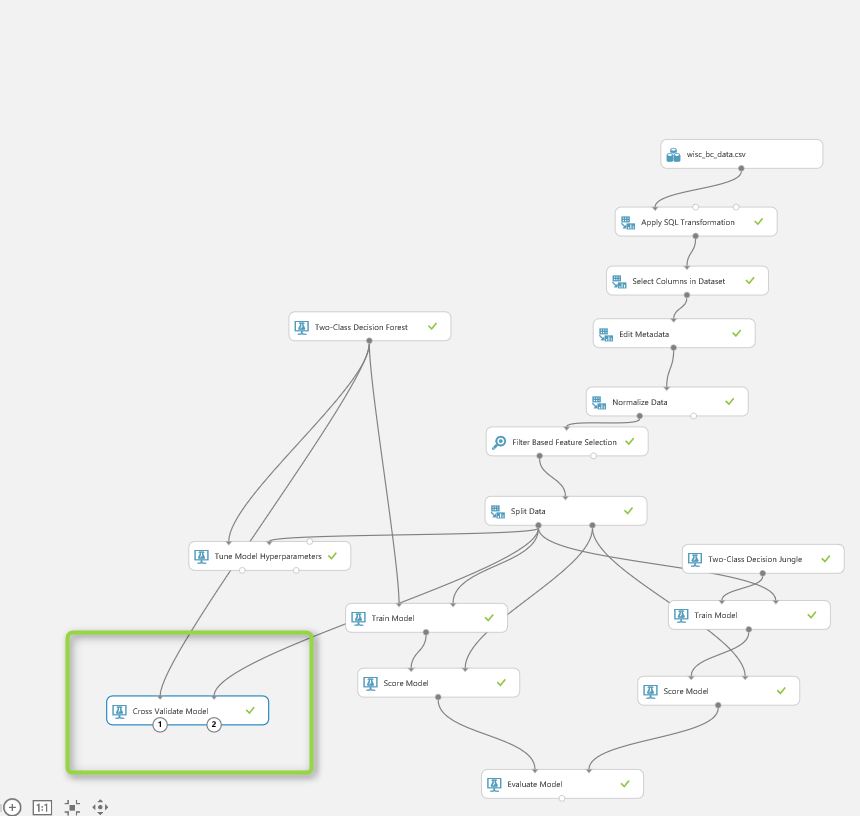
[](http://radacad.com/wp-content/uploads/2017/06/cross0.png)In the previous posts (from Part 1 to Part7), I have explained the whole process of doing machine learning inside the Azure ML, from import data, data cleaning, feature selection, training models, testing models, and evaluating. In the last [post,](http://radacad.com/azure-ml-part-8-tune-parameters-machine-learning-prediction-5) I have explained one of the main ways of improving the algorithms performance name as “**Tune the algorithm’s parameters**” using “**Tune Model Hyperparamrters**” . In this post am going to show another way for enhancing the performance using **“Cross Validation”.**

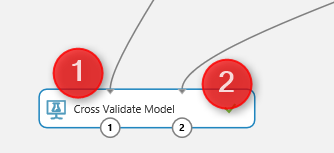
## So what is Cross Validation?

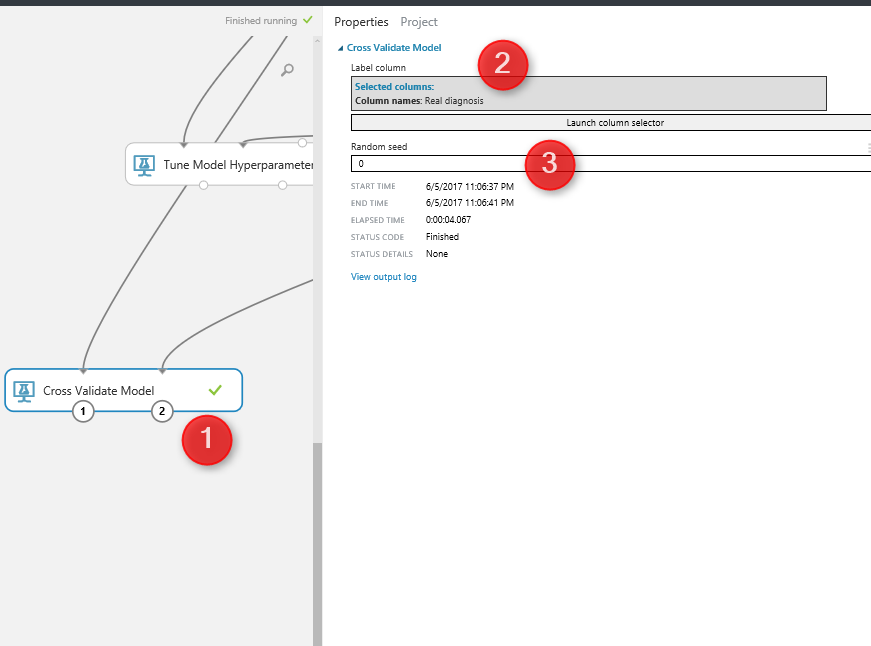
Cross Validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. So we are applying the algorithms in different portion of dataset to check the performance. so the process is 1. get the untrained dataset and model 2. partition data into some folds and sub datasets 3. applied the algorithms on these dataset separately it help us to see how algorithm perform for each dataset, also it helps us to identify the quality of the data set and understand whether the model is susceptible to variations in the data. so  initially, we follow the below process. just split dataset into 2 separate dataset as below. One for Training and one for testing. [](http://radacad.com/wp-content/uploads/2017/06/110.png)

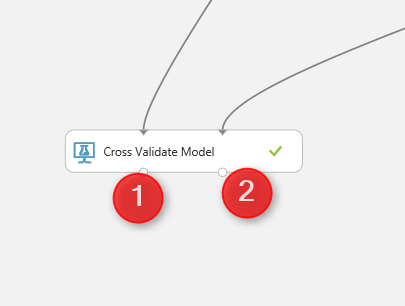
Now we are going to divide our datasets into multiple folds as below picture. As you can see in the below picture, I have 5 different datasets, which I applied a model to each of them to see what is the performance on each dataset

, [](http://radacad.com/wp-content/uploads/2017/06/210.png)

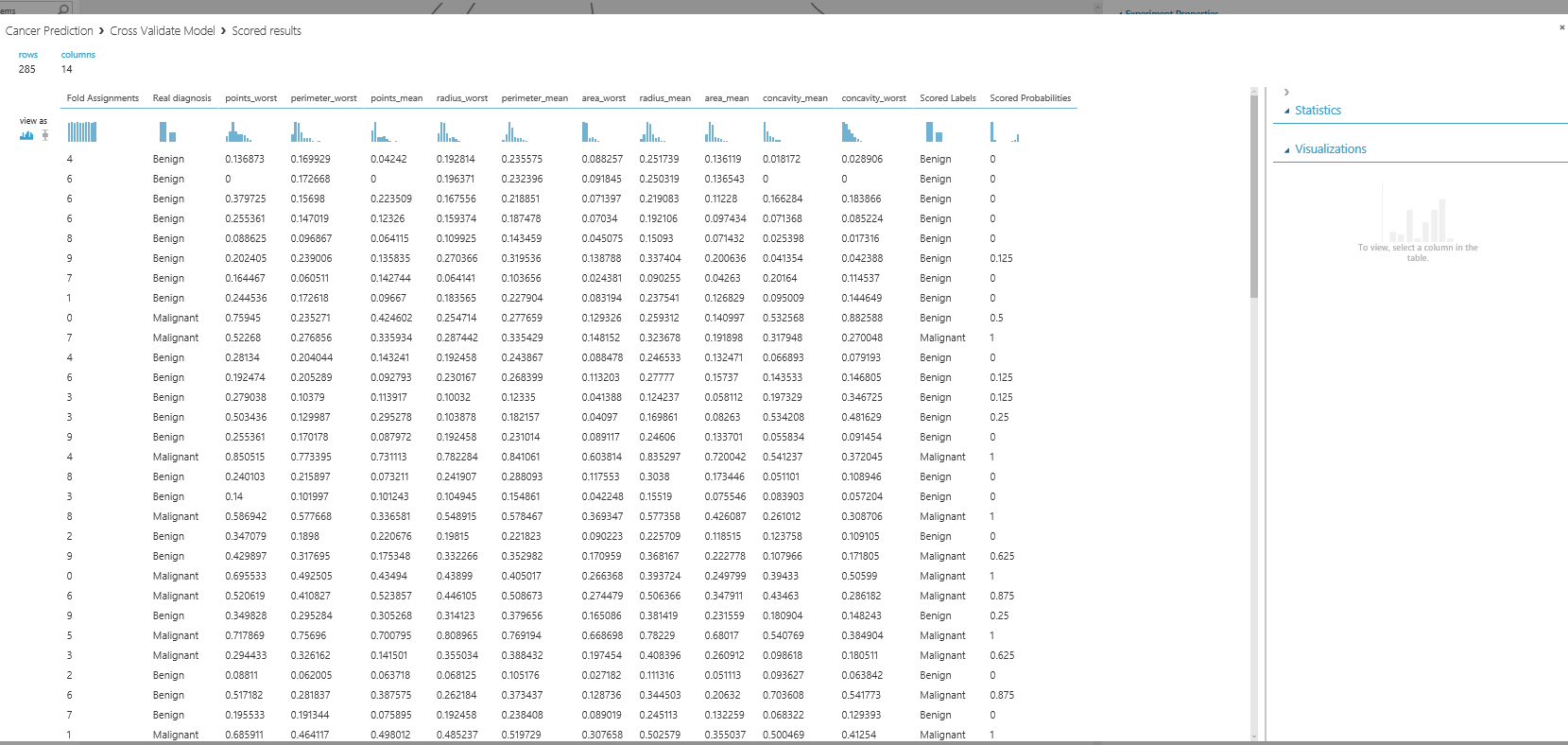
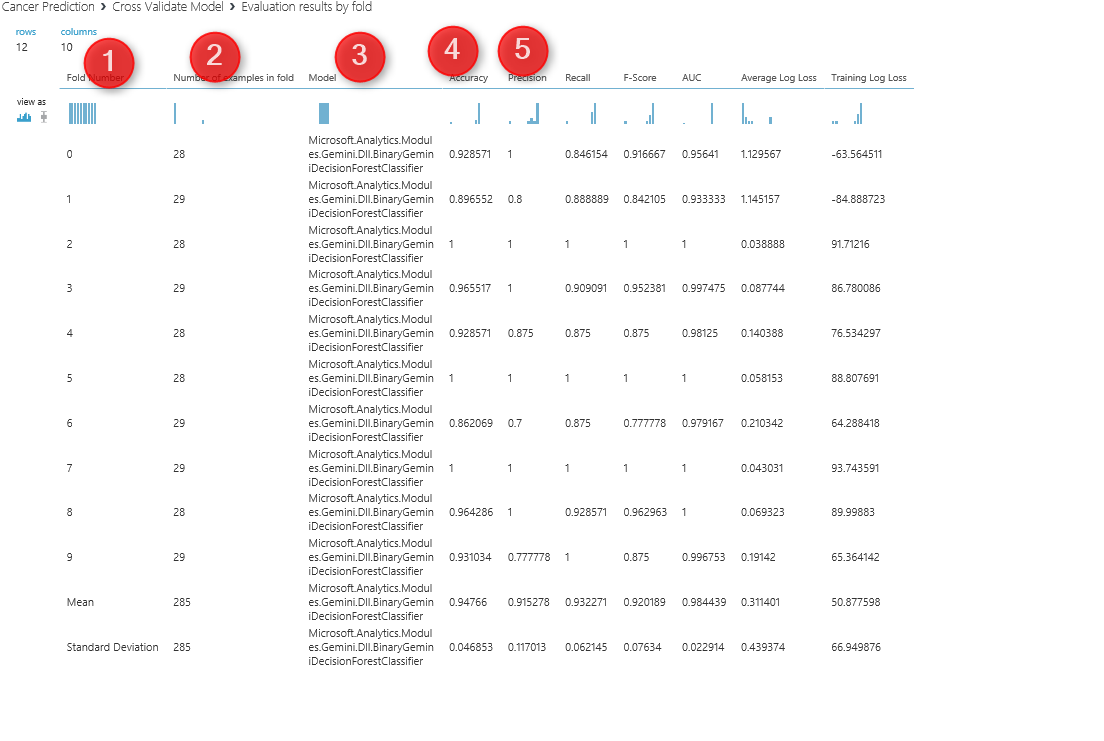
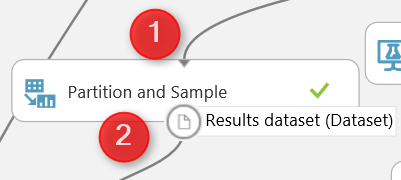
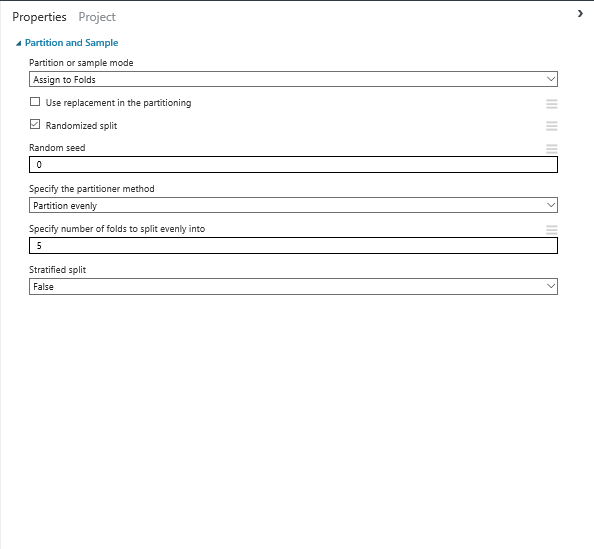
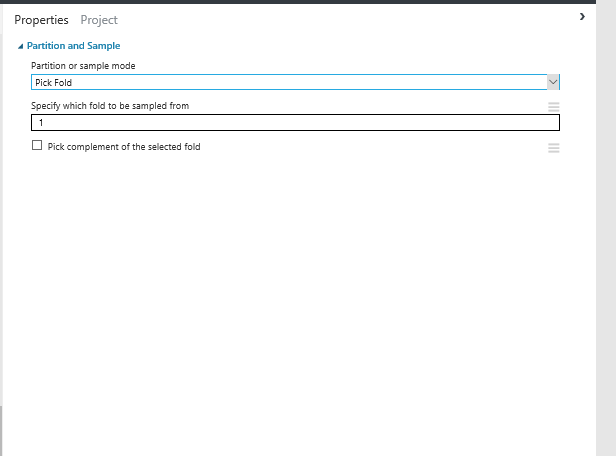
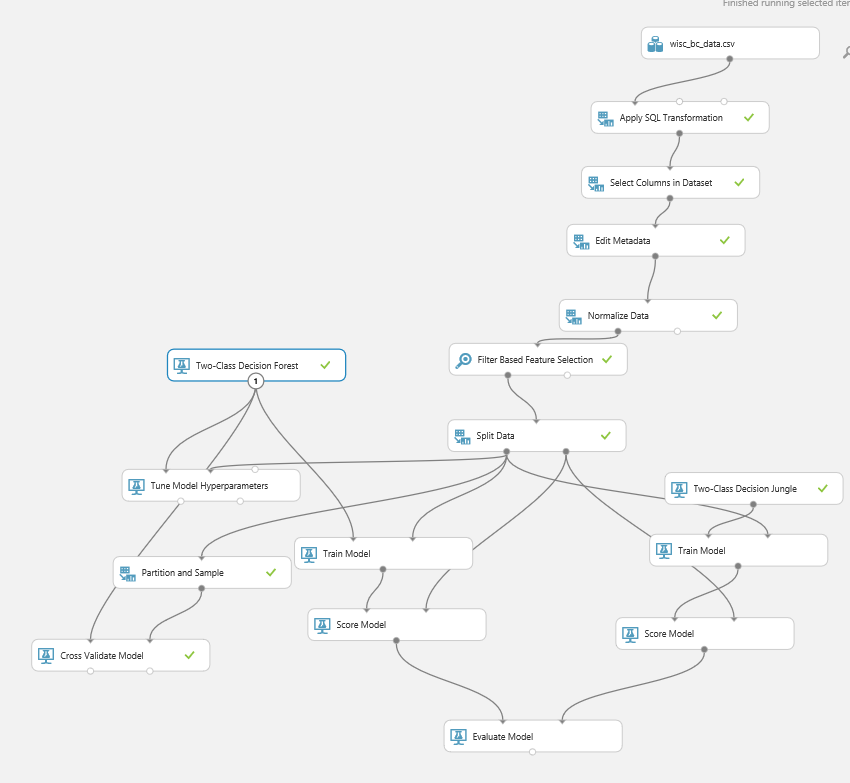
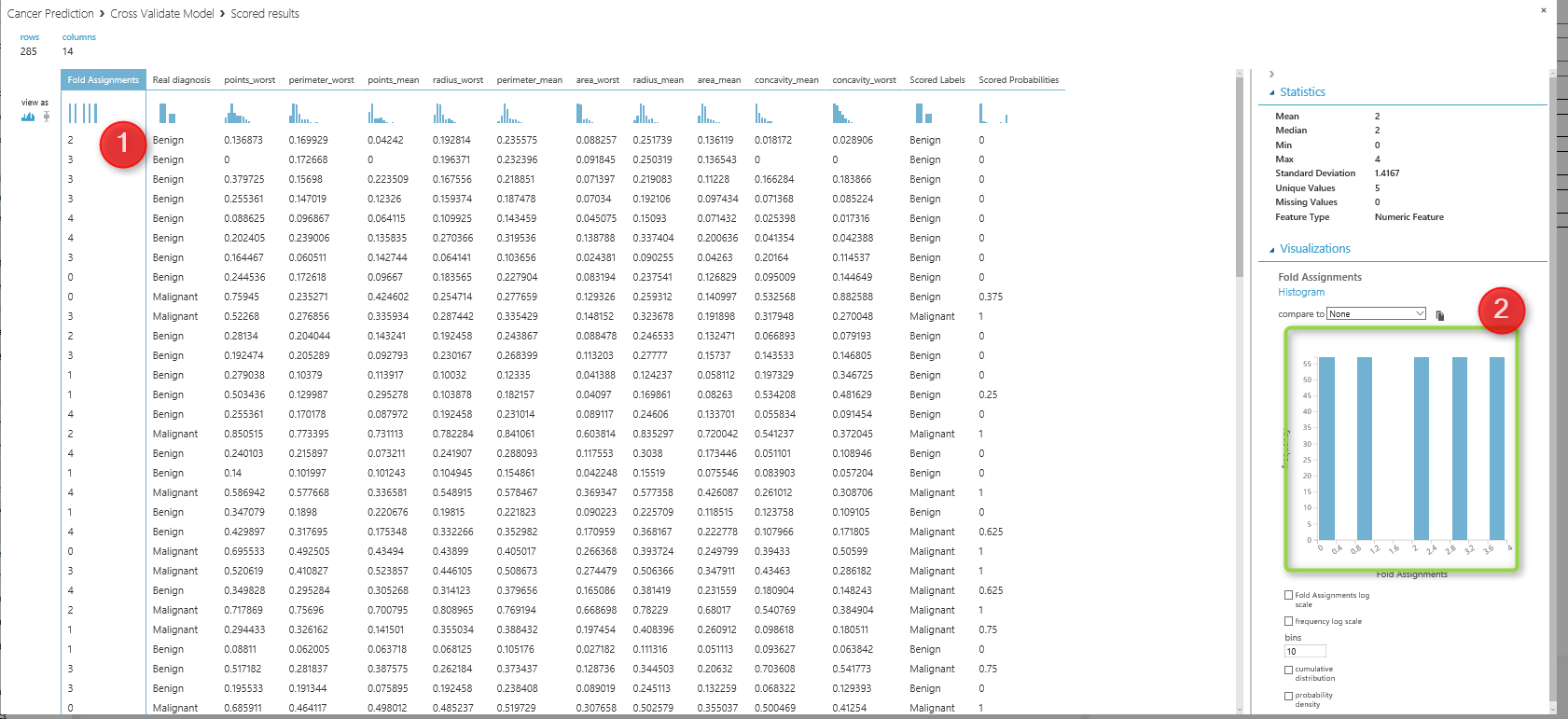
In Azure ML, we able to do the Cross Validation with the component with same name as “**Cross Validation Model**” . Cross Validation model works both for training and scoring (testing) the model, so there is no need to add these components (See the below picture) [](http://radacad.com/wp-content/uploads/2017/06/cross0.png)As you can see in above picture, it get one input from mode (decision forest) and the other input come from

the data for training. [](http://radacad.com/wp-content/uploads/2017/06/cross.png)

By clicking on the “**Cross Validation Model**” component, we will see a properties panel that has just two parameters to set, the prediction column that is “real Diagnosis” and the Random Seed.   [](http://radacad.com/wp-content/uploads/2017/06/cross5.png)

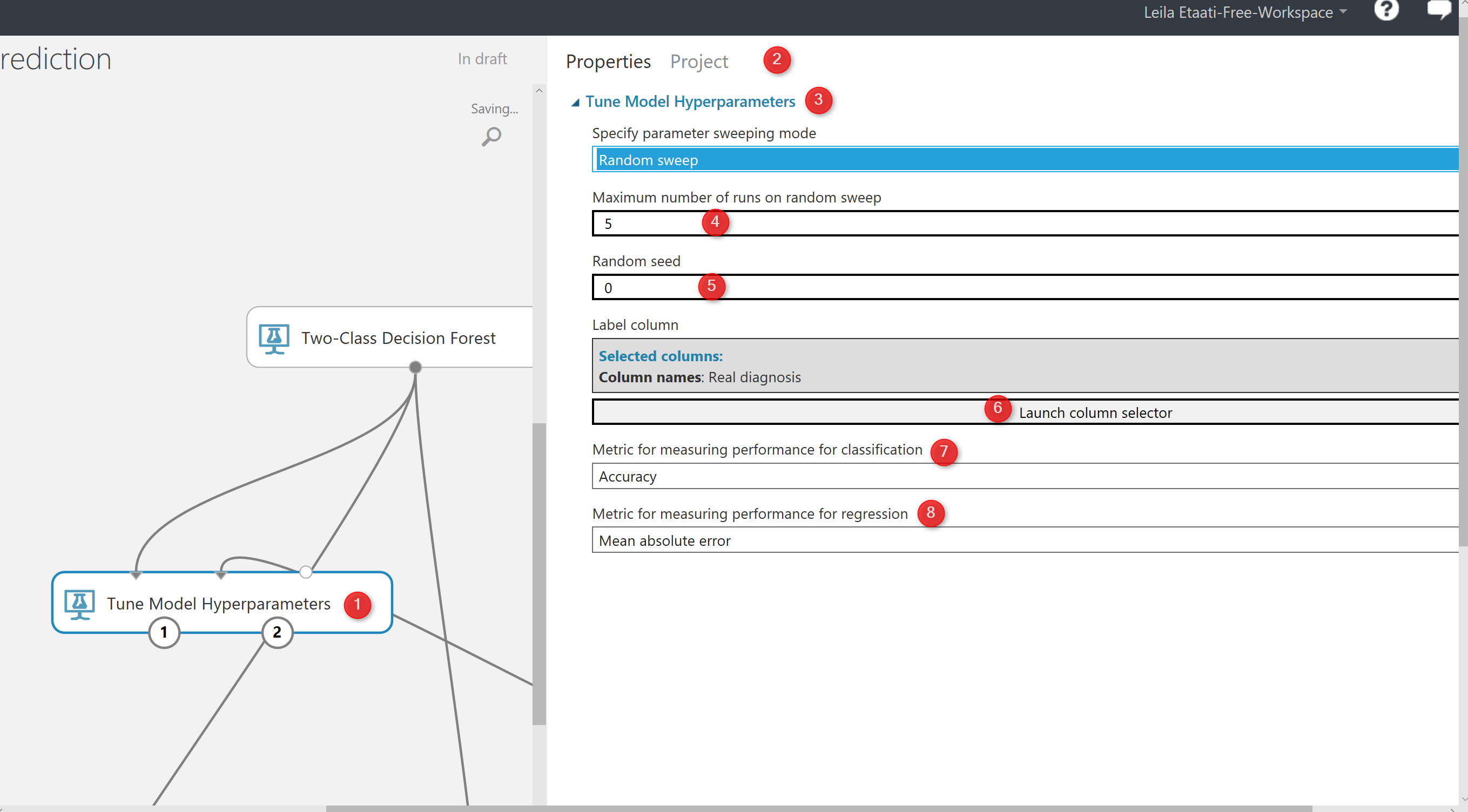
We run the experiment, Then ,by right click on the left side output node of Cross Validation model (below picture number 1), and click on the visualize icon, you will see the result for prediction.   [](http://radacad.com/wp-content/uploads/2017/06/cross1.png)

The below windows will show up, as you can see in the first column, the fold number (dataset number) is in Colum 1, and two last columns (column 13 and 14) show the prediction result and its probability to happen.

[](http://radacad.com/wp-content/uploads/2017/06/cross3.png)By clicking on the other output node of the cross validate  model, (number 2 in previous picture), you will see the analysis for each dataset (see below picture) so for first column (Fold) we have the dataset number or fold number, the second columns show the number of data in each fold. By default, we have 10 folds, but you can change it using the “Partition and Sampling”  component (will talk about it later in this post).The third column is about the model that we run. From column 4 to 10, you can see the accuracy measures that shows the performance of the algorithms on each dataset. so as you look into accuracy in column 4, you will see that the value range from 89 to 1 which good and shows the dataset is pretty well.     [](http://radacad.com/wp-content/uploads/2017/06/cross4.png)  if you wish to specify the number of folds yourself, then you have to use the a component name **” Partition and Sample”.** See below picture. This component get one input from split dataset and one output that has the data for cross validation [](http://radacad.com/wp-content/uploads/2017/06/sample3.png)if you click on the component, you will see the properties panel in right side. [](http://radacad.com/wp-content/uploads/2017/06/sample.png)As you see in the above picture, there are some parameters that we able to assign. First for Cross validation, I have choose the “**Assign to Folds**” value for partition and sample mode.  Also I have ticked the option for “**Random Split**” to decrease the possibilities of biased data selection.   And finally for Specify number of folds to split I choose 5. so instead of 10 folds of data in cross validation, now we have 5 folds. Also there is a possibility to pass one data fold to cross validation by selecting “**Pick Fold**” as you can see in below picture [](http://radacad.com/wp-content/uploads/2017/06/sample1.png)so we have below  process [](http://radacad.com/wp-content/uploads/2017/06/sample4.png)as you can see in above picture, I have connect the output of the partition and sample to the input of the cross validation method. and I run the experiment , so I saw below data. I click on the first column (fold assignment), and in the right side of the windows, There is a chart that shows the summary of data. you see there that we have 5 folds now  instead of 10. [](http://radacad.com/wp-content/uploads/2017/06/sample6.png)  In the next posts, I will talk about the creating web service from azure ml model, how we can use it in Excel or other application. https://en.wikipedia.org/wiki/Cross-validation\_(statistics) https://msdn.microsoft.com/library/azure/75fb875d-6b86-4d46-8bcc-74261ade5826

# Azure ML Part 8- Tune Parameters: Machine Learning Prediction (5)

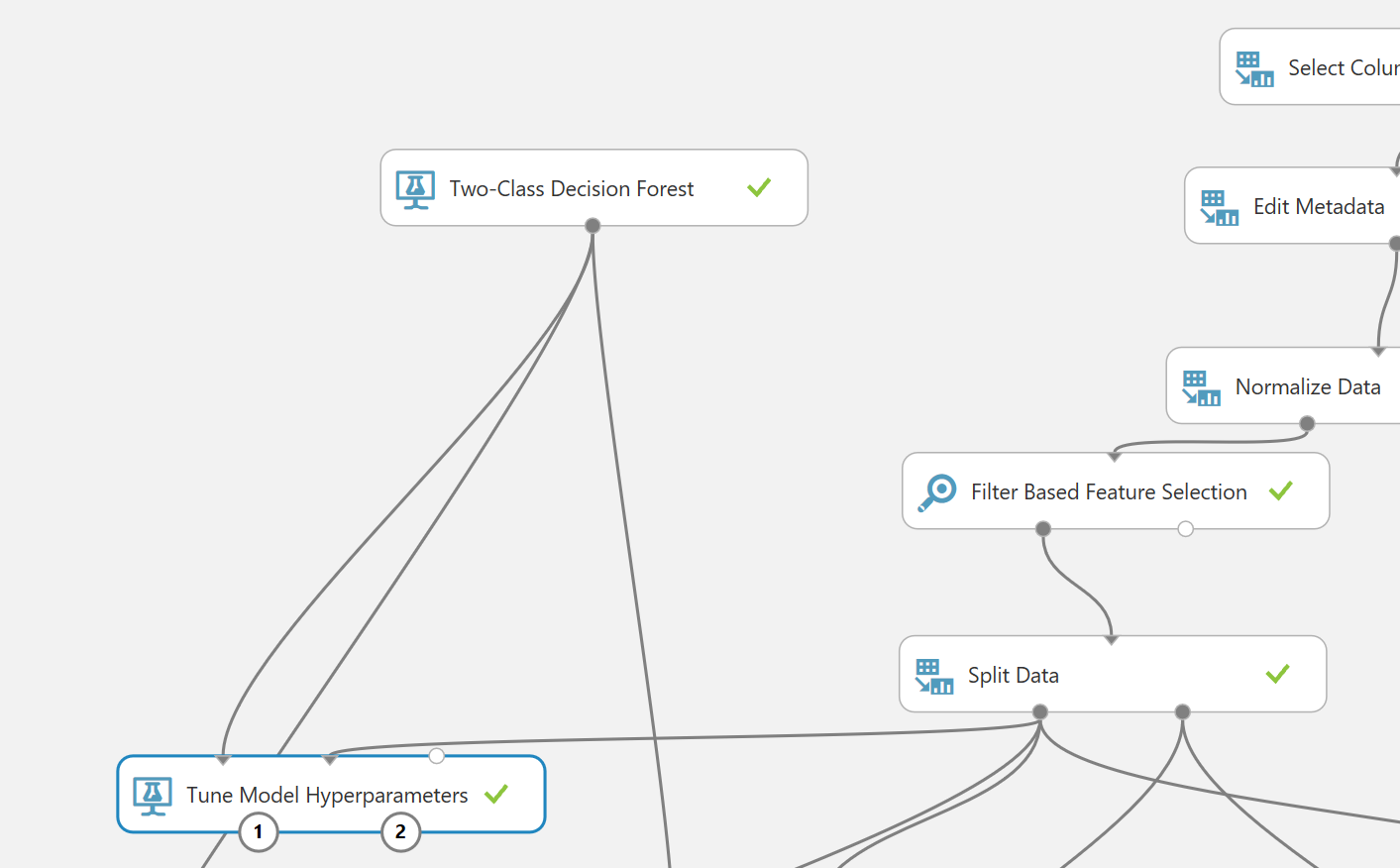
Published Date : June 19, 2017

[](http://radacad.com/wp-content/uploads/2017/06/random-sweep.png)

In the previous posts from Part 1 to [7](http://radacad.com/azure-ml-part-7-a-machine-learning-prediction-scenario-4), I have explained how to do machine learning with Azure ML. I have explained some of the main components in Azure ML that helps us to do data wrangling, train the model, feature selection and evaluating the result.

The data cleaning such as **SQL transformation**, **select specific columns, remove missing values,  Edit meta data**, and **normalize data**.  Also, I have explained how to find relevant attributes  using **Feature Selection Feature** to identify which feature are more important than the other. Then, I shown how to split data using **Split Data** component to train and test the model. Then, I show how to train and test the model. Moreover, in the last post I have explained the evaluation process **Evaluate Model.**

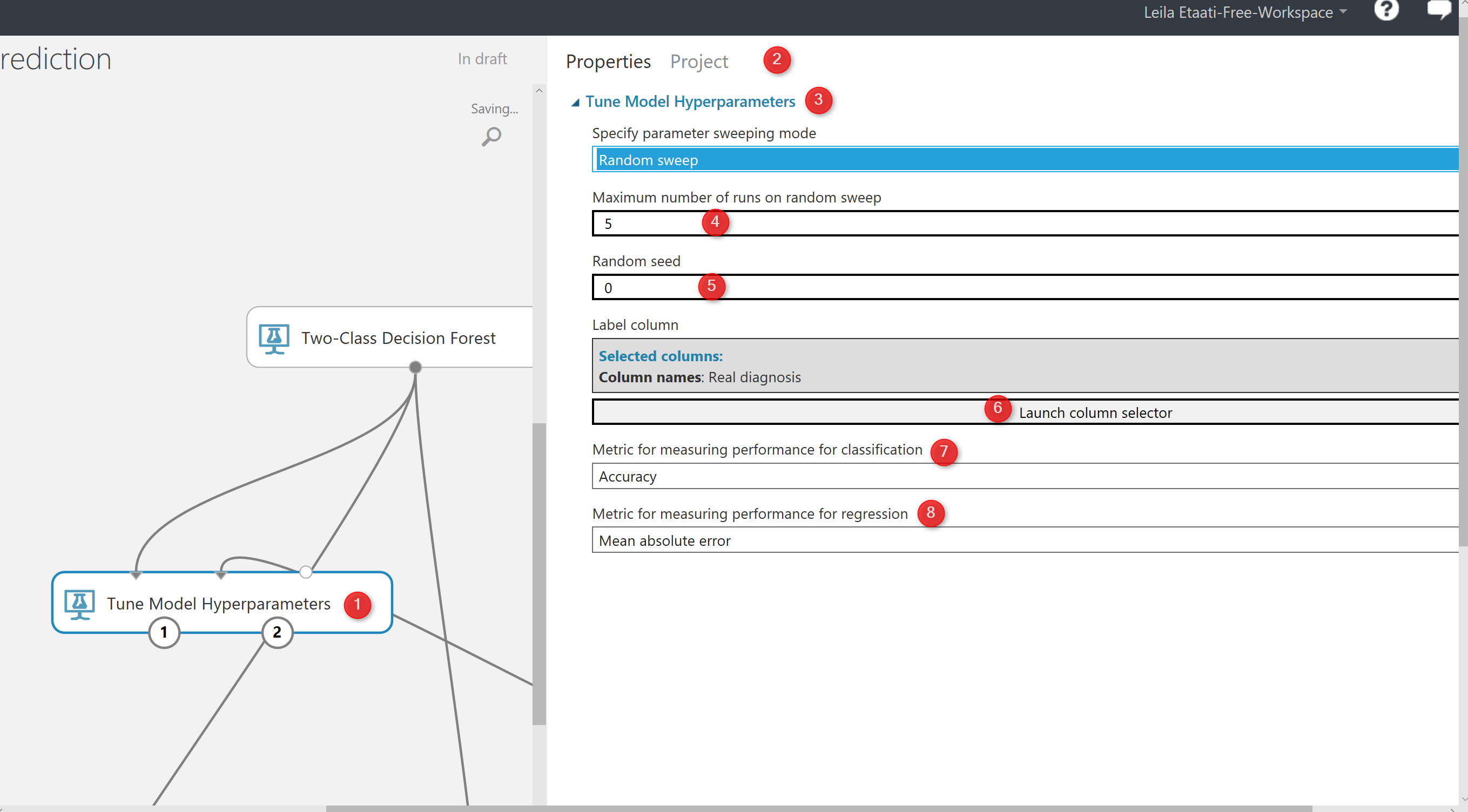
In this post, I am going to show the another way for enhancing the model as “**Try Different  parameters**“. Each algorithms has its own parameters. Choosing the right Parameters for each dataset and algorithms can improve the accuracy. In Azure ML, there is a component name “**Tune Hyper Parameters**” . This component will help us to better improve the accuracy.

[](http://radacad.com/wp-content/uploads/2017/06/hyper.png)

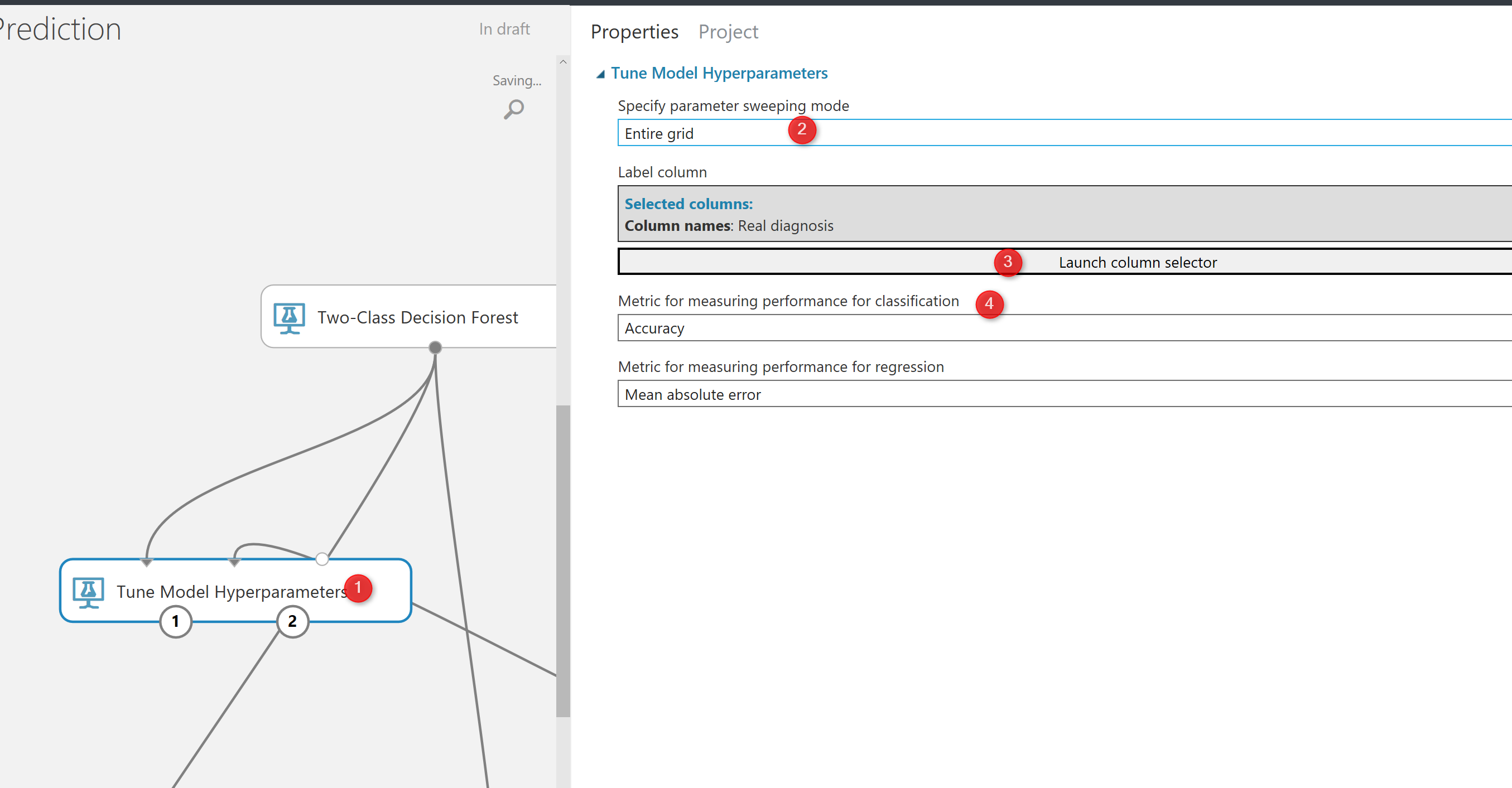
“**Tune Model Hyperparameters**” get two inputs: one for data and for aim of the training model, which comes from  the split data component, another one from algorithm. Actually this component can be replacement of **Train Model.**

if you click on the component, in the right side of experiment area,  you will see the properties panel. as you can see in the below picture, by clicking on the “**Tune Model Hyperparameters**“. We have couple of the parameters that we should set them up. the first parameter is about the method that we want to try different parameters value. In the below picture, first I have choose the “**Random Sweep**“. It performs a set number of training iterations by randomly choosing parameter’s value. So, this component will try different parameter’s value randomly and train the model based on them(Number 3). the second parameters to set up is (number 4) how to specify number of times we have to run the code. This help us to identify with which parameters what accuracy we will have.

Then in the number 6, we have to specify the item we are going to train the model and identify which column we are going to predict. For this example we choose “**Real Diagnosis**“. In number 7 and 8, we have to specify the metric for measuring the model performance. if the problem is classification, then we need go for accuracy or recall measures. If the model is about predicting a value and we using the regression models, then we should select one of the item from number 8 in the picture.

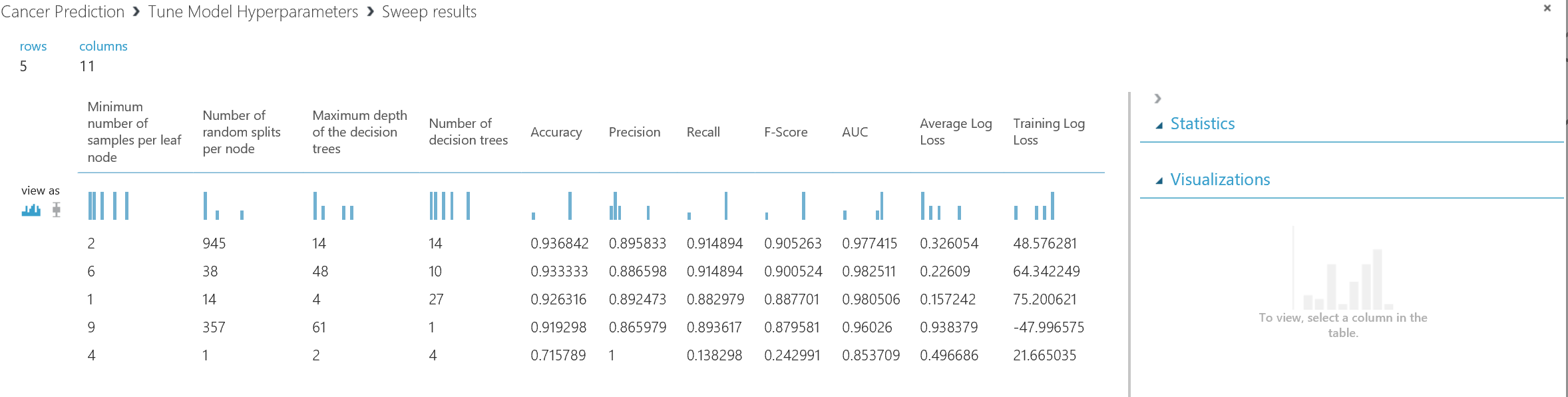
[](http://radacad.com/wp-content/uploads/2017/06/random-sweep.png)

The other option is to test all the parameters against each other that means trying different parameters combinations and identifying the best of them. we call this approach as trying “**Entire Grid**“. as you can see in the below picture in number 1, we select “Entire Grid” as the approach to find the best parameters. as you see here we are not to specify the number of times we run the training model.

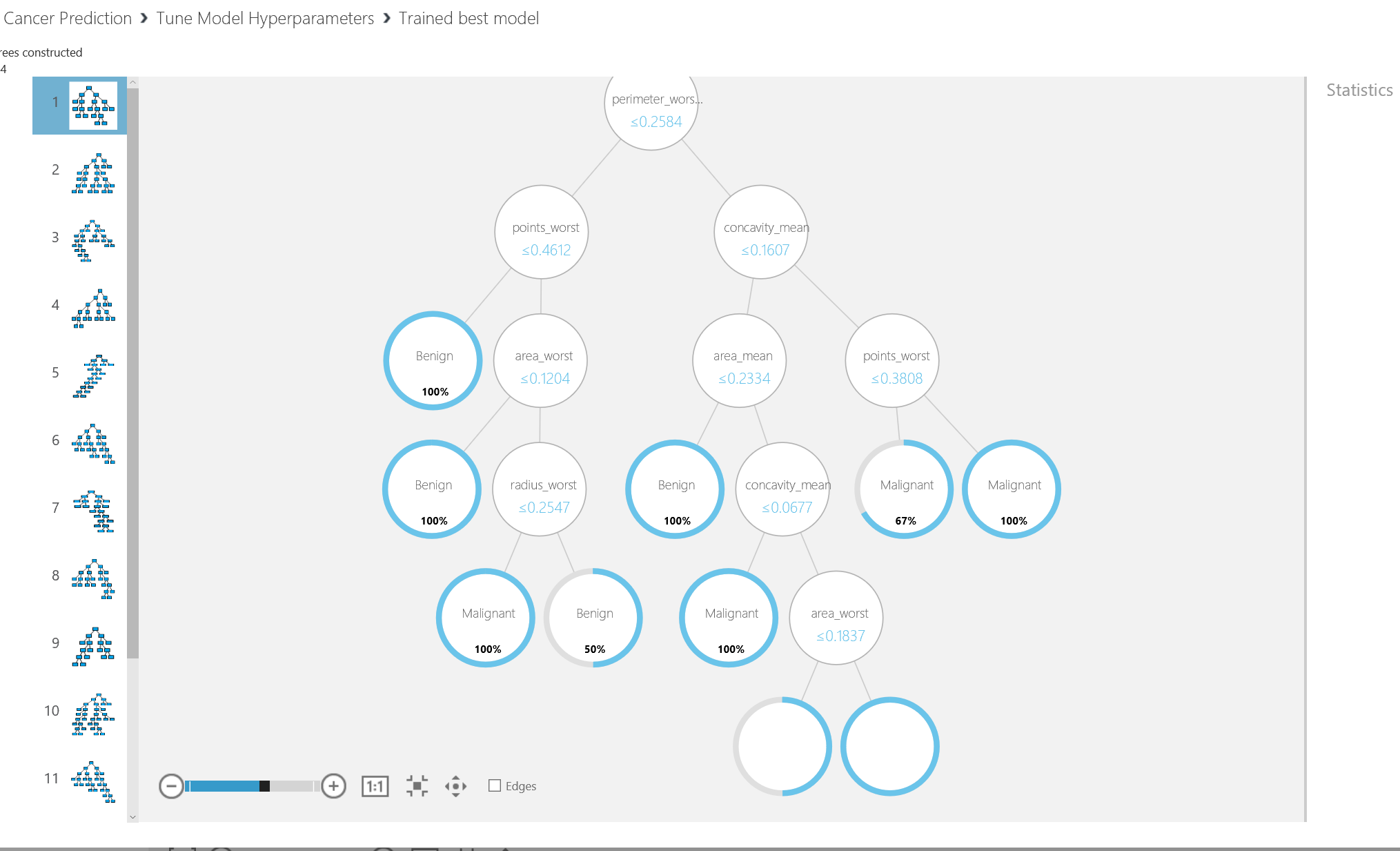
[](http://radacad.com/wp-content/uploads/2017/06/entiregrid.png)

Now, we setup the code, we are going to just run the experiment. after run the experiment you will see we have two output for the “Tune model Hyperparameters”. The first output at the left side has the dataset that shows the parameters value and relative accuracy to them (See below picture).

As you can see in the below picture, the first4 first column are parameters for decision forest algorithm that we have. Such as number of sample per leaf, depth of the decision tree, number of decision tree. as you can see from the below picture, we have 6 other column that shows the accuracy, Precision, Recall, and so forth value that we got based on the selected parameters. so it help us to assign better parameters for our model.

[](http://radacad.com/wp-content/uploads/2017/06/accuracy1.png)

The other output of the “Tune Model Hyperparameter” is to  visualize of the decision tree (see below picture). So you see the different models that has been trained.

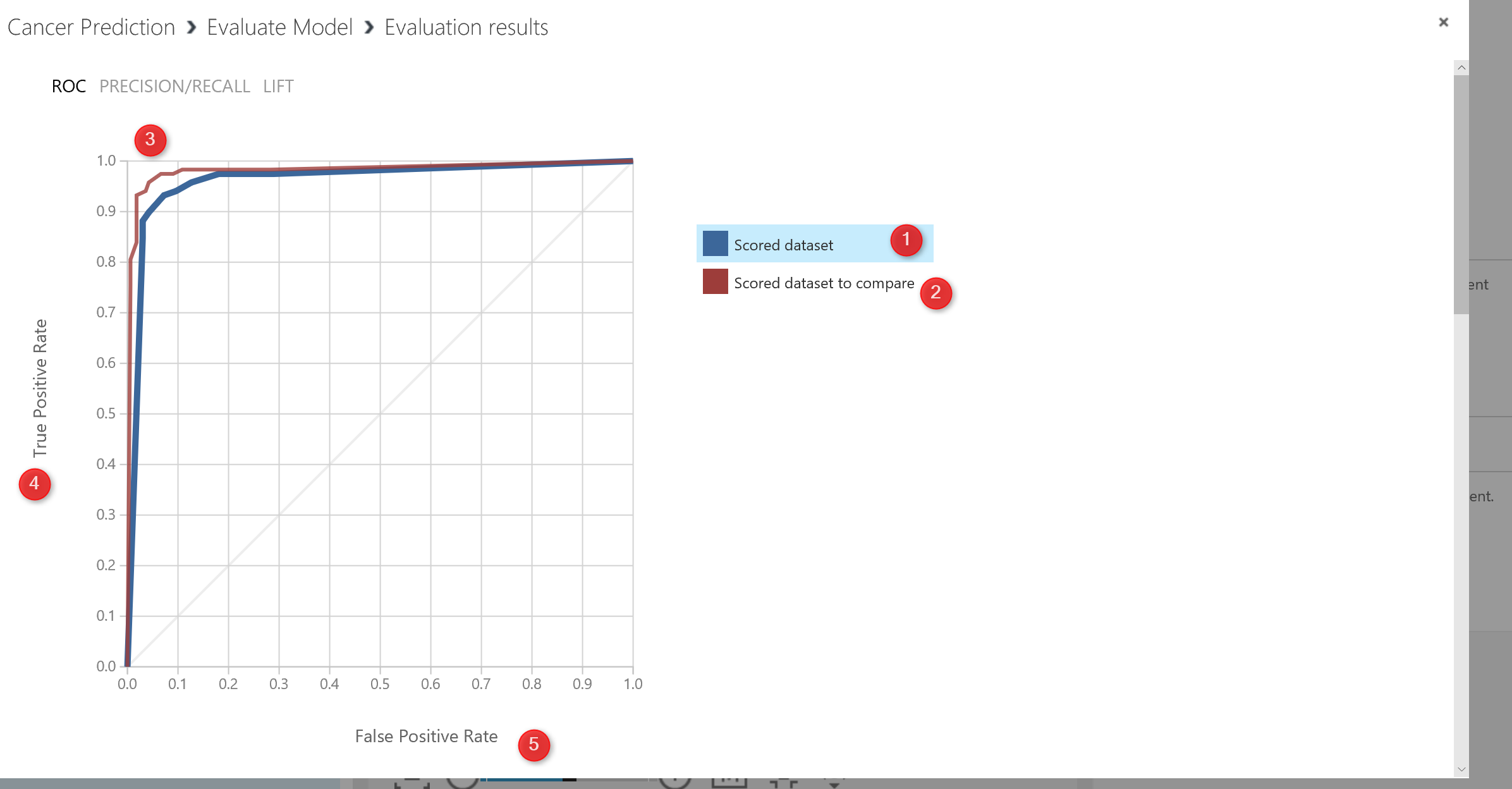
[](http://radacad.com/wp-content/uploads/2017/06/dt.png)

In next post I will talk about the “**Cross Validation**” component that is another way of enhancing the accuracy by trying different datasets.

https://msdn.microsoft.com/library/azure/038d91b6-c2f2-42a1-9215-1f2c20ed1b40

# Azure ML Part 7: A Machine Learning Prediction scenario (4)

Published Date : June 8, 2017

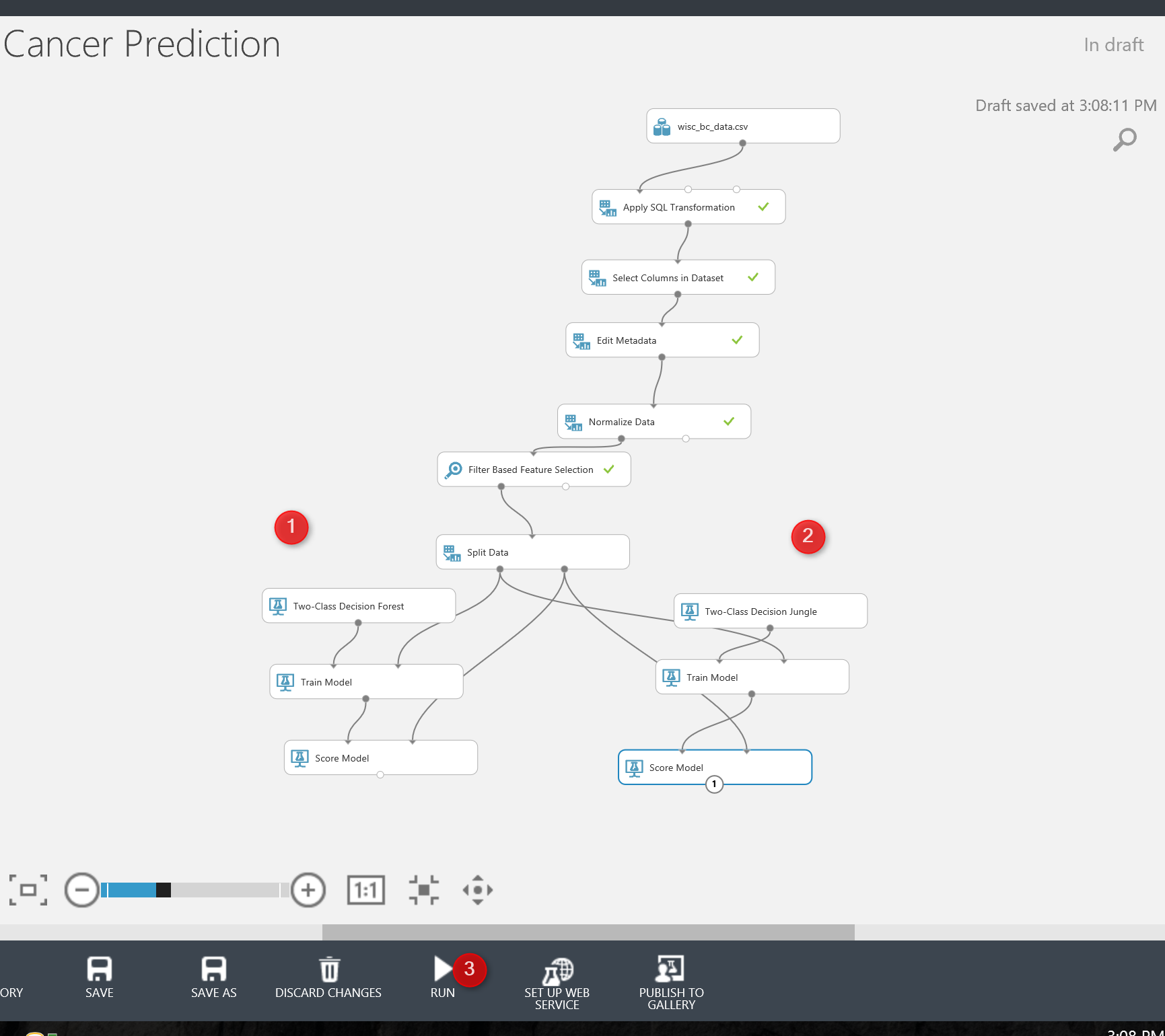
[](http://radacad.com/wp-content/uploads/2017/06/28.png)

In the previous posts from Part 1 to 6, I have explained how to do machine learning process.

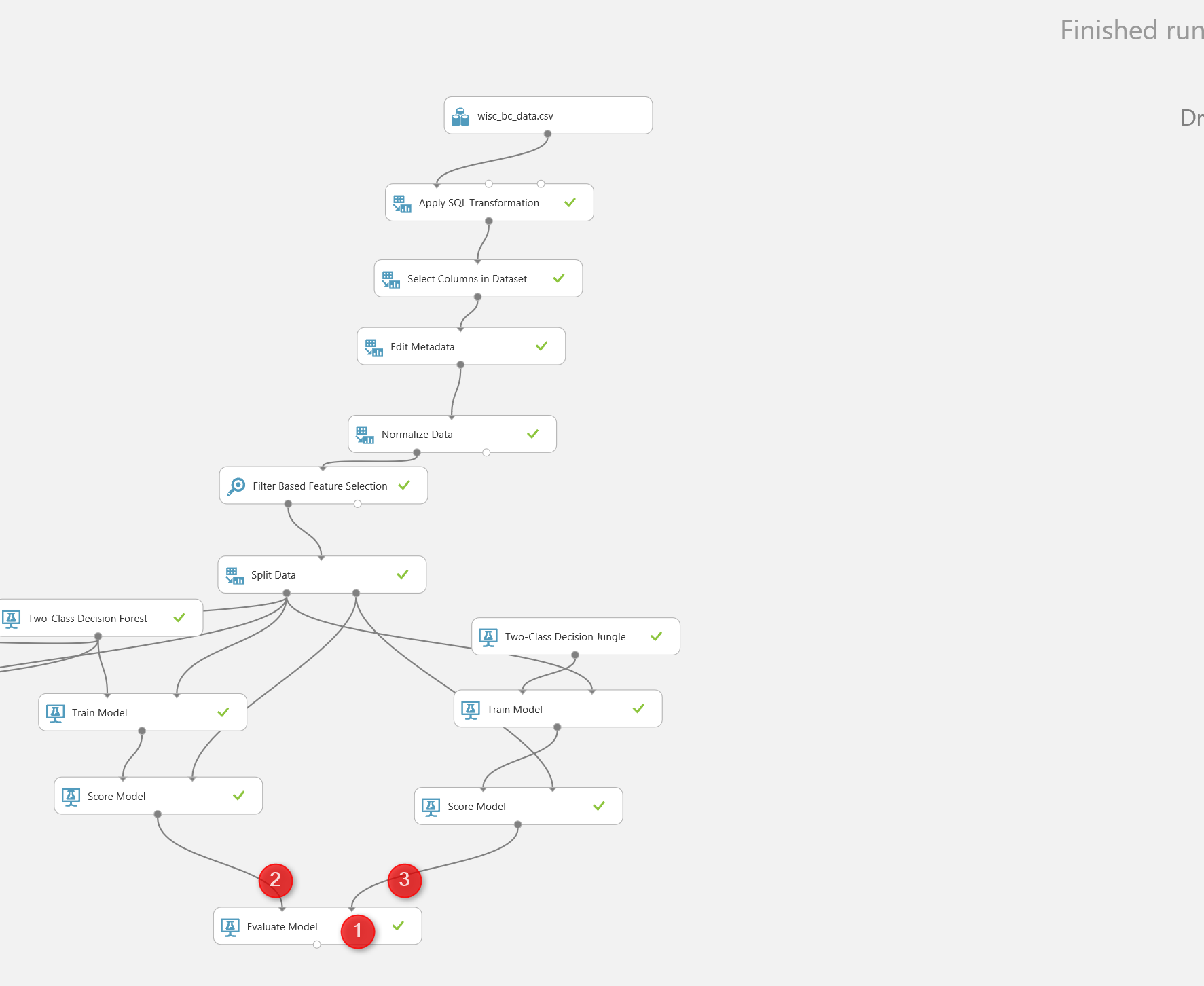
The data cleaning such as **SQL transformation**, **select specific columns, remove missing values,  Edit meta data**, and **normalize data**.  Also, I have explained how to find relevant attributes  using **Feature Selection Feature** to identify which feature are more important than the other. Then I shown how to split data using **Split Data** component to train and test the model. Then, I show how to train and test the model.

In this post I will show how to evaluate the results and also how to interpret the evaluation measures.

So in the previous Part we came up with the below process. we just create a model (see below picture)

[](http://radacad.com/wp-content/uploads/2017/06/26.png)

Now we need to evaluate the model. There is a component name **Evaluate Model.** Just drag and drop it to the experiment area (see below picture) . Evaluate model has two input node. In this experiment, we have two models, so we have two score model (as you can see in the below picture). Then connect the output node of Score Model (Testing) to the input of **Evaluate Model** (see number 2 and 3)**.**

[](http://radacad.com/wp-content/uploads/2017/06/evaluate.png)

Now lets check the Evaluate model components.

just right click on the output node of the **Evaluate Model t**o see the result of the evaluation via **Visualize icon**. You will see the below image right after.  We have a chart as below. the legend show to colour as blue and red. The blue one is related to the left side algorithm (in our experiment) and the red one related to the right side algorithm)used to check the performance a two-class algorithm.

in this chart the first chart is “ROC” is **[receiver operating characteristic curve](https://en.wikipedia.org/wiki/Receiver_operating_characteristic).** The plot is use to check the performance of the binary classification.  we have two value in x and Y  axis. **True Positive and False Positive Rate.**

But, what is **True Positive and False Positive** ?

In a Two-Classification Problem, the prediction and real world scenario may have below situation. The evaluation will look at the test dataset to compare the

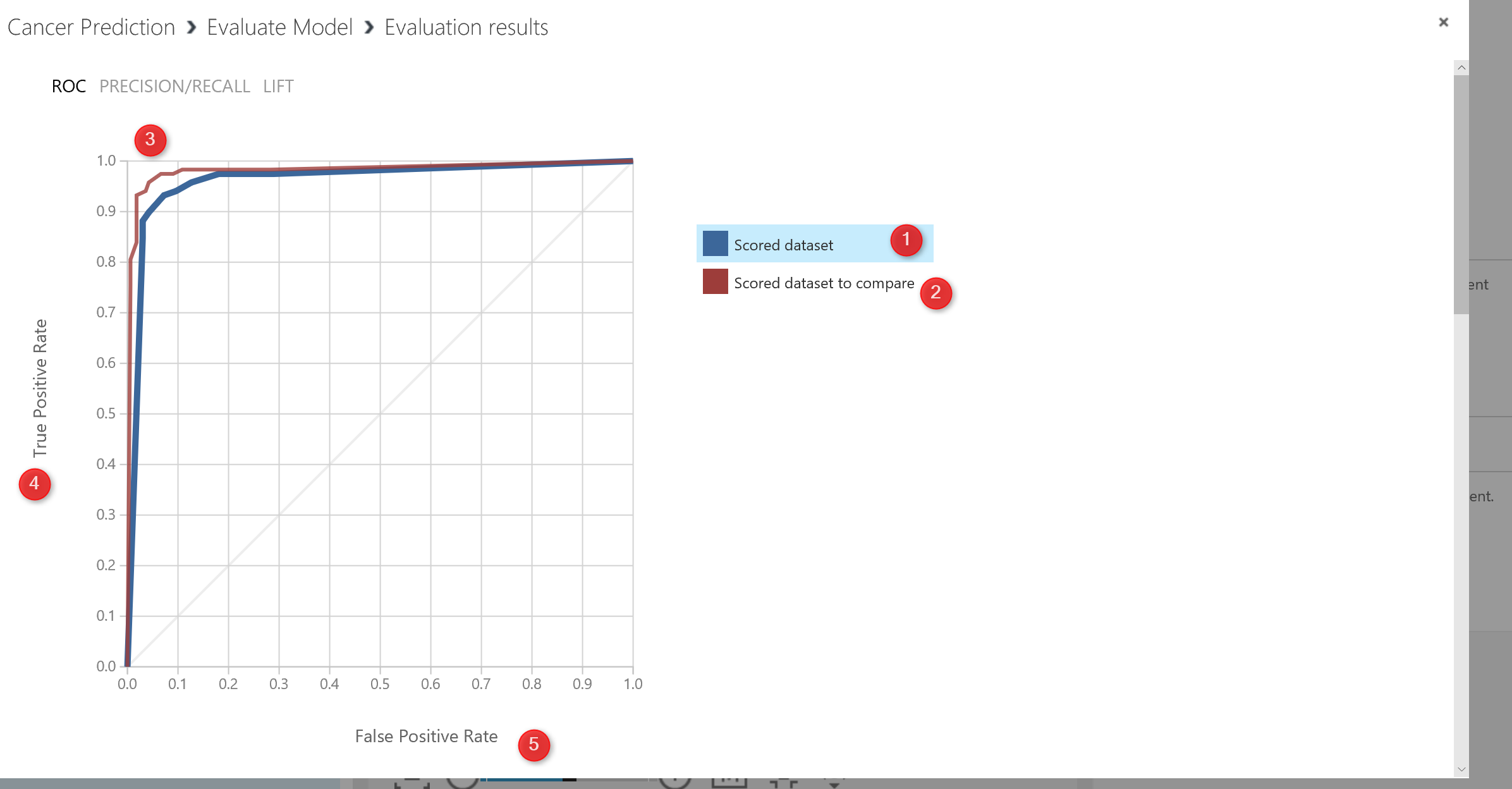
**True Positive**: if values in real world is 1 (for instance in our example, Benign) and out algorithm correctly predict patients will Benign so we call it true positive (the higher number better) :)

**False Negative:** In the real world scenarios, the patient become Benign but in our algorithm predict they will Malignant. which means in real case it is true but in Prediction it is Negative. :(

**False Positive:** In real world scenario, the patient become Malignant but in our model, we predict they are  Benign which is incorrect so we call it False Positive. :(

**True Negative:** In real word scenario, the patient become  Malignant and our machine learning pr3ediction is that they will become Malignant :)

So in **ROC** chart as below we are going to see the relation between True Positive and False Positive.  As you can see below picture, as the line close to the Y axis, it means that it is has better performance. That in below picture both charts has good performance.

[](http://radacad.com/wp-content/uploads/2017/06/28.png)

if you scroll down the evaluate model page, you will see the below picture . as you can see we have some values that help us to evaluate the mode’s performance.

first we have the True Positive, False Negative, true negative and false positive value in the right side of the page (below picture number 1).

Also we have some other measure for evaluating the model performance:

**Accuracy**:  The accuracy is about the (∑TP+∑TN)/Total population

So in the below picture in number 2, you will see it is 93%. as much as accuracy higher better.

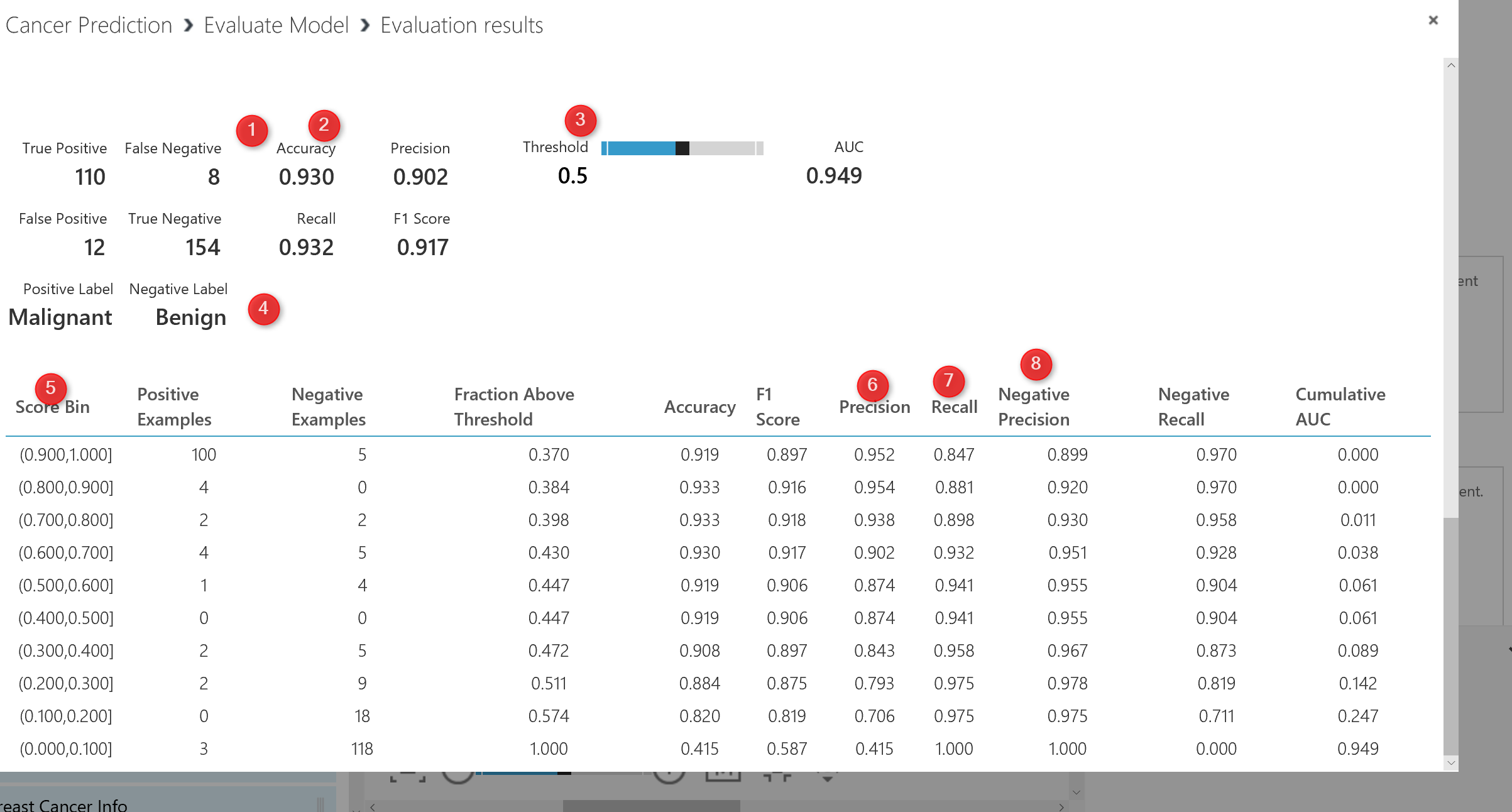
**Recall:** The recall is the ∑TP/(∑TP+∑FN). That measure how model is good in predicting the true positive cases comparing the total case that happen true. in our example is 93%.

**Precision:** The precision is the ∑TP/(∑TP+∑FP) . in our example is 90%

**F1 Score:** The F1 Score is the ∑2TP/(∑2TP+∑FP+∑FN) which is 91%.

The bar chart for **Threshold and AUC** (the area under the ROC curve) shown in below picture. It shows 94% which is great. The Threshold for 50% . so it always good to have a AUC more than

threshold, which means more True positive than False Positive.

[](http://radacad.com/wp-content/uploads/2017/06/29.png)

Over all, all the two algorithm perform well.

As I have mentioned it, to enhance the performance algorithms, there are three main approaches:

**1.** it always recommend to run **multiple algorithms** to see which algorithms best fir the dataset and able to predict the data better.

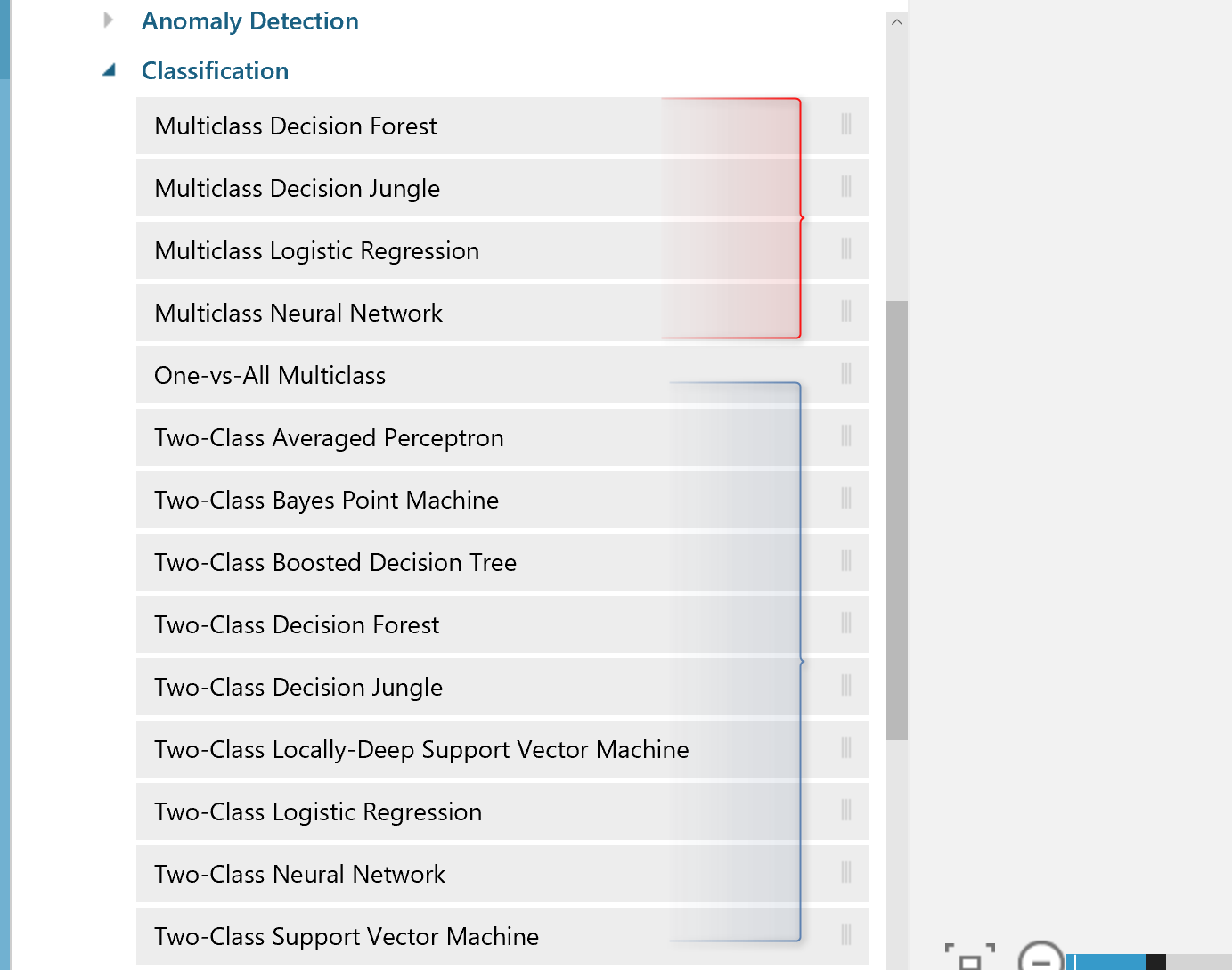
2. Do **Cross Validation,** to check the result of Machine Learning on each dataset sample

3. Check the different parameters for algorithms. each algorithms has its own parameters list, it is important to find the best parameters that cause better performance. This can be done in Azure ML via **Hyper Tuning Parameters.**

In the next Post I will show how to enhance the performance of model via **Cross Validation and Hyper Tune Parameter**  Component.

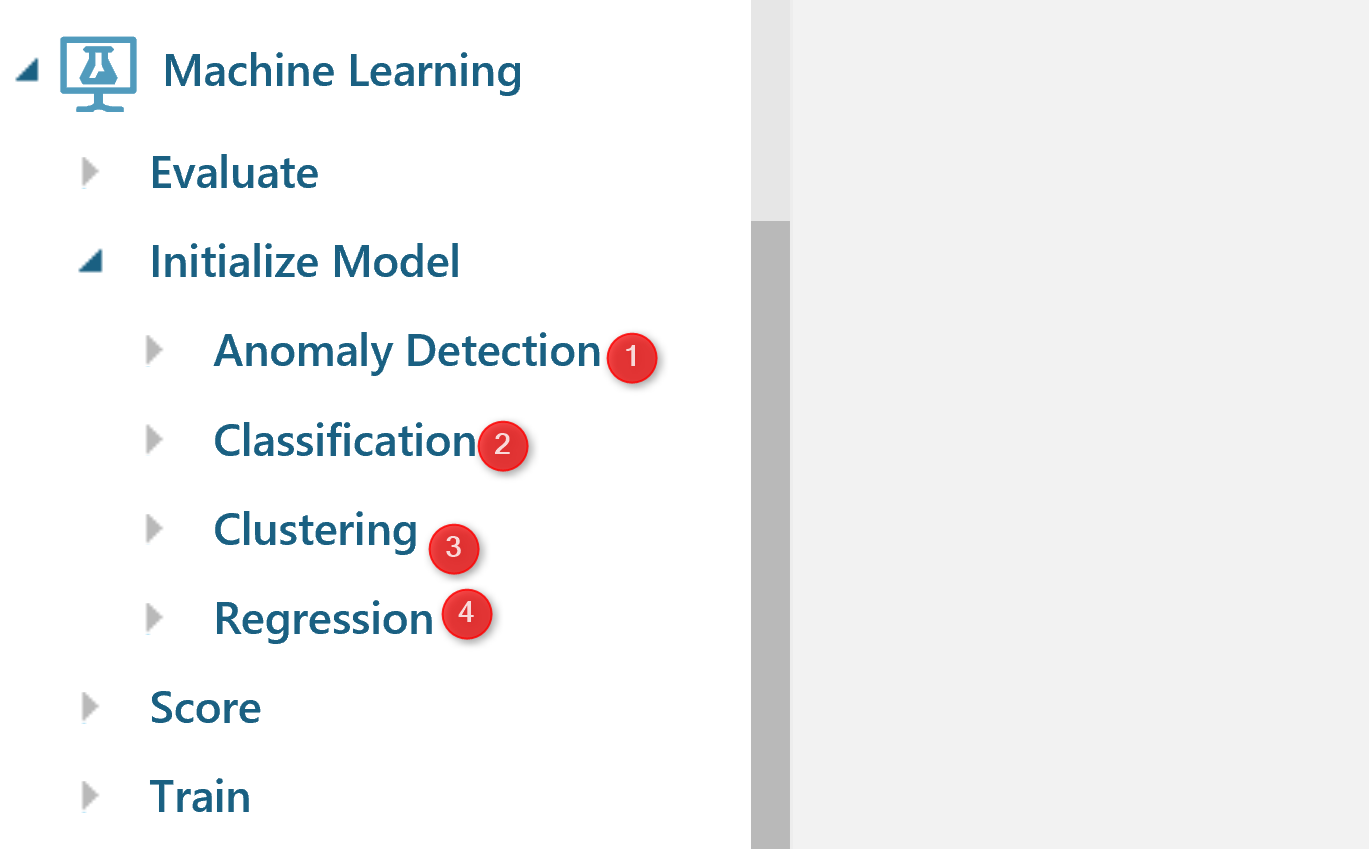
# Azure ML Part 6: A Machine Learning Prediction scenario (3)

Published Date : June 6, 2017

[](http://radacad.com/wp-content/uploads/2017/06/56.png)

In previous posts ([Part 4](http://radacad.com/azure-ml-part-4-a-machine-learning-prediction-scenario-1) and[Part 5](http://radacad.com/azure-ml-part-4-a-machine-learning-prediction-scenario-2)), I have explained some of the main components of Azure ML via a prediction scenario. In post one the process of data cleaning (using **SQL Transformation, Cleaning Missing Value, Select specific Columns, and Edit Meta Data)** has been explained. and in the second Post, I have explained how to apply **Data normalization, Feature selection** and how to **split Data** for creating **Training and Testing** Datasets has been explained.

In this post I will continue the scenario of predicting patient diagnosis to show how to select **appropriate algorithms**,  **Train Models, and Test Models (Score),**In the next post I will show how to evaluate a model using **“Evaluate Model” component,**also how to find the best parameters for algorithms using **Tune Model Hyperparameters,** and also check the algorithm evaluation result on different portion of the dataset using **Cross Validate** component.

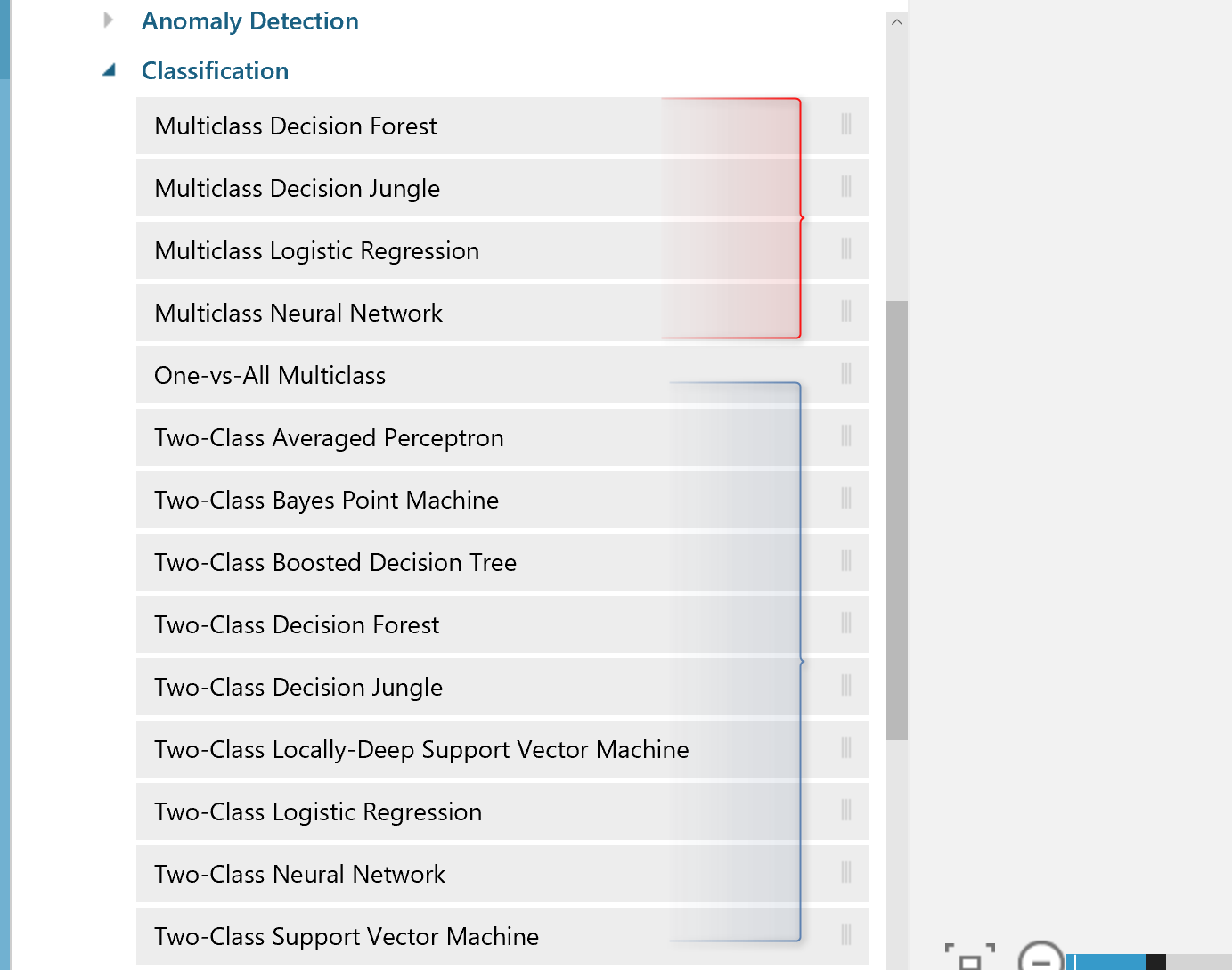
[](http://radacad.com/wp-content/uploads/2017/06/azurealgo.png)

**How to choose Algorithms**

There are many algorithms that address the different types of machine learning problems. In Azure ML, there are four types of the algorithms  : **Anomaly Detection, Classification, Clustering, and Regression.**

**Anomaly Detection** : this types of algorithms are helpful for finding cases that not follow the normal pattern of the data such as finding the fraud in credit cards (above pictur3e Number 1).

**Classifications:** Classification can be used for predicting a group. For instance, we want to predict whether a customer will **stay** with us **or not**, or a customer will get a **Bronze, Silver, or Gold** membership. In Azure ML we have two sets of algorithms for classification as : **Two-Class and Multi-class algorithms** (See below Picture)

[](http://radacad.com/wp-content/uploads/2017/06/56.png)

**Clustering**

Clustering algorithms like k-mean clustering (I have a post on it) more used for finding the natural pattern in data without doing a prediction.

**Regression**

Regression algorithms is used for predicting a value. for instance predicting the sales amount in a company.

In this scenario we are going to predict whether a  patient will be **Benign or Malignant.** So I have to use a two-class algorithms.

So I am going to use **two-class classification algorithms**.

To get a better result, it always recommend to :

**Try different  algorithms** on a dataset to see which one better able to predict.

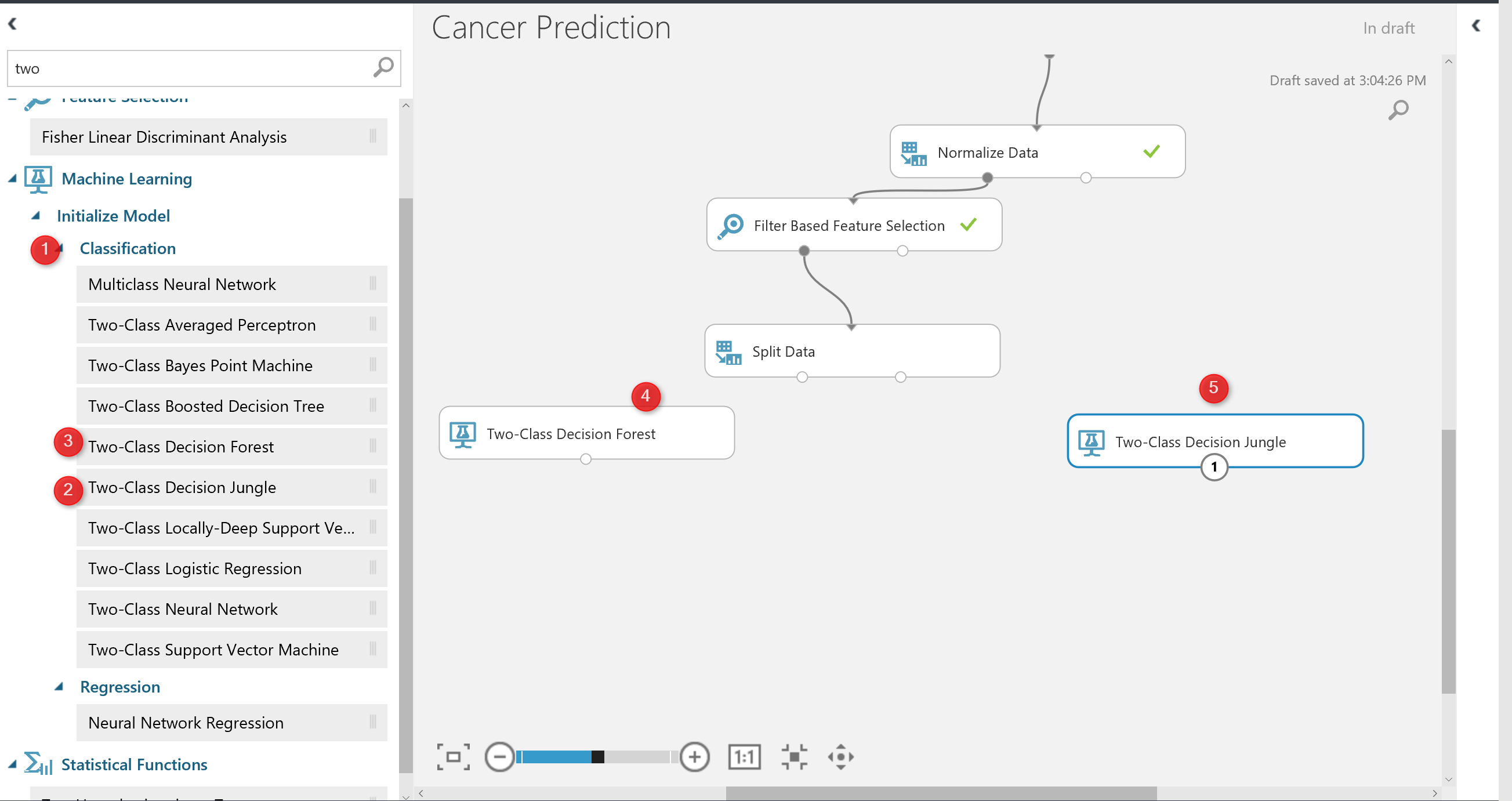
**Try different Dataset** it is good idea to divide a current dataset to multiple one, and then check the algorithms evaluation in different parts of a dataset. in this scenario I am going to use a component name “**Cross Validate Model”**

**Try different algorithms Parameters.** Each algorithms has specific parameter’s values, It always recommended to try the different parameter’s value. there is a component in Azure Ml called “**Tune Model Hyperparameters”,** that run the algorithms against different parameter’s value.

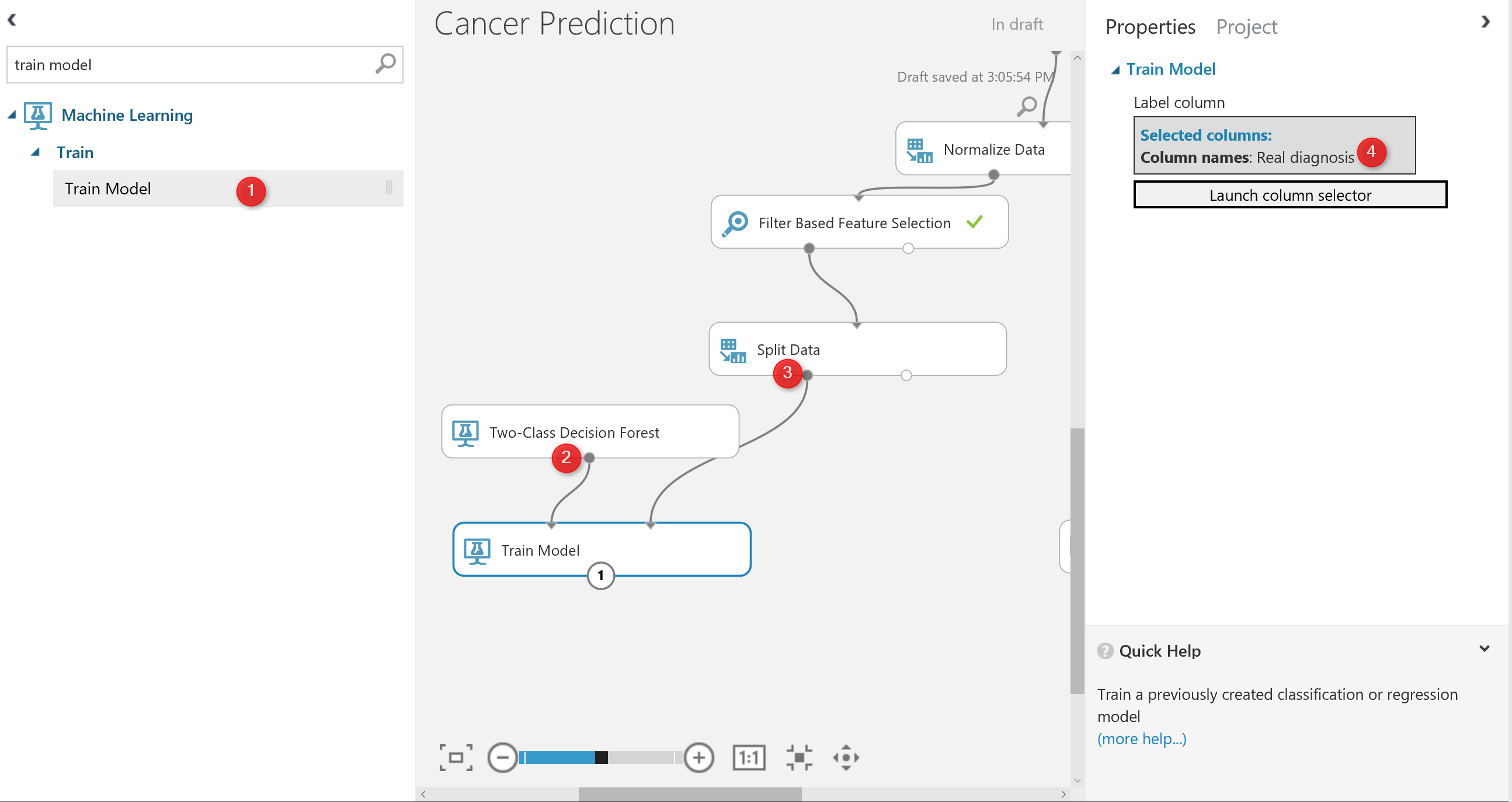
first I am going to try two algorithms on a single dataset to see which one better predict the patient’s diagnosis.

I select the “**Two-Class Decision Forest”** and “**Two-class Decision Jungle”** to see which one better predict.

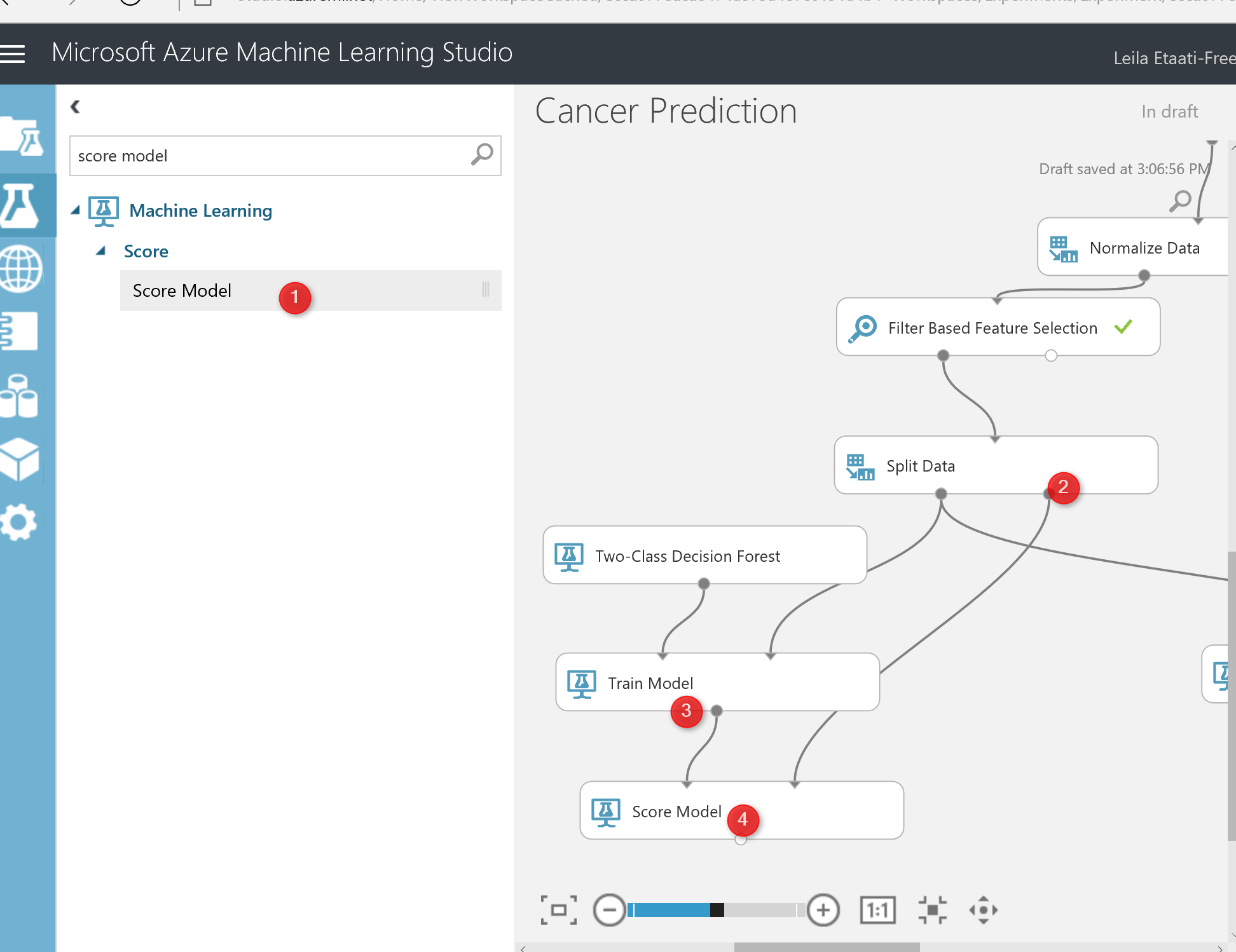
I just simply search for them in left side menu, and drag and drop them into experiment area.

[](http://radacad.com/wp-content/uploads/2017/06/23.png)

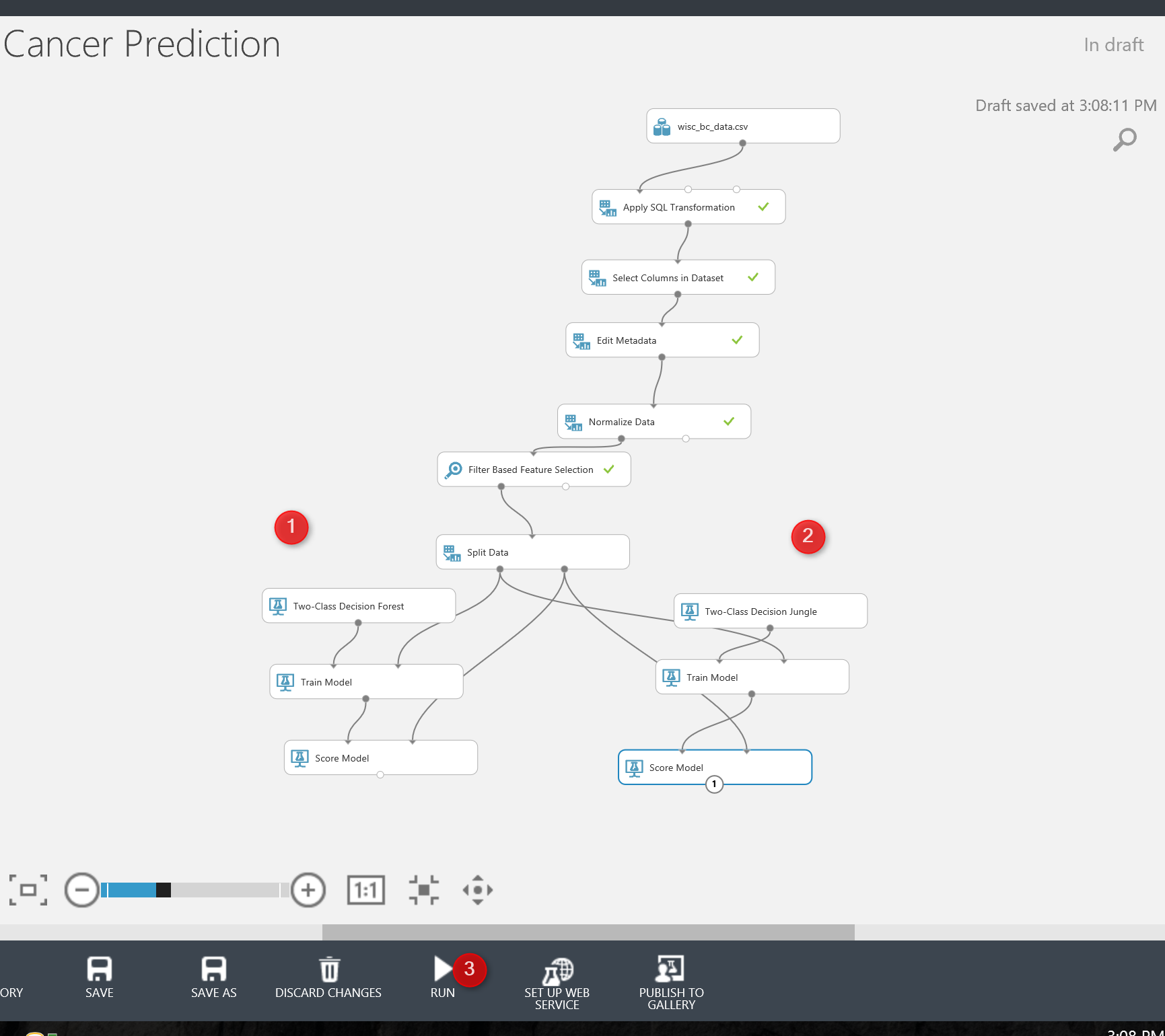
to work with thee models we need to train them first, there is a component name “**Train Model”** that help us to train a model. Train model component gets two input: one from algorithm and the other from dataset (in out example from split output node at the left side) see below picture. Moreover, just click on the train model components, and in the right side of the experiment in properties (number 4), we have to specify the columns that we want to predict. in this example is “Real Diagnosis”.

[](http://radacad.com/wp-content/uploads/2017/06/24.png)

Finally, we have to test the model. To test a model in Azure ML, there is component name “**Score Model**“. search for it and drag and drop it into experiment area. The score model gets first input from train model component, that is a machine learning algorithm, and the right side input from the split component the output for testing dataset.

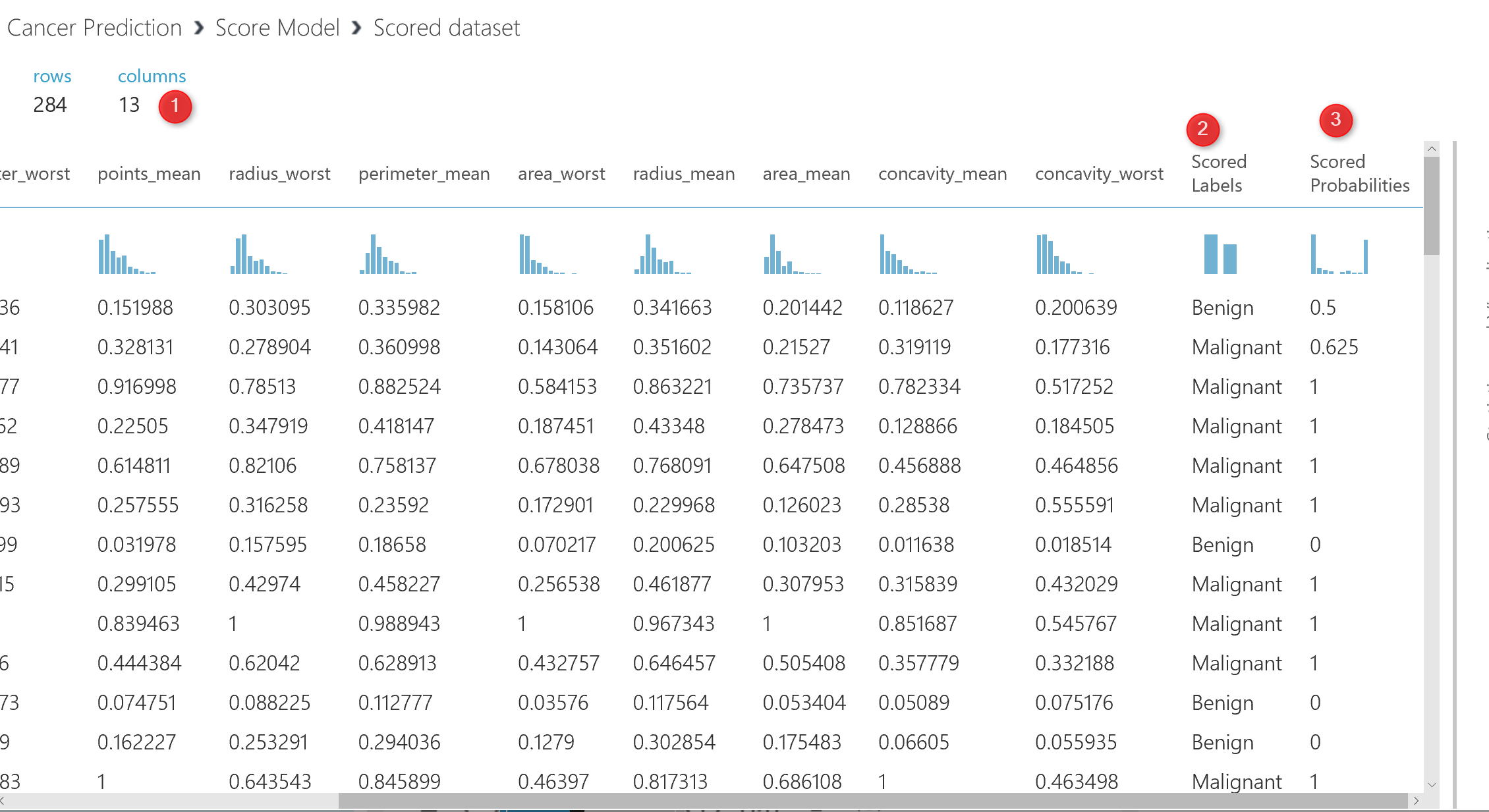
[](http://radacad.com/wp-content/uploads/2017/06/25.png)

I just created the same process for “**two-class decision jungle and forest**” (see below picture). Now I am going to run the experiment, to see the result.

[](http://radacad.com/wp-content/uploads/2017/06/26.png)

after the running the experiment, just click on the output node of one of the score model and visualize the dataset. in this dataset, you will see that we have 284 rows (that is the testing dataset), also you will see that we have 13 columns instead of 11. the 2 other columns are for the result of predictions as:**Score Label** and **Score Probabilities.**

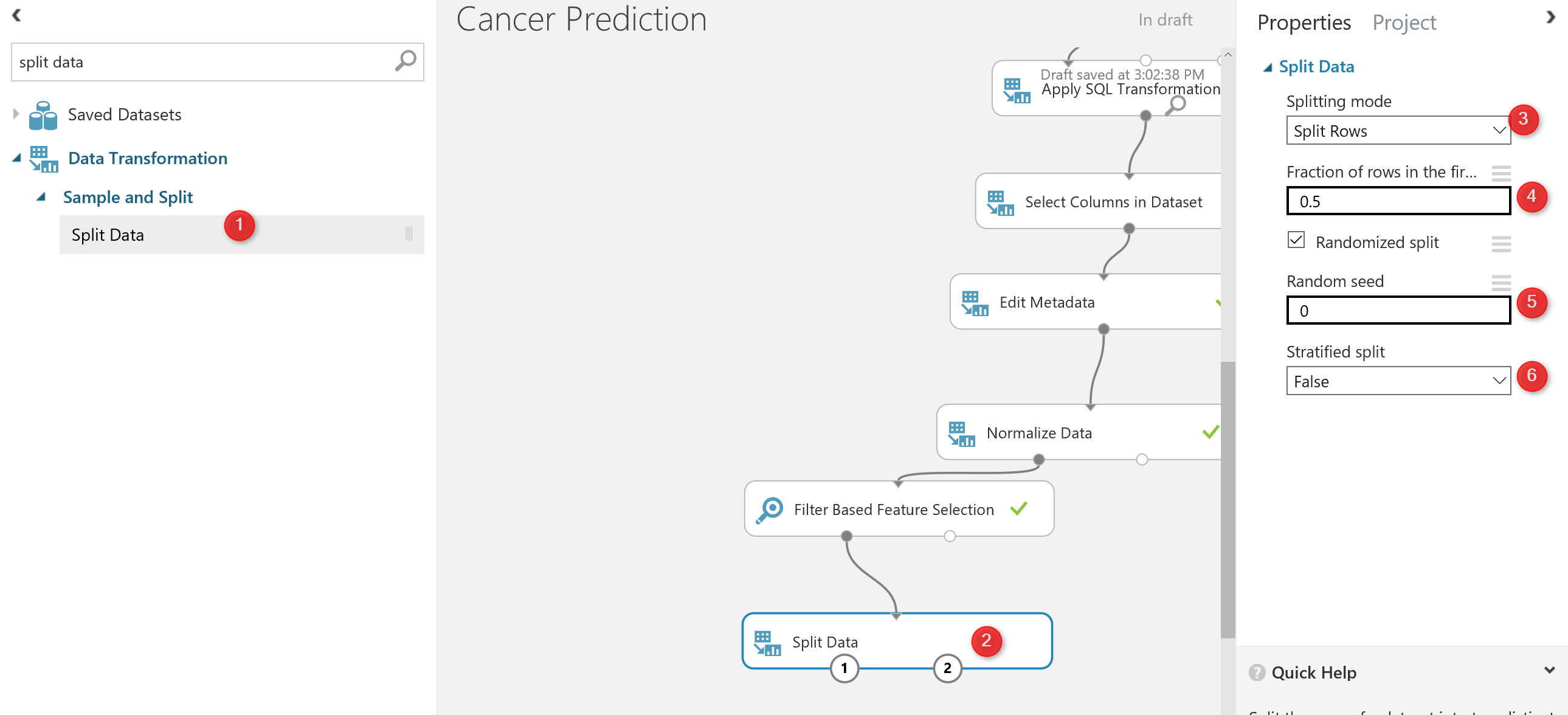
The Score Label: show the prediction of the diagnosis  for patients . The score Probabilities show the probability of predictions. For instance, for the first row  the prediction of the model was patient with these laboratory result will be **Benign** with 50% **probability**

[](http://radacad.com/wp-content/uploads/2017/06/27.png)

In the next post I will show How to **evaluate** the result of the prediction, how to try different algorithm **parameters** using tune hyper parameters, and also how to check different dataset using **Cross Validation**.

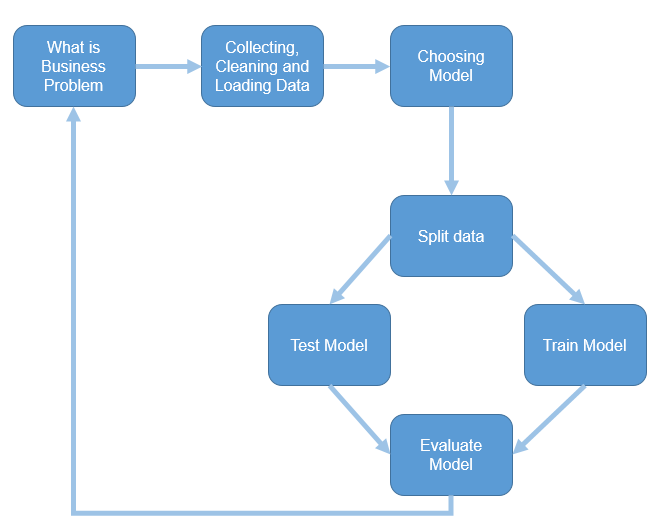
# Azure ML Part 5: A Machine Learning Prediction scenario (2)

Published Date : June 2, 2017

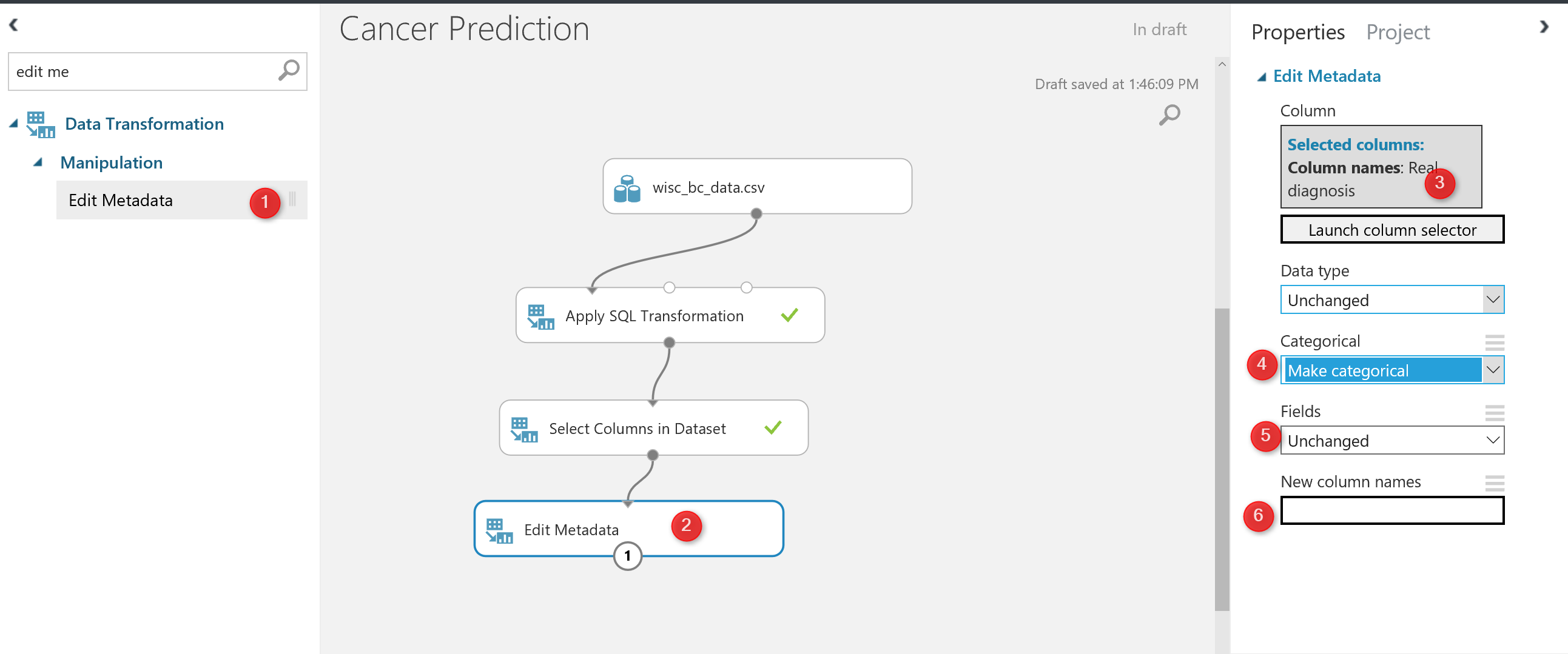
[](http://radacad.com/wp-content/uploads/2017/06/22.png)

In the [previous Post](http://radacad.com/azure-ml-part-4-a-machine-learning-prediction-scenario-1) , I start to do prediction the cancer diagnosis using some laboratory data. I have explained some of the main components for doing the data cleaning such as “**SQL Transformation”**, **“Edit Meta Data”, “Select Columns” and “Missing Values”.**

In this post I am going to show the rest of data cleaning process using Azure ML components and how to split data for training.

[](http://radacad.com/wp-content/uploads/2017/03/a4.png)

In the last post, we come up with the below process.

[](http://radacad.com/wp-content/uploads/2017/06/9.png)

Now we are going to do some more data cleaning  as “**Normalization of Data”.**

look at the out put of the data from “Edit metadata”:

[](http://radacad.com/wp-content/uploads/2017/06/13.png)

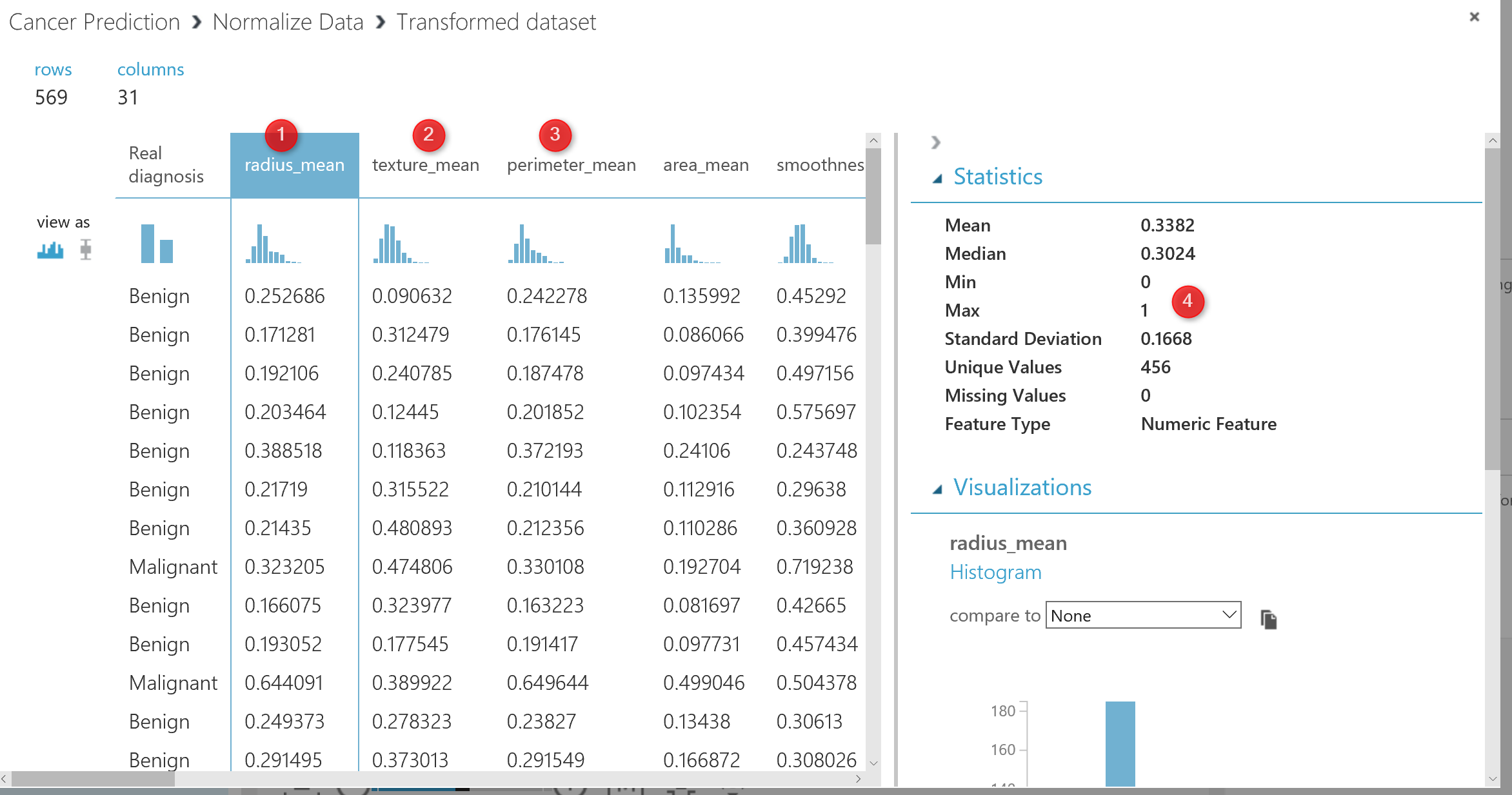
If you look at the data, you will see that each column has its own data range, for instance column 2 (number 1 in above picture ) has data range between 9 to 20, whilst the column number 5 (number 2 in picture) has values between 100 to 500, the same for the column 6 (number 3), the data range is between 0.01 to 0.1. So the data is not in the same range. To do the machine learning, it is important that all data be in the same range. I am going to bring data in range of 0 to 1 using “**Min Max**” algorithm. There is a component in Azure Ml name “**Normalize Data**“.

As you can see in the below picture, Normalize data component exits under “**Data Transformation**” Component.

[](http://radacad.com/wp-content/uploads/2017/06/14.png)

I just drag and drop the component to the experiment area, in the right side of the experiment, we able to specify the normalization method (number 4 in above picture), for this experiment I have choose the “**Min Max**” method. Also, we able to select which column we want to normalize (number 5 in picture).

After running the experiment we will have below data set that is totally normalized in comparison with the previous one.

[](http://radacad.com/wp-content/uploads/2017/06/16.png)

So, now we have enough data cleaning and data wrangling in our dataset. The next step is about the “Choosing the right data for prediction” that we call it **” Feature Selection”.**

**Feature selection**

is the process of finding which attribute ha more impact on the prediction columns. In our example, we are going to see which laboratory measure has more impact on the Diagnosis result.

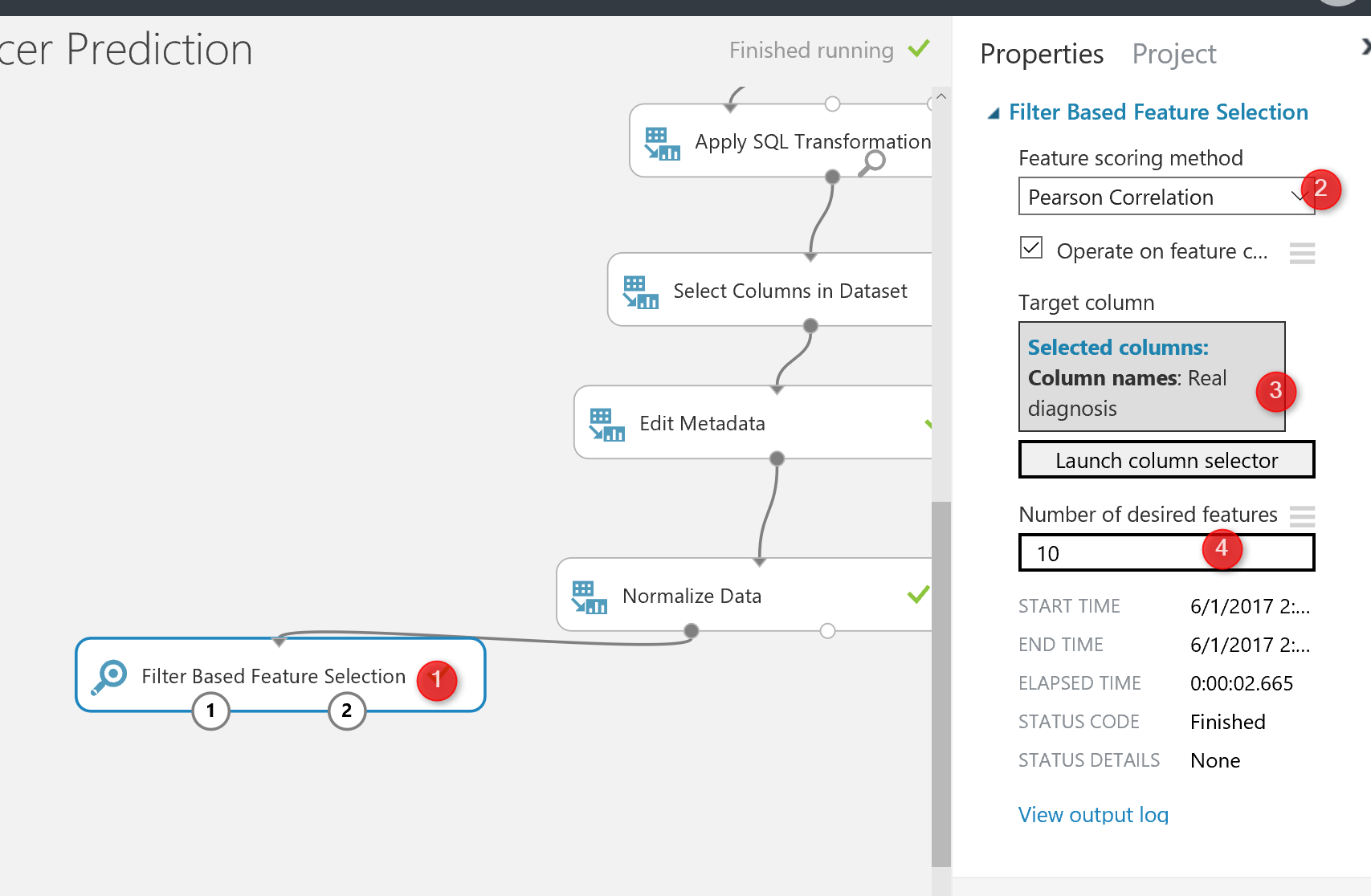
There are many approach in machine learning to do that using algorithms like “**regression, decision tree”, correlation analysis** will help.

in Azure Ml there is a component name **“Filter Based Feature Selection”.**

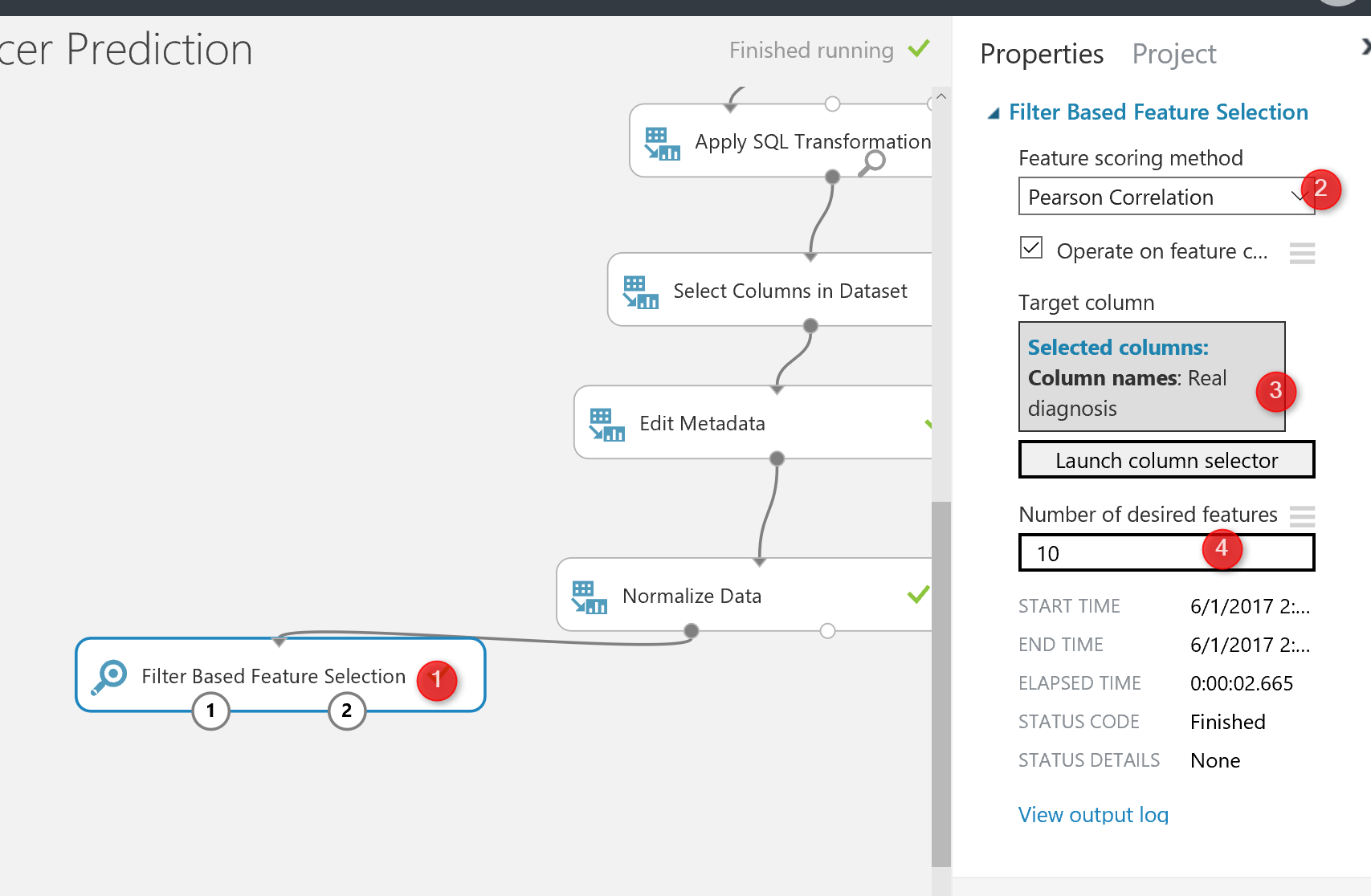
in the below picture, I have shown how I use it to find which attributes has more impact on the diagnosis of the cancer condition.

As you can see in the below picture, I have connected the output of the normalization component to the input node of the “**Filter Based Feature Selection**“. by clicking on this node, in the right side, you will see some options that you have to set them up first.

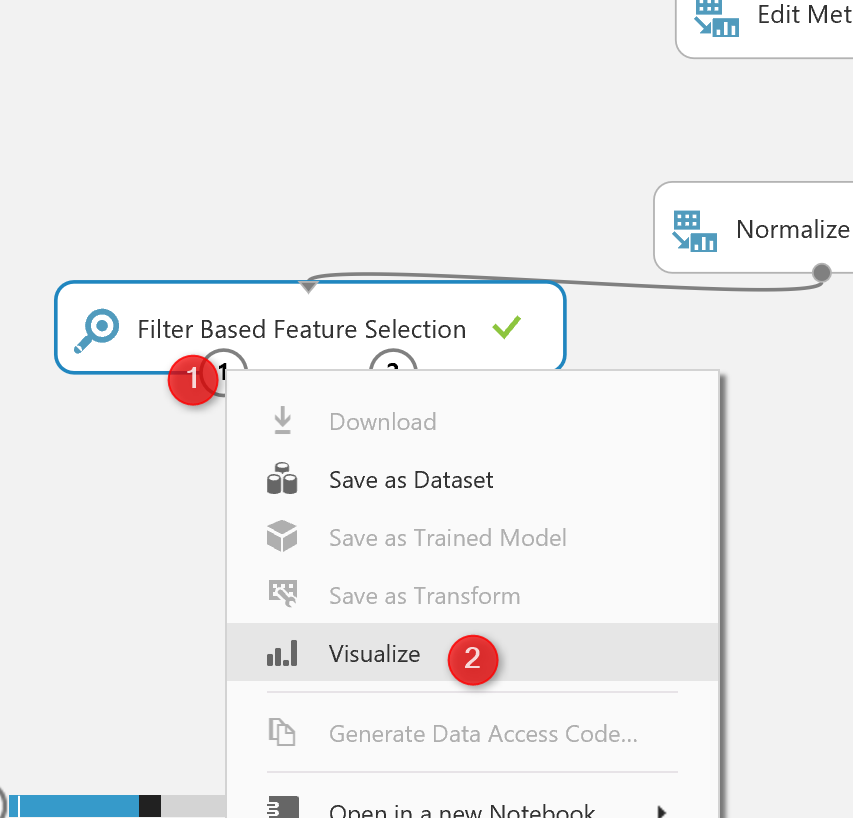
First of all, you should choose the algorithm for the aim of feature selection. In this experiment, I have choose the “**Pearson Correlation**” analysis (Number 3) . However, there are many other approaches that I will talk about them later. Then, in the next textbox (number 4), I have identified the columns that I want to predict, which in our example is “**Diagnosis column**“.

[](http://radacad.com/wp-content/uploads/2017/06/18.png)

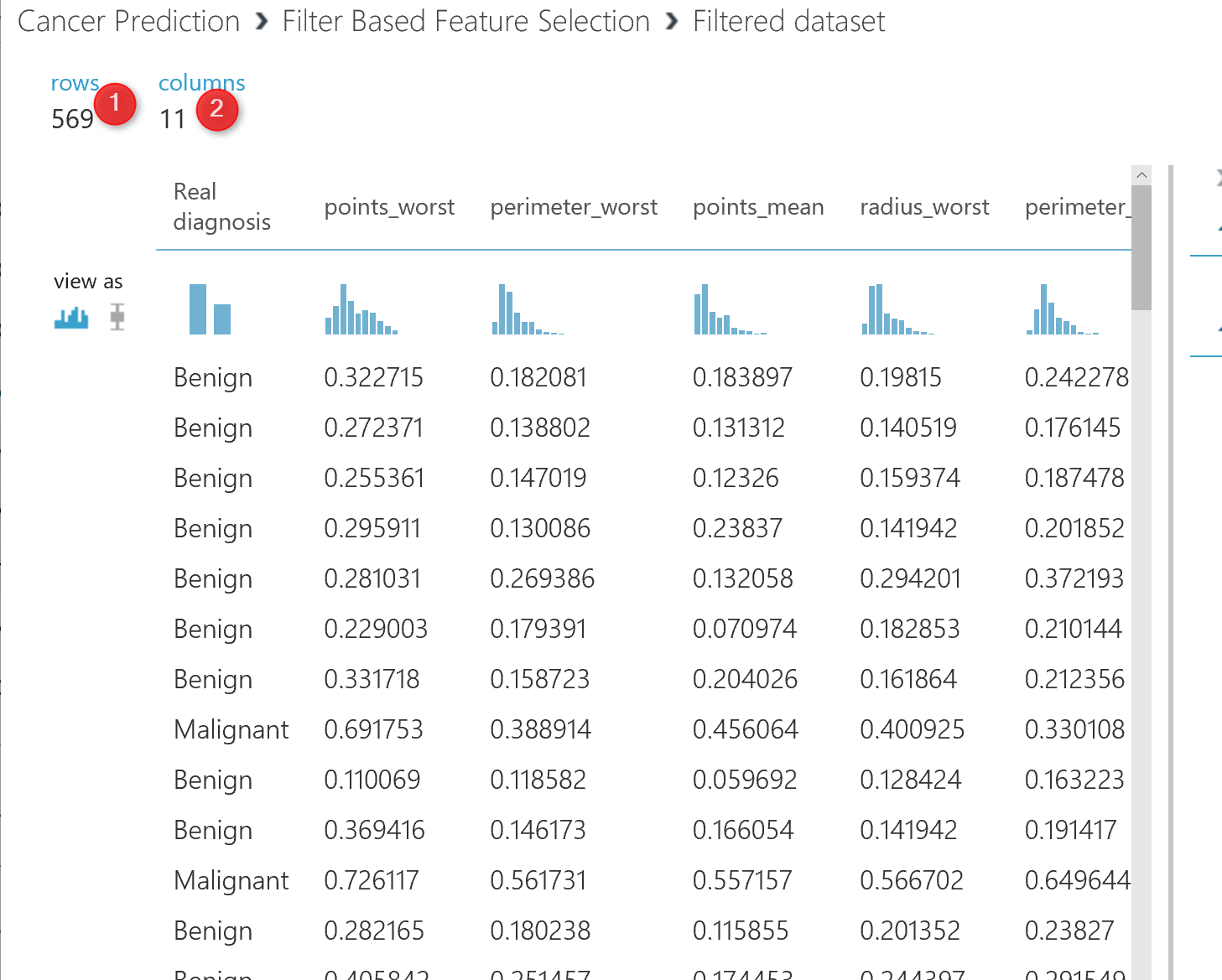
Finally, in the last textbox, I specify the number of features that I am interested to have for prediction among 32 columns, the default value is 1, but I specify it as 10 (see below picture).

[](http://radacad.com/wp-content/uploads/2017/06/18.png)

Then, I run the experiment to see the result of the feature selection by right click on the left output of the node (see number 1 in below picture)

[](http://radacad.com/wp-content/uploads/2017/06/19.png)

The below result will be shown, as you can see, now we have 11 columns instead of the 32 that means these are columns that have more impact on the predicting of cancer diagnosis.

[](http://radacad.com/wp-content/uploads/2017/06/20.png)

If you right click on the right side output node of  the “feature selection” and visualize the dataset, you will see below data:

[](http://radacad.com/wp-content/uploads/2017/06/21.png)

This data shows which factor has impact more on the “real diagnosis” column, for instance, column “Point\_Worst” has 79% impact on the diagnosis, or “Perimeter\_worst” has 78% impact. all of these  analysis has been done by correlation analysis to see the impact of each attributes on predictable column.

We done by data cleaning and feature selection. Now we clean our data, we identify which factor has more impact on the “caner Diagnosis”.

The next step according the machine learning process is so spit data for training and testing purpose

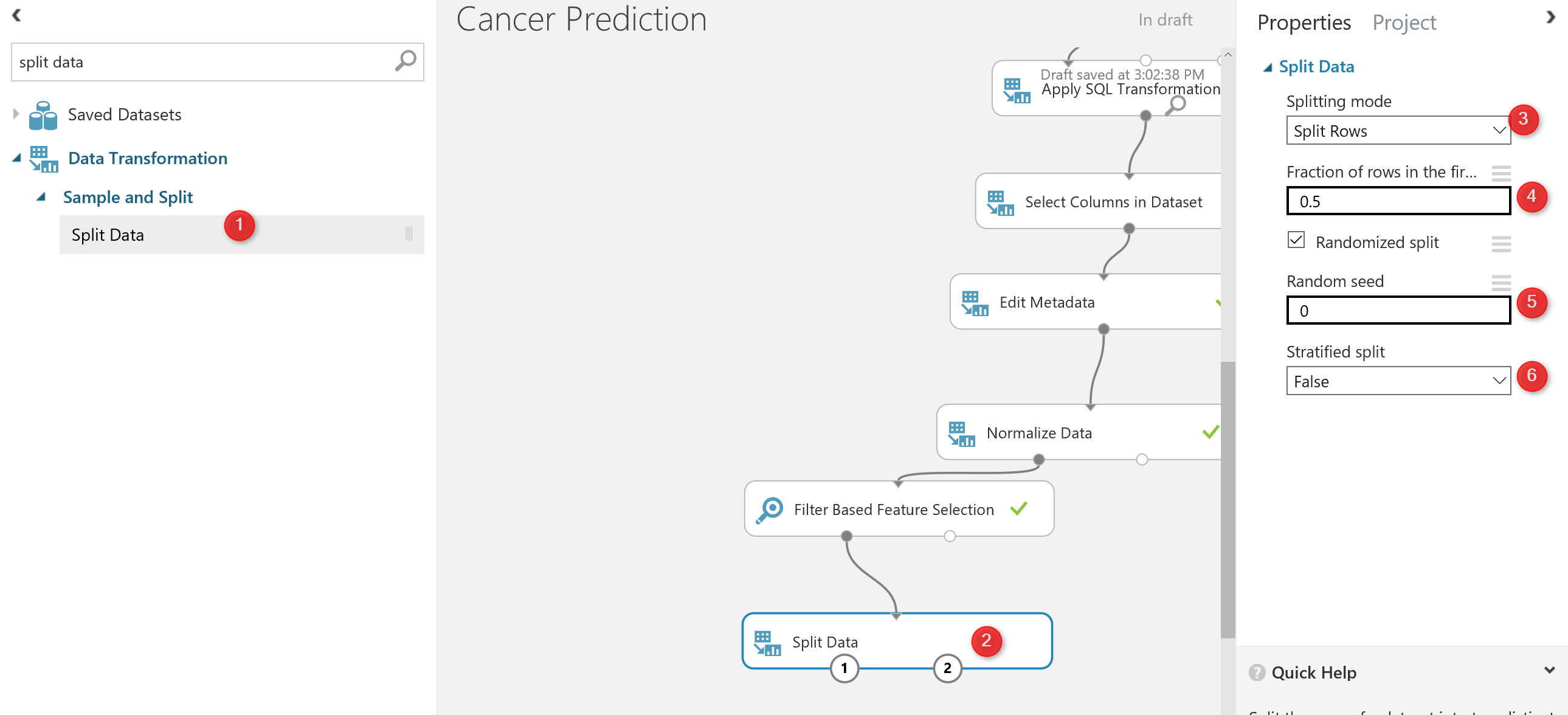
**Split Data**

The main aim of the machine learning is to learn from past data, so we have to provide a set of data to train the model. The training dataset, helps an algorithm to better understand the data behaviour, so able to learn from past data and predict the future data.

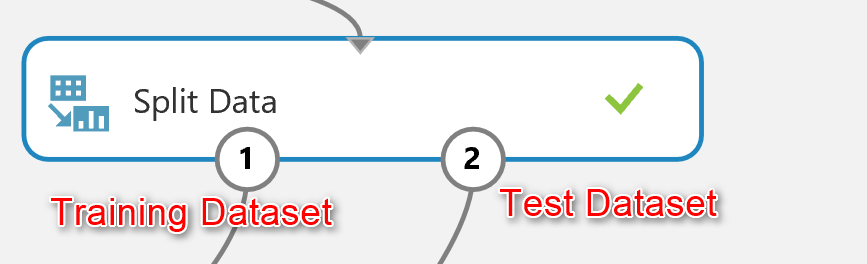
Also, after creating the model, we should test it to see whether they predict well or not, so we have to provide a Test dataset from what we already have to check the results.

there is a component in Azure ml that help us to create a test and train dataset name “**Split Data**“. split data can be found under the “**Data Transformation**” component. (number 1). Just drag and drop it to the experiment and connect it to the output of “**feature selection**” ( the dataset output that is in left side).

Then in the right side of the experiment, you will see a windows that show the parameter list.  the first parameter identifies how to split the dataset, which I choose the “Split Rows” , there are other approach like using regular expression for dividing the dataset into train and test, which hopefully I will talk about them later. Next in the number 4, you see that I specify that 0.5 % of data should go for testing and 0.5% for training. always this percentage should be above 70% that provides more data for training. in number 5 and 6 we can set a value for seed to make the experiment consistence for each run.

[](http://radacad.com/wp-content/uploads/2017/06/22.png)

Now by running the code , we have two datasets: Training dataset which located in the left side of the “Split Data”. The test data has been located in the right side of the “split Node”

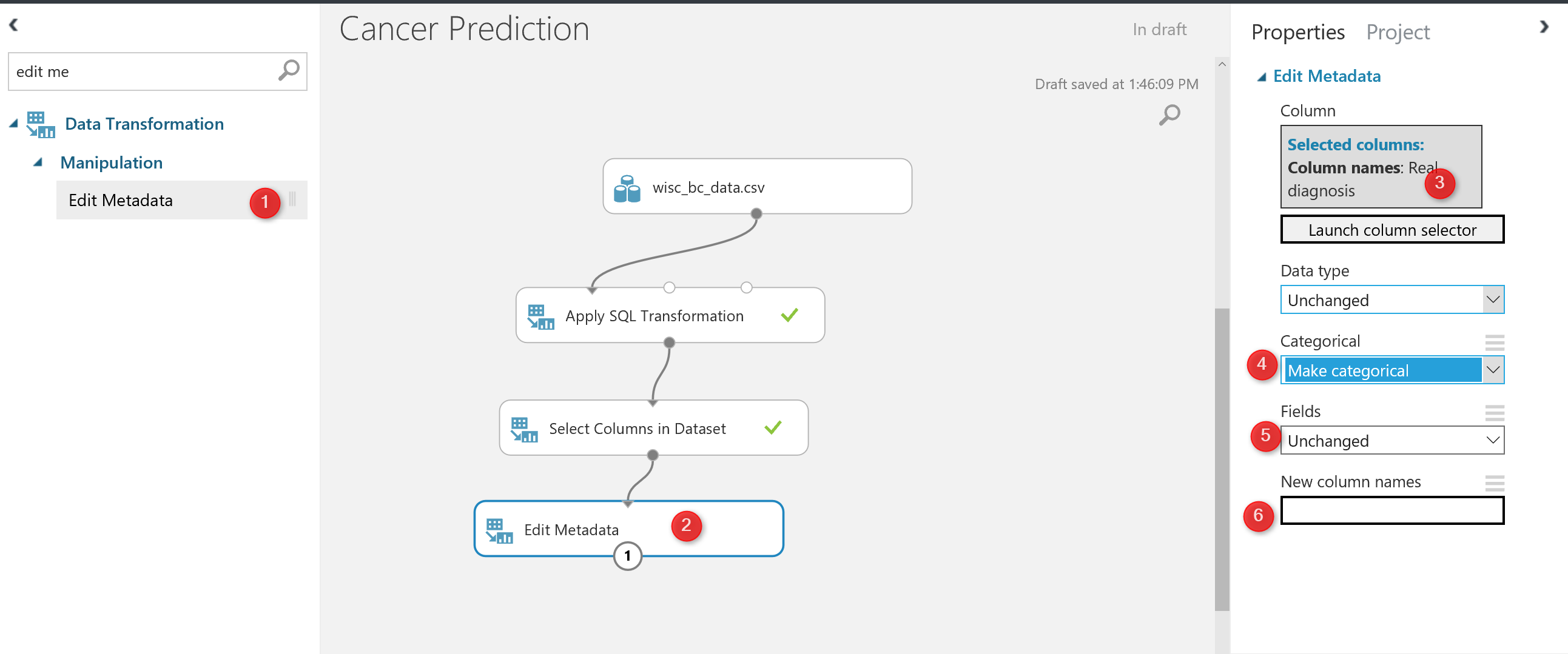
[](http://radacad.com/wp-content/uploads/2017/06/55.png)

Now we have “Training dataset” and “Testing Dataset”.

In the next post I will show how to choose algorithms and also how to train, test, and evaluate the model.

# Azure ML Part 4: A Machine Learning Prediction scenario (1)

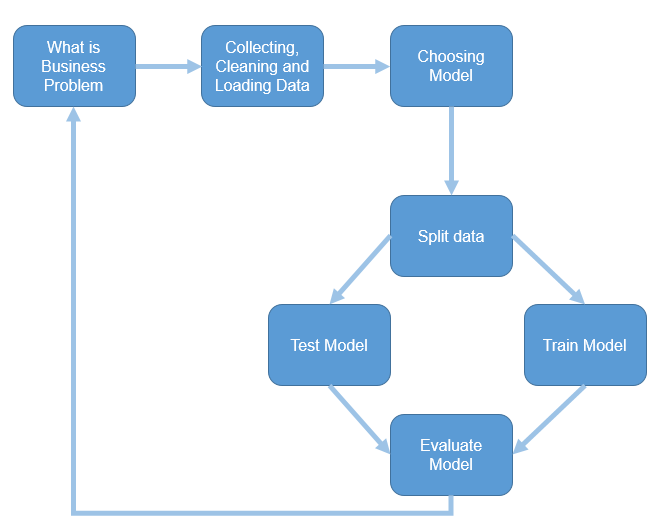
Published Date : June 1, 2017

[](http://radacad.com/wp-content/uploads/2017/06/9.png)

In previous Posts [Part 1](http://radacad.com/azure-machine-learning-part-1-introduction" \o "Part 1), [Part 2](http://radacad.com/azure-machine-learning-part-2-azure-ml-environment) and [Part3](http://radacad.com/azure-machine-learning-part-3-data-transformation-in-r" \o "Part 3) I have explain some about the azure Ml environment, how to import data into it and finally how to do data transformation using Azure ML component.

In this post and the next one I am going to show how to do a Machine Learning in Azure ML using different components via a scenario.

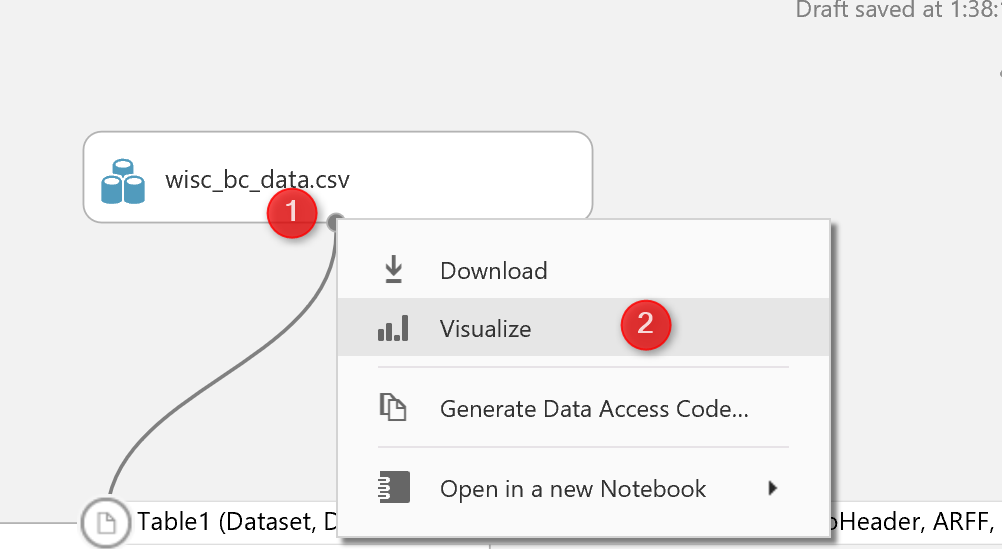
First based on the Machine Learning Process (see below image), the first step is to identify the business problems

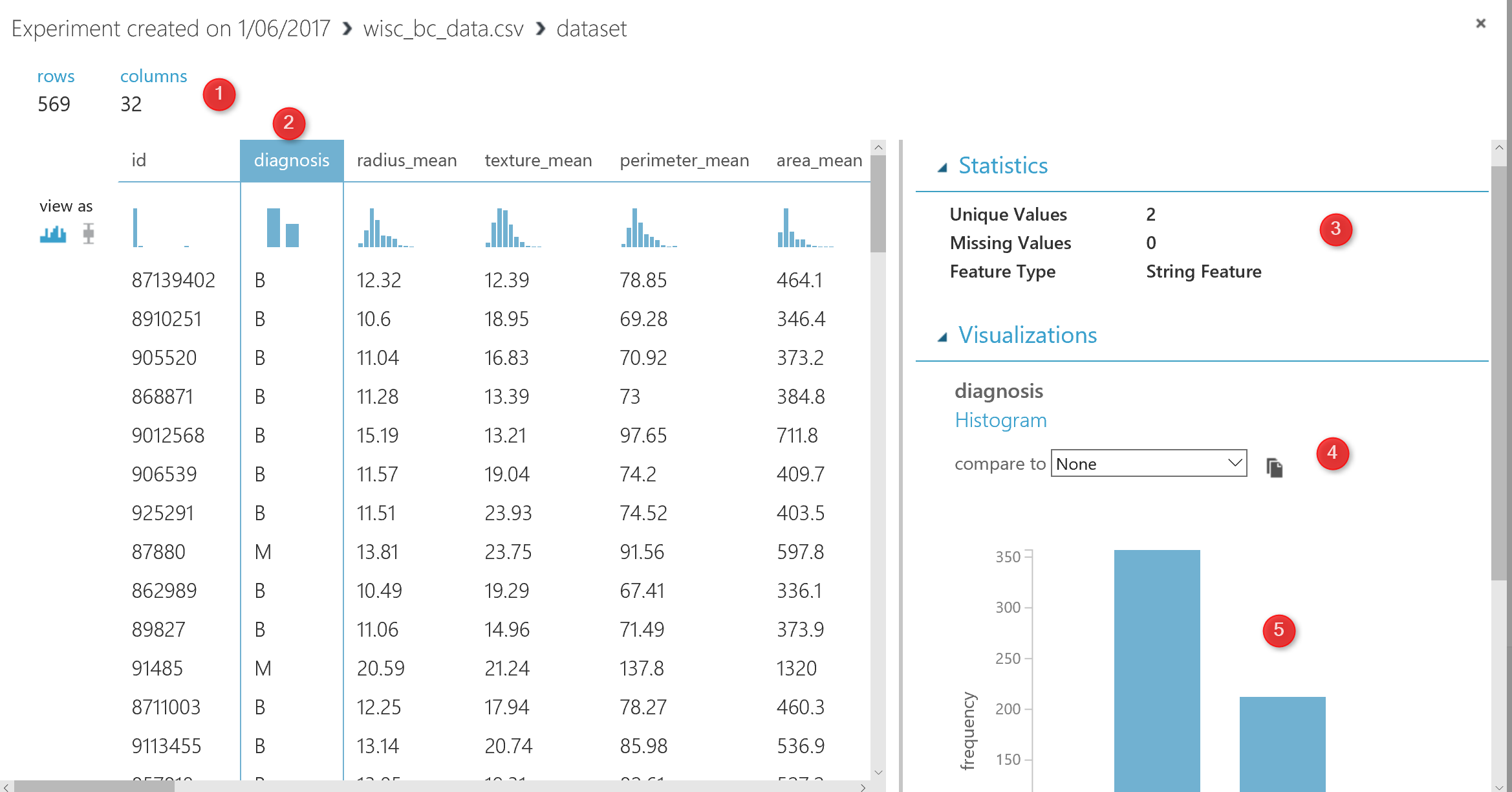
[](http://radacad.com/wp-content/uploads/2017/03/a4.png)

**What is Business Problem ?**

I am going to predict whether a customer will be **Benign or Malignant.**I have a dataset that shows the laboratory results of cancer cells, also the patients ID , and the final doctor’s  diagnosis as “B” and “M”.

I have imported the data as a “CSV” file, the CSV file has **569 records** and **32 columns** (see below picture). In the right side of the picture, you will see a menu  that shows the statistics about the data (number 3). Moreover, in the number 4 and 5,  you able to see the chart that shows the histogram of the diagnosis column data.

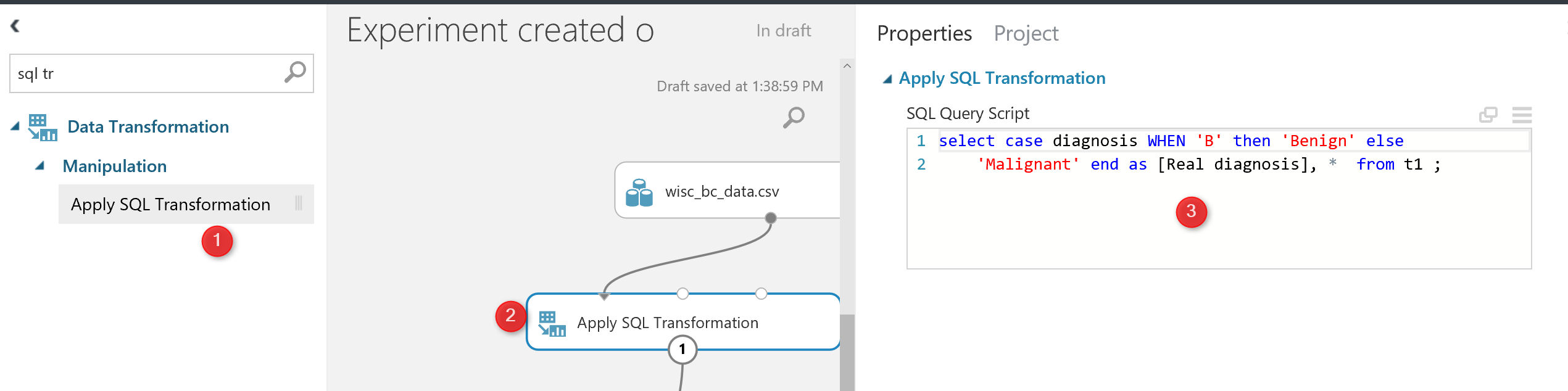
[](http://radacad.com/wp-content/uploads/2017/06/2.png)

[](http://radacad.com/wp-content/uploads/2017/06/1.png)

I am going to change the data description of the diagnosis. I want to replace “**B**” with “**Benign**” and “**M**” with “**Malignant**“. This can be done using “**SQL Transformation**“. SQL Transformation component can be access under the “**Data Transformation**” component. Just drag and drop it to the experiment area, then connect it to the dataset as number 2 in the below picture. To write SQL code just click on the component, then write a normal SQL Statement to transfer data as below

“select case diagnosis WHEN ‘B’ then ‘Benign’ else  
‘Malignant’ end as [Real diagnosis], \* from t1 ;”

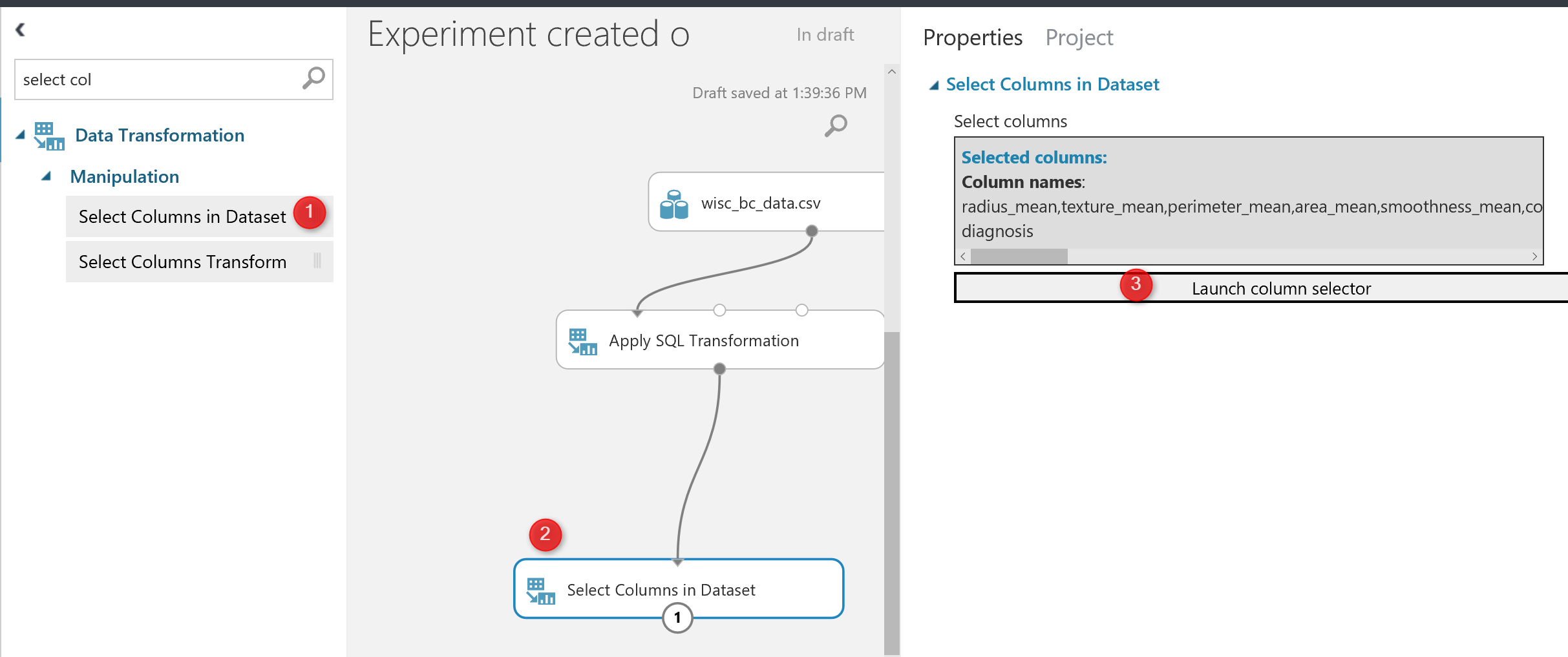
The code will replace the value of column “diagnosis”. as you can see in the above code, the data has been selected from “t1″ table. So what is “T1″ table , As you can see in the below picture (number 2). SQL transformation component has three main input nodes. The first node is called “t1″ as we connect the dataset to it. so if you connect you dataset to the second node then in SQL scripts you should select your data from “t2″

[](http://radacad.com/wp-content/uploads/2017/06/3.png)

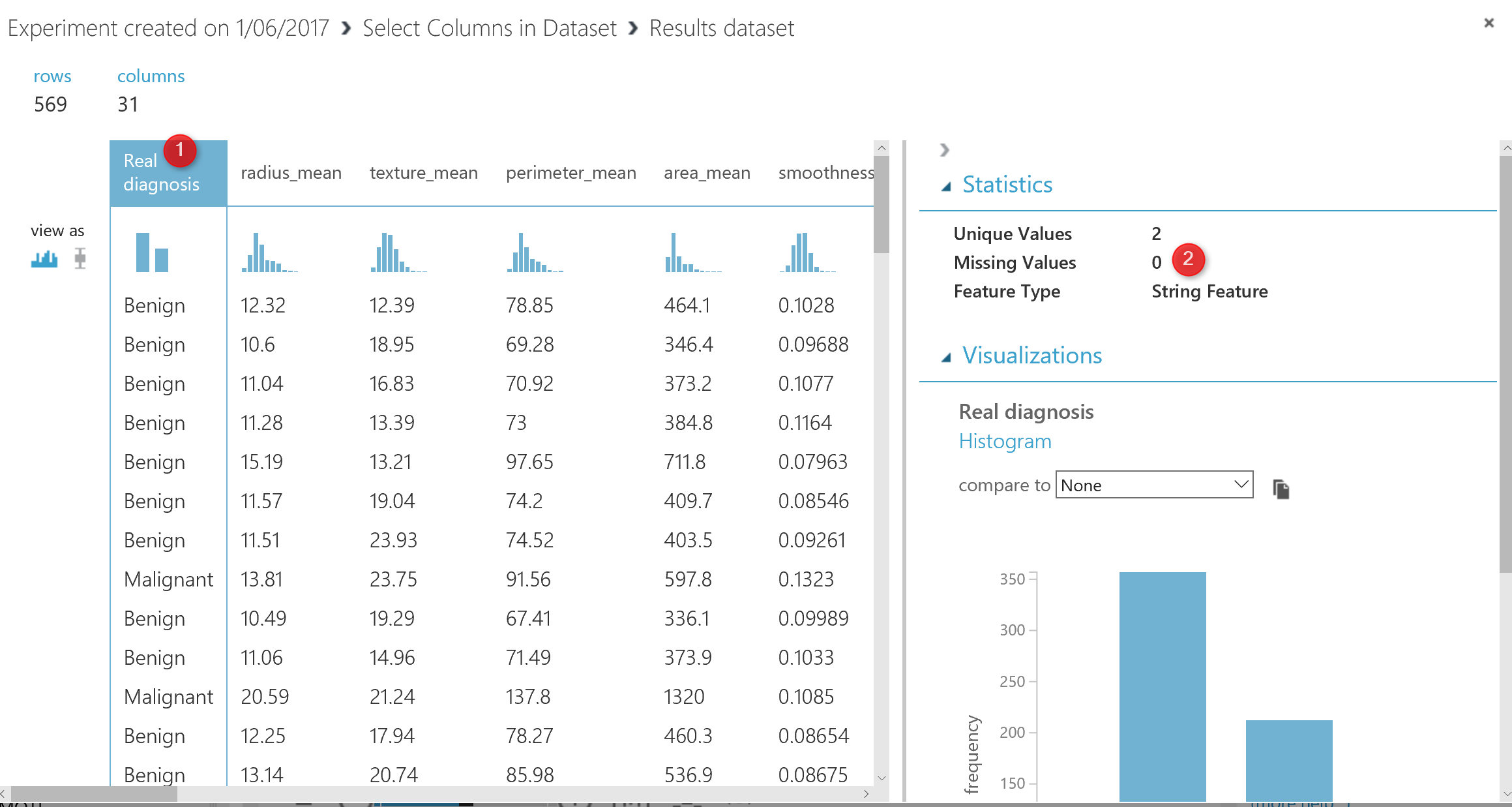
The next step for **Data cleaning** is to remove the columns that we think does not have impact on the prediction.

**Select Column Data**

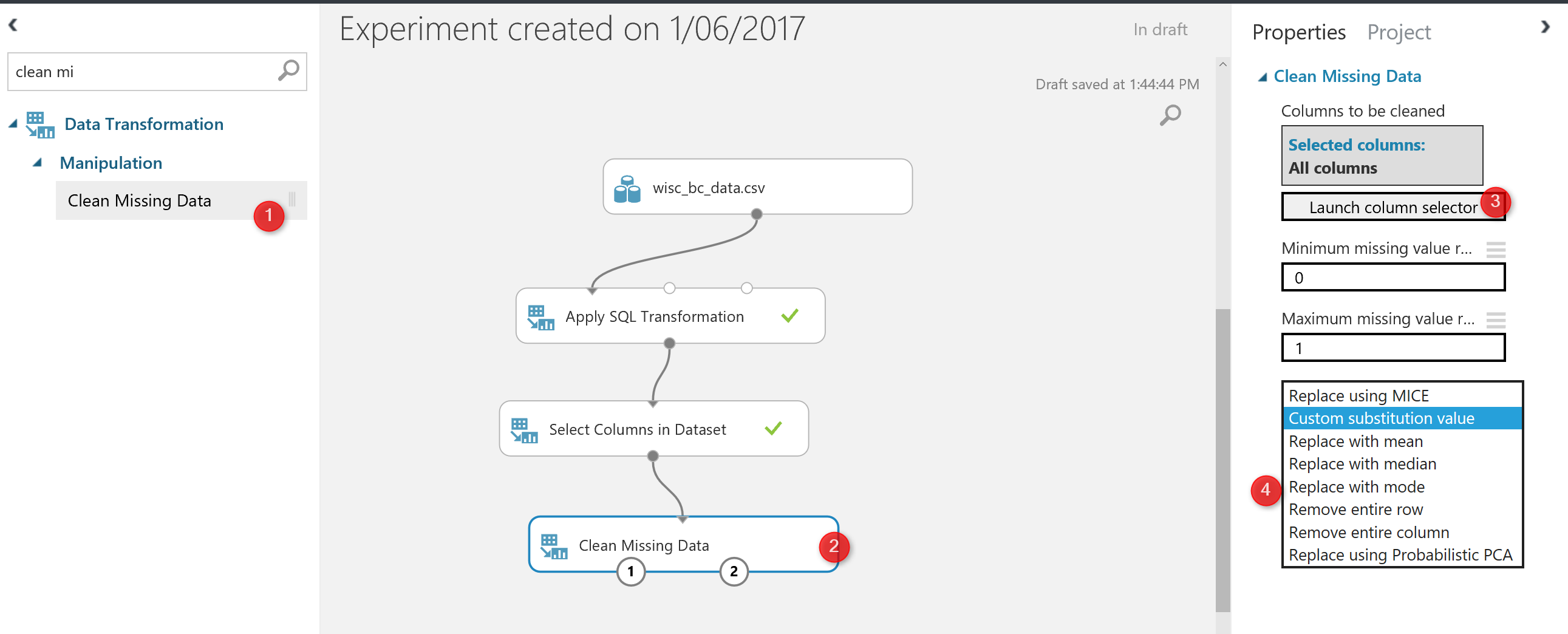
I am going to remove the “**Patient ID**” from dataset using “**Select Columns in Dataset**” to select the columns that I need.

[](http://radacad.com/wp-content/uploads/2017/06/4.png)

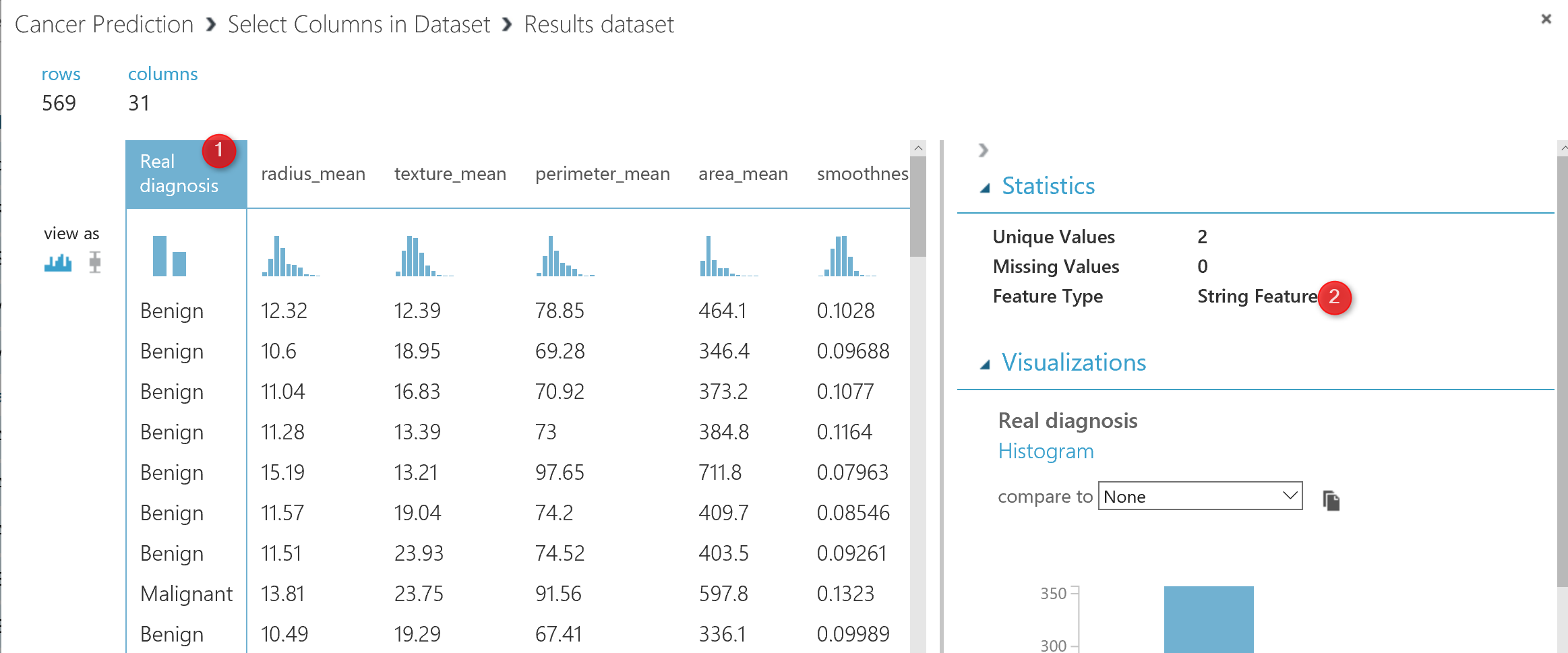
after run the code, we will see the below results, so the “ID” columns has been r3emoved and we have now 31 columns instead.

[](http://radacad.com/wp-content/uploads/2017/06/7.png)

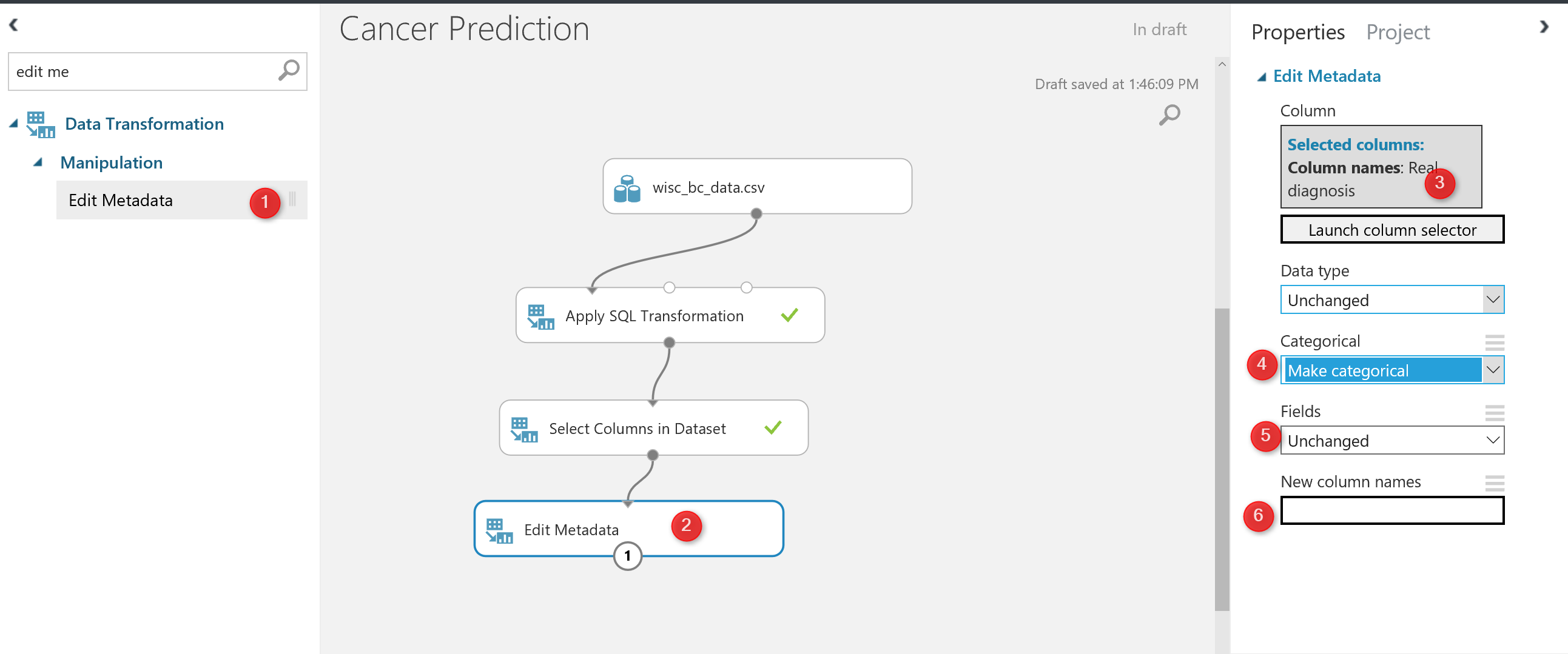
There are some possibility to have “**Missing Values**“. Missing value may impacts on the pr3ediction results, so always recommend to remove them. There is a component in Azure ML called “**Clean Missing Values**”  as shown below. This component able to remove data that does not have at value. (see below picture).

[](http://radacad.com/wp-content/uploads/2017/06/8.png)

The next step for data cleaning is to change the data structure, the diagnosis data column was “**string** ” type ( see the below picture ), before adding the edit meta data component, the diagnosis column was “String Feature”.

[](http://radacad.com/wp-content/uploads/2017/06/11.png)

for doing the prediction, we are going to use “**Two-Class**” algorithms that need a “**categorical** ” data, hence we need to change the diagnosis data type.to change the data type in Azure Ml we have a component that helps us name “**Edit Metadata”.**

[](http://radacad.com/wp-content/uploads/2017/06/9.png)

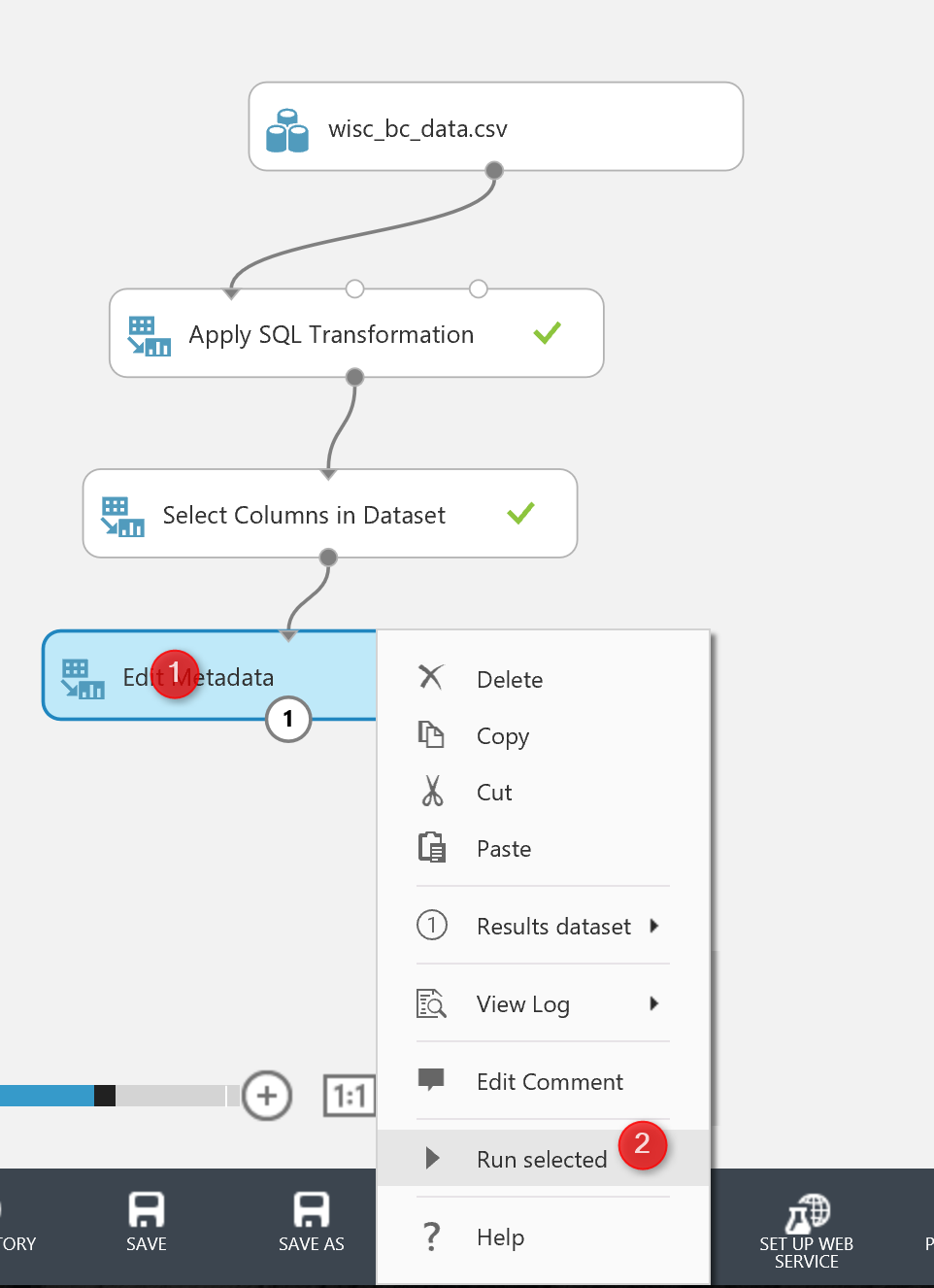
as you can see in above picture, I have drag and drop this component into the experiment area, then in number 3, I specify the column that I need to change the data structure. Following, in number 4, I change the column to **categorical data**.

you also able to change the name of the column in Edit meta data component as shown in number 6.

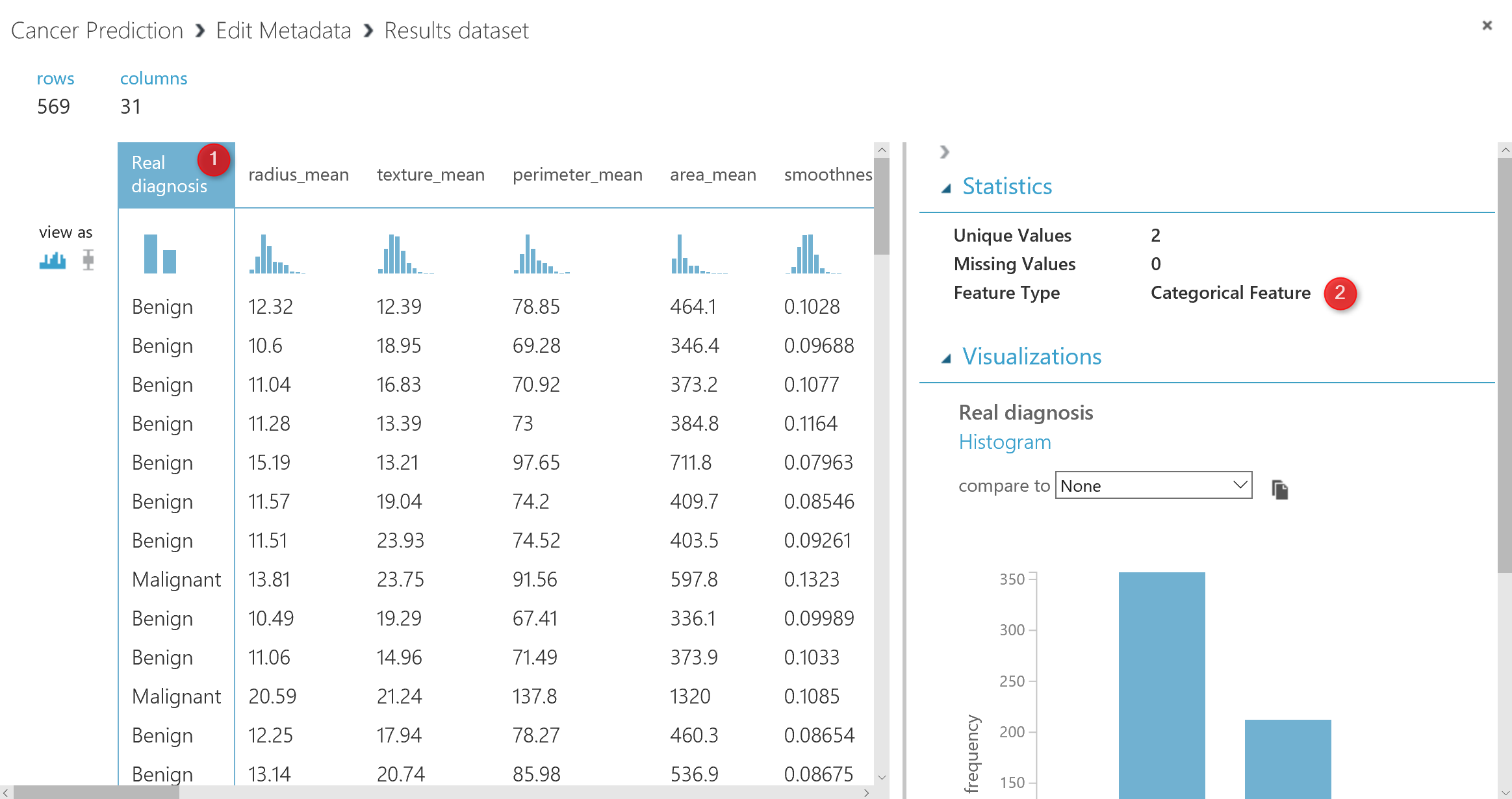
Now I am going to run the code. However I just want to see the final result of adding “Edit Metadata”.

There is a possibility to run just one node at the time in Azure Ml to make experiment creation faster.

For instance, after adding the “**Edit Metadata**” I just right click on the node and I have selected the ” Run selected node”.

[](http://radacad.com/wp-content/uploads/2017/06/10.png)

Now If I visualized the data at the node” **Edit Meta Data**” you will see that the diagnosis column has a  ” categorical  Feature” data type which good for doing the prediction.

[](http://radacad.com/wp-content/uploads/2017/06/12.png)

there are still some steps for doing the data cleaning and feature election, that I will explain in the next posts.

to sum up, in this post, I have explain how to work with “**SQL Transformation”** Component**, “Select specific columns”** for remove a attributes. How to remove missing value using **“Clean Missing Data”** component. Also edit meta data to change the name and type of columns.

I next post I will explain how to **normalize** these data, how to find which attributes has more impacts on the “diagnosis ” columns prediction, also how to **split data** , **train model, score morel, evaluate model**.