AI BASED DIABETES PREDICTION SYSTEM

TEAM MEMBER

211521106077: KARTHIKEYAN B

PANIMALAR INISTUTE OF TECHNOLGY

**PHASE 2 SUBMISSION DOCUMENT**

**INTRODUCITON**

* AI-based diabetes prediction systems can be used to identify people who are at risk of developing diabetes long before they experience any symptoms. This allows for early intervention and prevention of the disease.
* AI-based systems can be used to create personalized risk assessments and treatment plans for each patient. This ensures that patients receive the care that they need to prevent or manage diabetes.
* AI-based systems can automate many of the tasks involved in diabetes screening and diagnosis. This can help to improve the efficiency of healthcare systems and make diabetes care more accessible to everyone.

Benefits of using AI-based diabetes prediction systems:

* Early detection of diabetes
* Personalized care
* Improved efficiency

**CONTENT FOR PROJECT PHASE 2:**

 Exploring innovative techniques such as ensemble methods and deep learning architectures to improve the prediction system's accuracy and robustness.

**DATA SOURCE:**

**DATASET LINK:(** **https://www.kaggle.com/datasets/mathchi/diabetes-data-set)**

**CONTENT FOR INNOVATION:**

Collect a large dataset of medical records, including information about patients' age, weight, height, blood pressure, cholesterol levels, and other health factors.

This is the first and most important step in developing any AI-based system. The dataset should be as large and diverse as possible, in order to ensure that the system can learn to predict diabetes risk accurately for a wide range of patients.

Split the dataset into training and testing sets.

The training set is used to train the machine learning models, while the testing set is used to evaluate their performance on unseen data. This is important to ensure that the models are not overfitting to the training data, and that they will be able to generalize to new patients.

Train a variety of machine learning models, such as random forests, logistic regression, support vector machines, CNNs, and RNNs, on the training set.

Each of these machine learning algorithms has its own strengths and weaknesses. By training a variety of different algorithms, you can increase the chances of finding one that performs well on your dataset.

Evaluate the performance of the individual models on the testing set.

This will give you an idea of how well each model generalizes to unseen data. You can use a variety of metrics to evaluate the models, such as accuracy, precision, recall, and F1 score.

Select a subset of the best performing models.

Once you have evaluated the individual models, you can select a subset of the best performing models. These models will be used to train the ensemble model.

Train an ensemble model, such as a stacking model or a bagging model, on the predictions of the individual models.

Ensemble models combine the predictions of multiple machine learning models to improve the overall prediction accuracy. This is done by averaging the predictions of the individual models, or by using a more sophisticated voting scheme.

Evaluate the performance of the ensemble model on the testing set.

Once you have trained the ensemble model, you need to evaluate its performance on the testing set. This will give you a final idea of how well the model generalizes to unseen data.

Deploy the ensemble model to production.

Once you are satisfied with the performance of the ensemble model, you can deploy it to production. This means making it available to users so that they can use it to predict their own risk of developing diabetes.

**DECISION TREE:**

A decision tree is a type of supervised machine learning algorithm that can be used for classification and regression tasks. It works by creating a tree-like structure, where each node in the tree represents a decision. The leaves of the tree represent the final predictions.

To train a decision tree model for diabetes prediction, we would first need to collect a dataset of medical records, including information about patients' age, weight, height, blood pressure, cholesterol levels, and other health factors. We would then need to split the dataset into training and testing sets.

The training set would be used to train the decision tree model. The model would learn to make decisions about whether a patient has diabetes based on the features in the training data.

Once the model is trained, we would evaluate its performance on the testing set. This would give us an idea of how well the model generalizes to unseen data.

If the model performs well on the testing set, we can then deploy it to production. This means making it available to users so that they can use it to predict their own risk of developing diabetes.

Here is an example of how a decision tree model might predict whether a patient has diabetes:

IF age > 50 AND weight > 100kg

THEN diabetes = True

ELSE

IF blood pressure > 140/90

THEN diabetes = True

ELSE

IF cholesterol > 200mg/dL

THEN diabetes = True

ELSE

diabetes = False

This decision tree model would first check the patient's age. If the patient is over 50 years old, and they weigh more than 100kg, then the model would predict that they have diabetes. If the patient is not over 50 years old, or they do not weigh more than 100kg, then the model would check the patient's blood pressure. If the patient's blood pressure is over 140/90, then the model would predict that they have diabetes. If the patient's blood pressure is not over 140/90, then the model would check the patient's cholesterol. If the patient's cholesterol is over 200mg/dL, then the model would predict that they have diabetes. Otherwise, the model would predict that the patient does not have diabetes.

Decision tree models are a relatively simple type of machine learning algorithm, but they can be very effective for predicting diabetes risk. They are also relatively easy to interpret, which makes them a good choice for applications where explainability is important.

Benefits of using decision trees for diabetes prediction:

* Decision trees are relatively simple to interpret, which makes them a good choice for applications where explainability is important.
* Decision trees can be trained on a variety of data types, including numerical, categorical, and ordinal data.
* Decision trees are relatively fast to train and evaluate.
* Decision trees are robust to overfitting.

Limitations of using decision trees for diabetes prediction:

* Decision trees can be biased, depending on the data that they are trained on.
* Decision trees can be oversensitive to small changes in the data.
* Decision trees can be computationally expensive to train on large datasets.

Overall, decision trees are a powerful tool for diabetes prediction. They are relatively simple to use and interpret, and they can be trained on a variety of data types. However, it is important to be aware of the potential limitations of decision trees, such as bias and computational complexity.

**LOGISTIC REGRESSION:**

Logistic regression is a supervised machine learning algorithm that can be used for classification tasks. It works by predicting the probability of a binary outcome, such as whether a patient has diabetes or not.

To train a logistic regression model for diabetes prediction, we would first need to collect a dataset of medical records, including information about patients' age, weight, height, blood pressure, cholesterol levels, and other health factors. We would then need to label each patient as either having diabetes or not having diabetes.

Once we have labeled the data, we can train the logistic regression model. The model will learn to predict the probability of a patient having diabetes based on the features in the data.

Once the model is trained, we can evaluate its performance on a held-out test set. This will give us an idea of how well the model generalizes to unseen data.

If the model performs well on the test set, we can then deploy it to production. This means making it available to users so that they can use it to predict their own risk of developing diabetes.

Here is an example of how a logistic regression model might predict the probability of a patient having diabetes:

Logit(p) = b0 + b1 \* age + b2 \* weight + b3 \* height + b4 \* blood\_pressure + b5 \* cholesterol + ...

p = 1 / (1 + exp(-logit(p)))

The coefficients b0, b1, b2, ... are learned by the model during the training process. The logit function is a sigmoid function that converts the linear combination of features into a probability.

The logistic regression model would first calculate the logit of the patient's probability of having diabetes. The logit is a linear combination of the patient's features, weighted by the learned coefficients. The model would then convert the logit to a probability using the sigmoid function.

If the probability is greater than a certain threshold (typically 0.5), then the model would predict that the patient has diabetes. Otherwise, the model would predict that the patient does not have diabetes.

Logistic regression models are a powerful tool for diabetes prediction. They are relatively simple to use and interpret, and they can be trained on a variety of data types. However, it is important to be aware of the potential limitations of logistic regression models, such as overfitting and bias.

Benefits of using logistic regression for diabetes prediction:

* Logistic regression is a relatively simple algorithm to understand and implement.
* Logistic regression models are robust to overfitting.
* Logistic regression models can be trained on a variety of data types, including numerical, categorical, and ordinal data.

Limitations of using logistic regression for diabetes prediction:

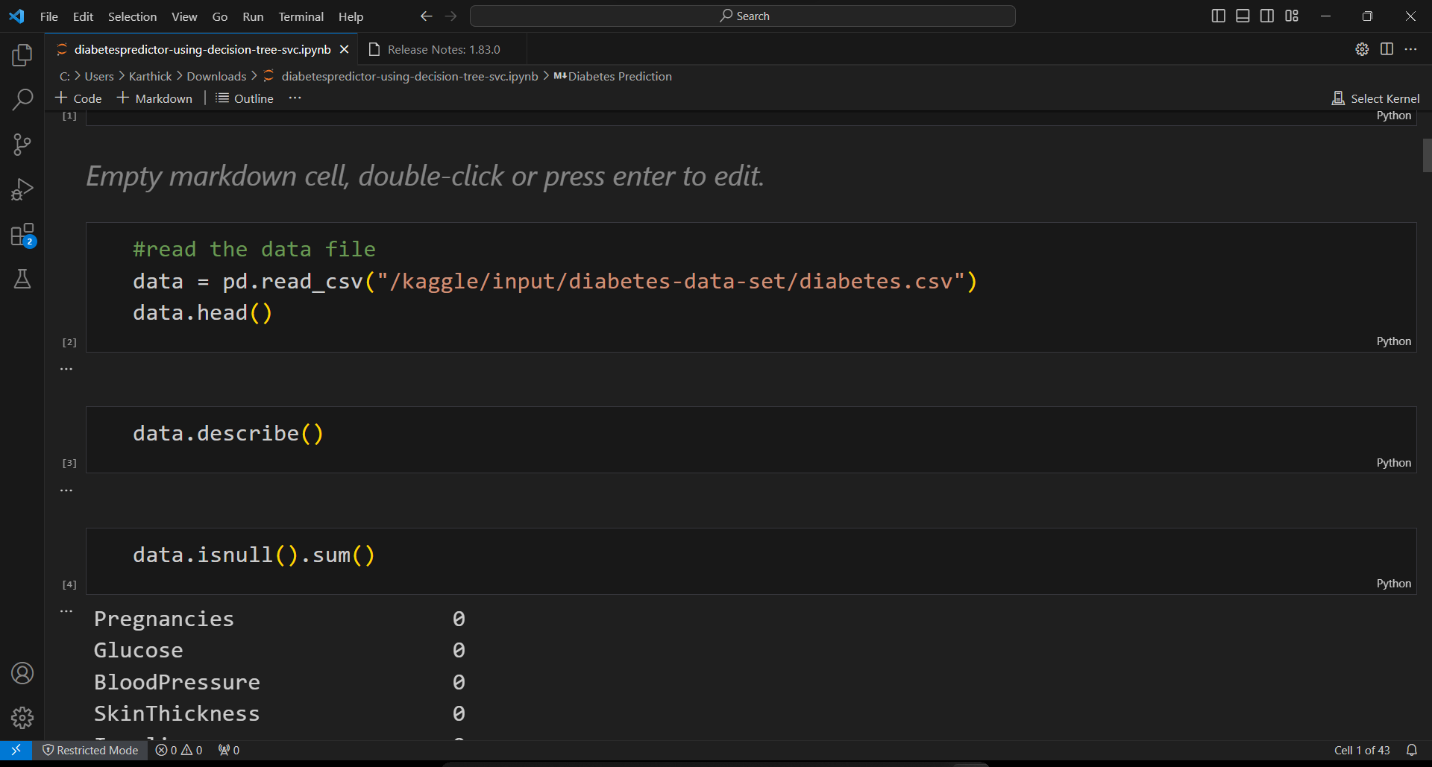
* Logistic regression models can be biased, depending on the data that they are trained on.
* Logistic regression models may not be able to capture complex relationships between the features and the outcome.
* Logistic regression models can be computationally expensive to train on large datasets.

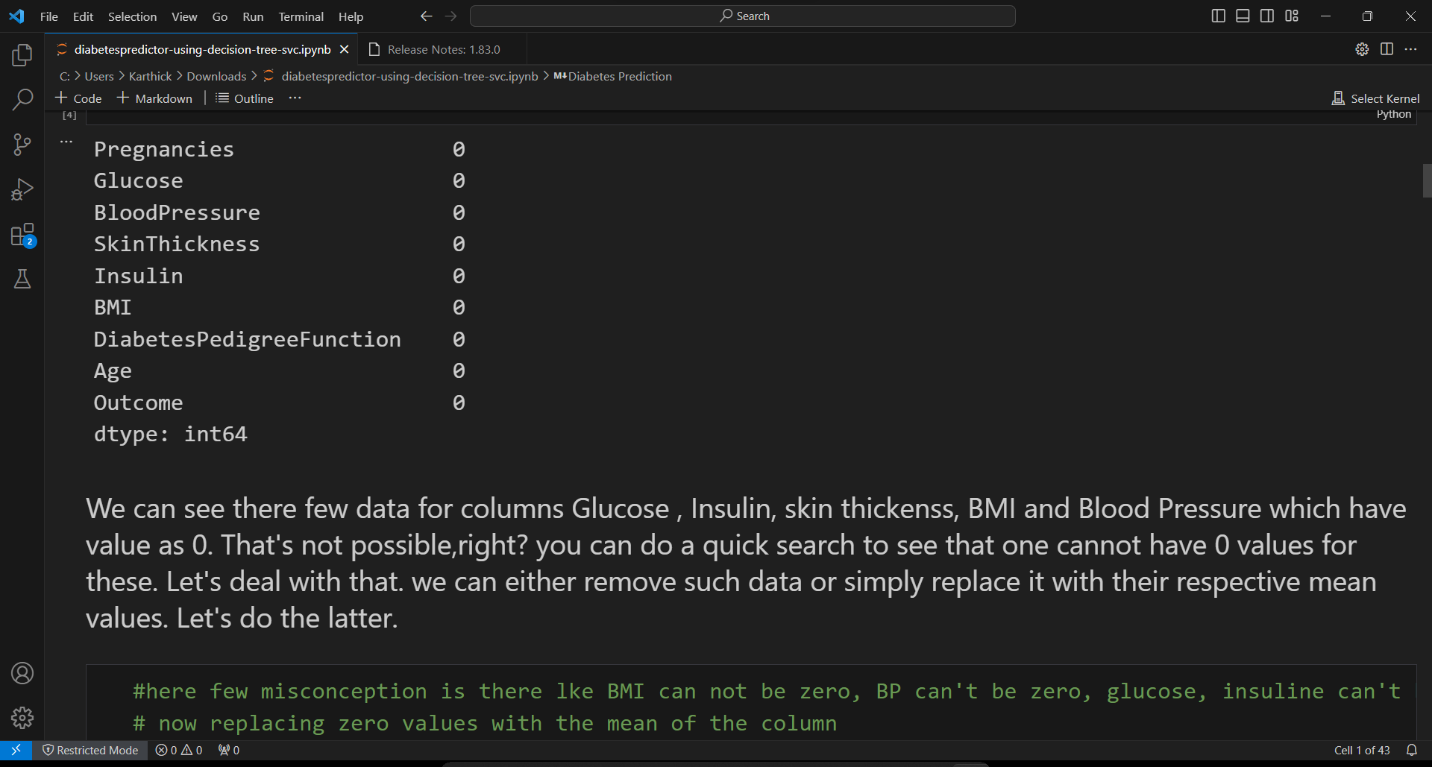
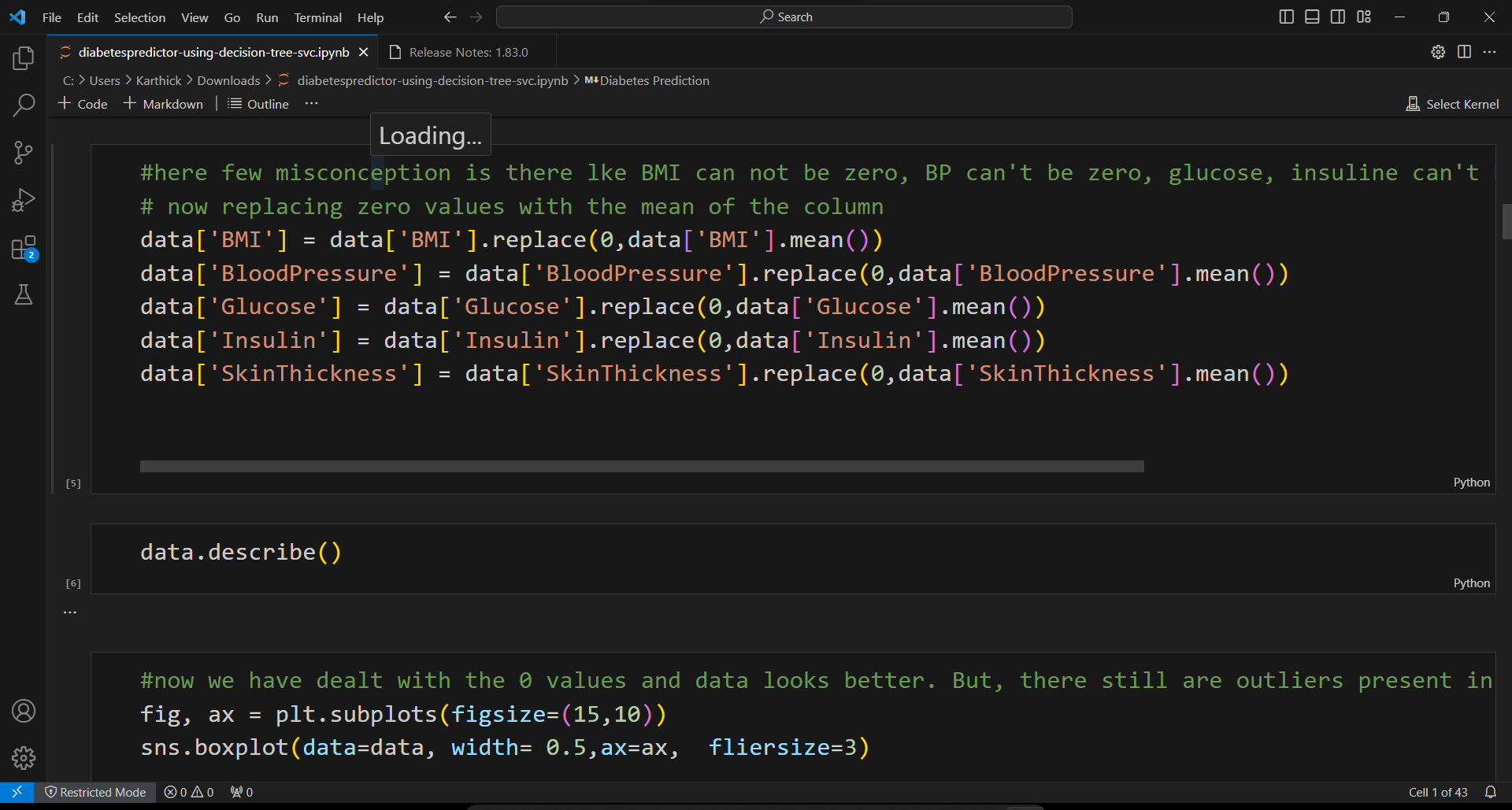
Overall, logistic regression is a powerful and versatile tool for diabetes prediction. It is a good choice for applications where simplicity, interpretability, and robustness are important.

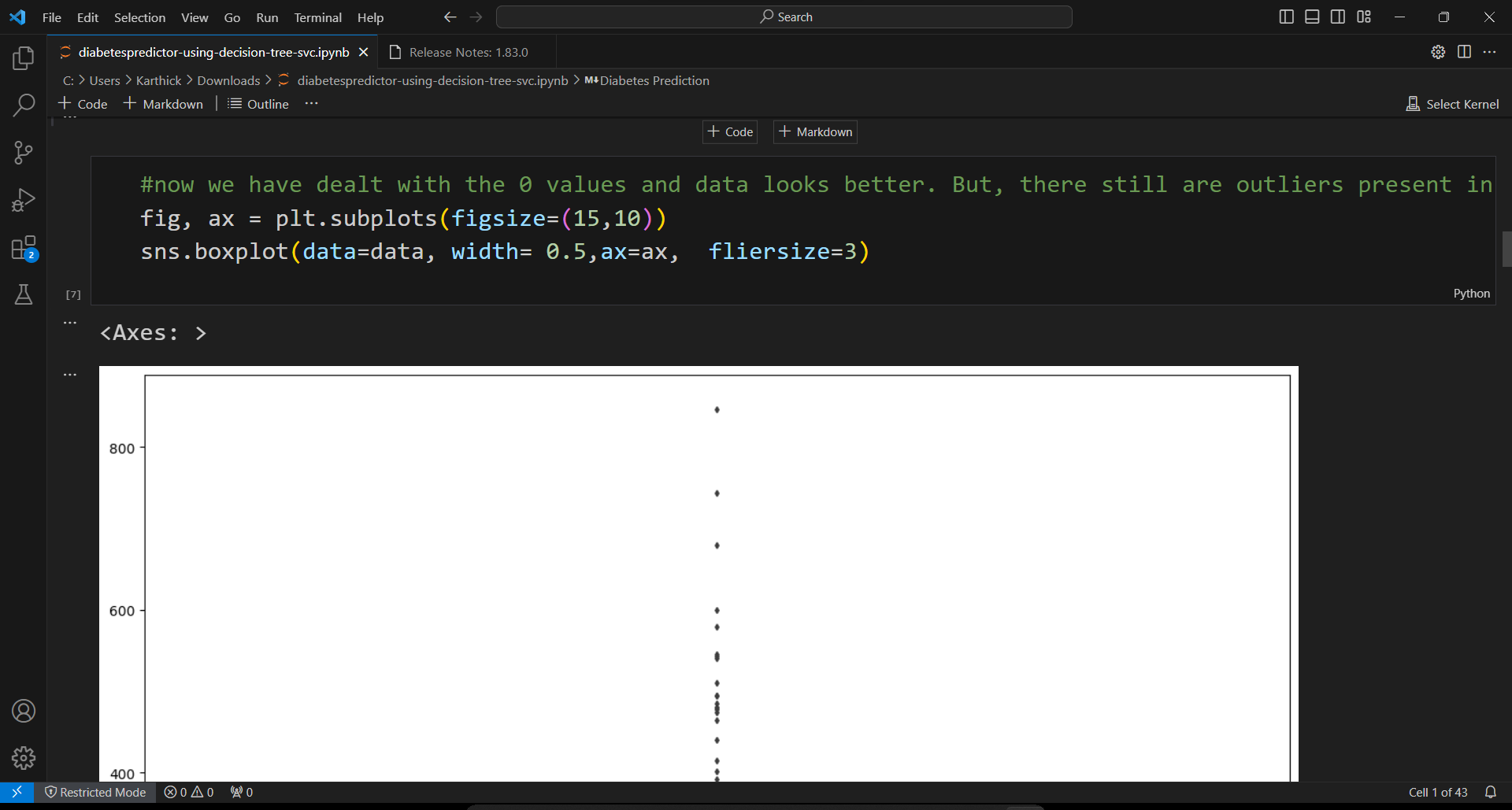
**PROGRAM:**

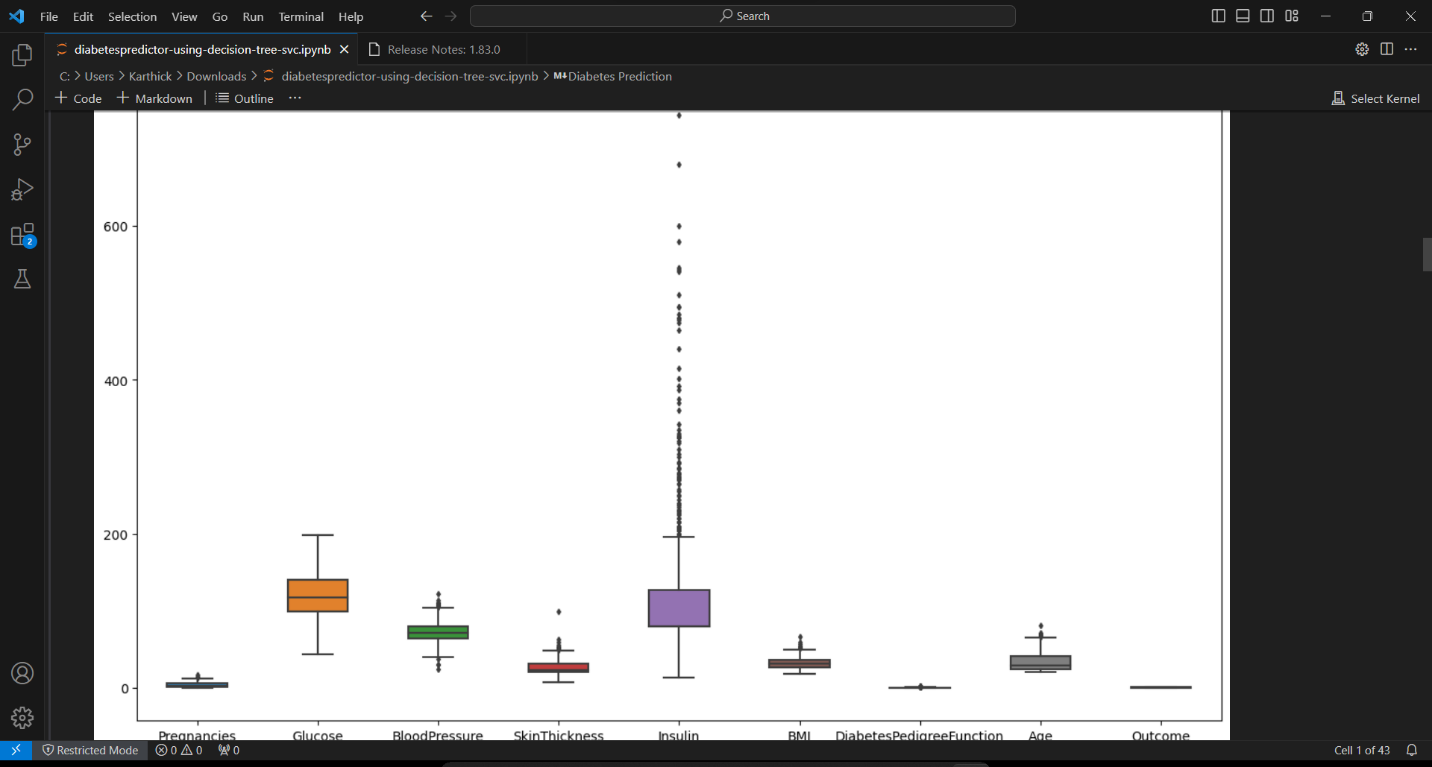
**DIABETES PREDICTION USING DECISION TREE TECHNIQUE:**

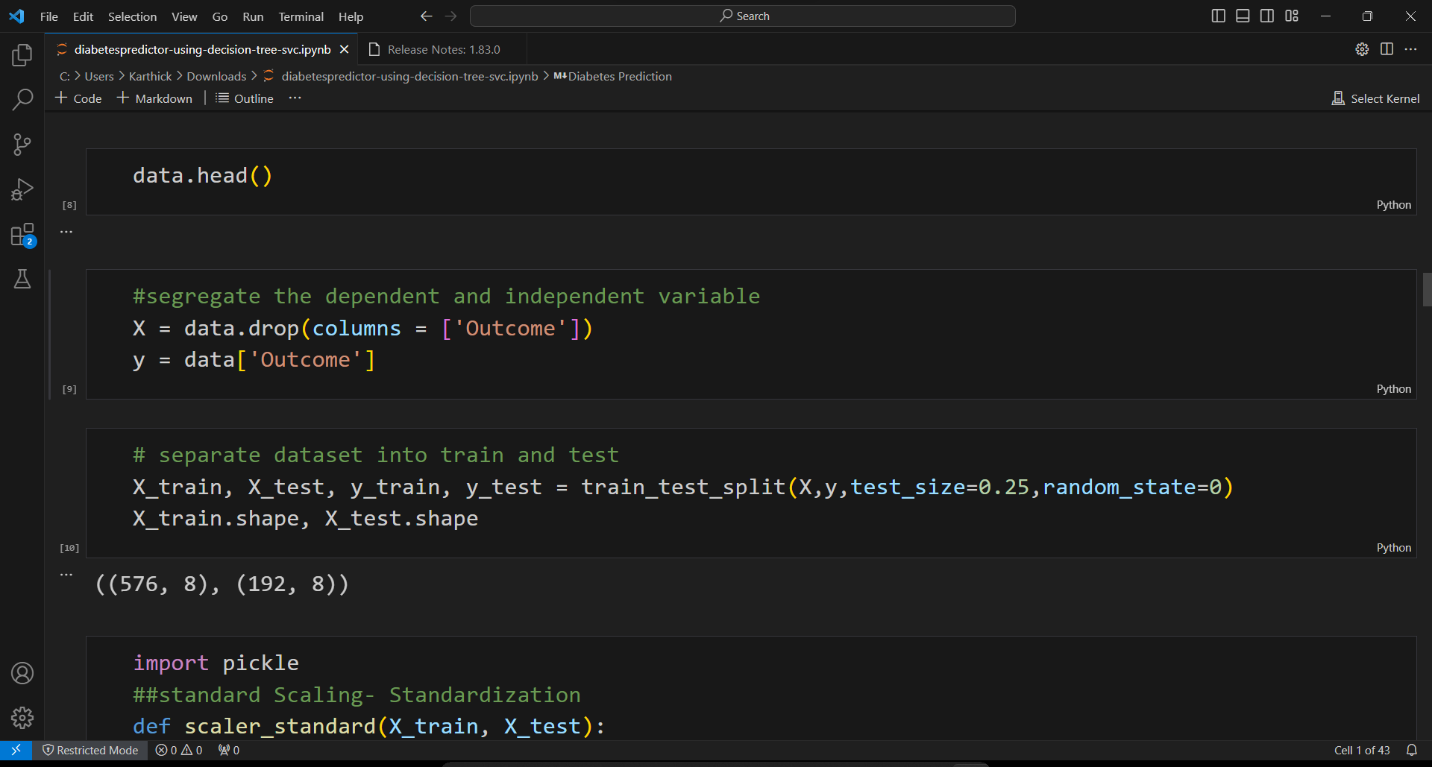
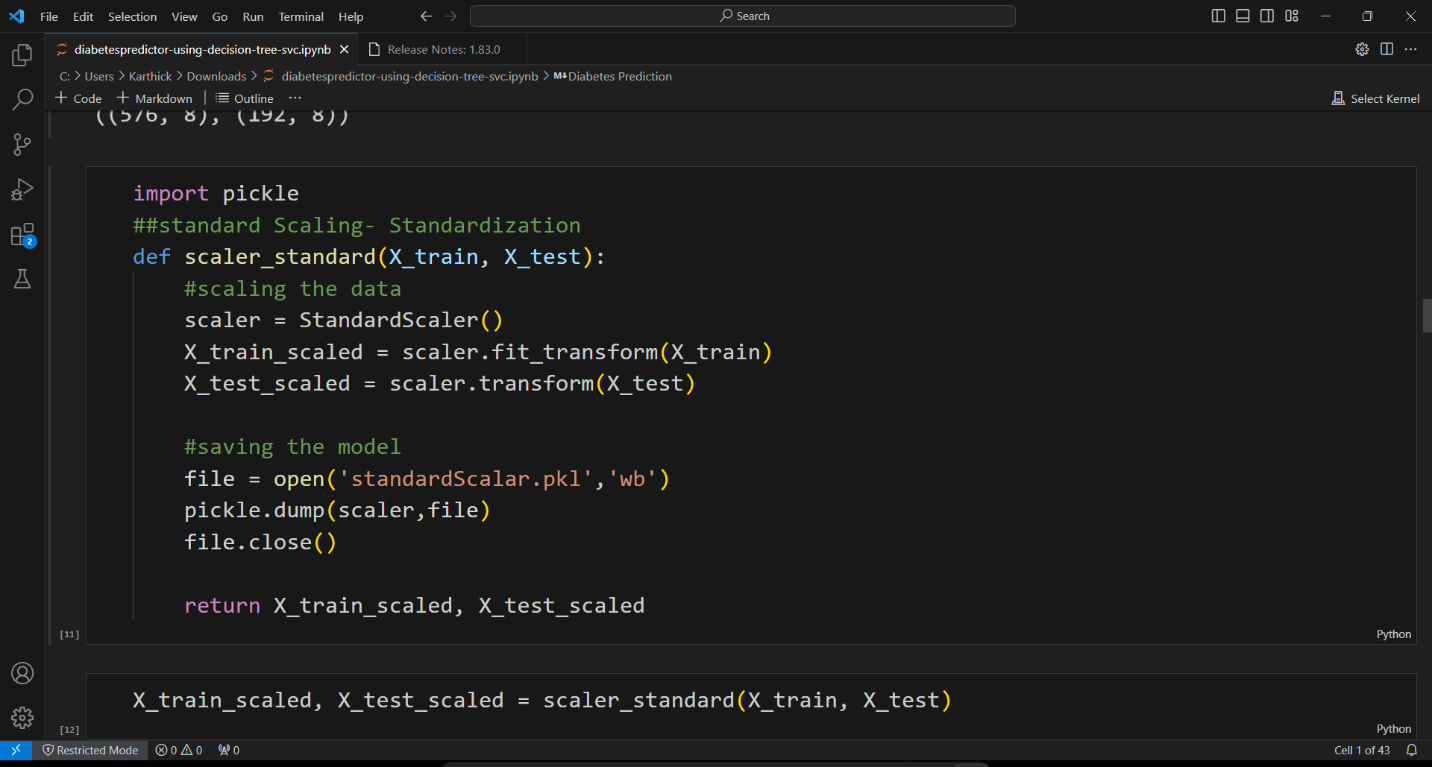
****

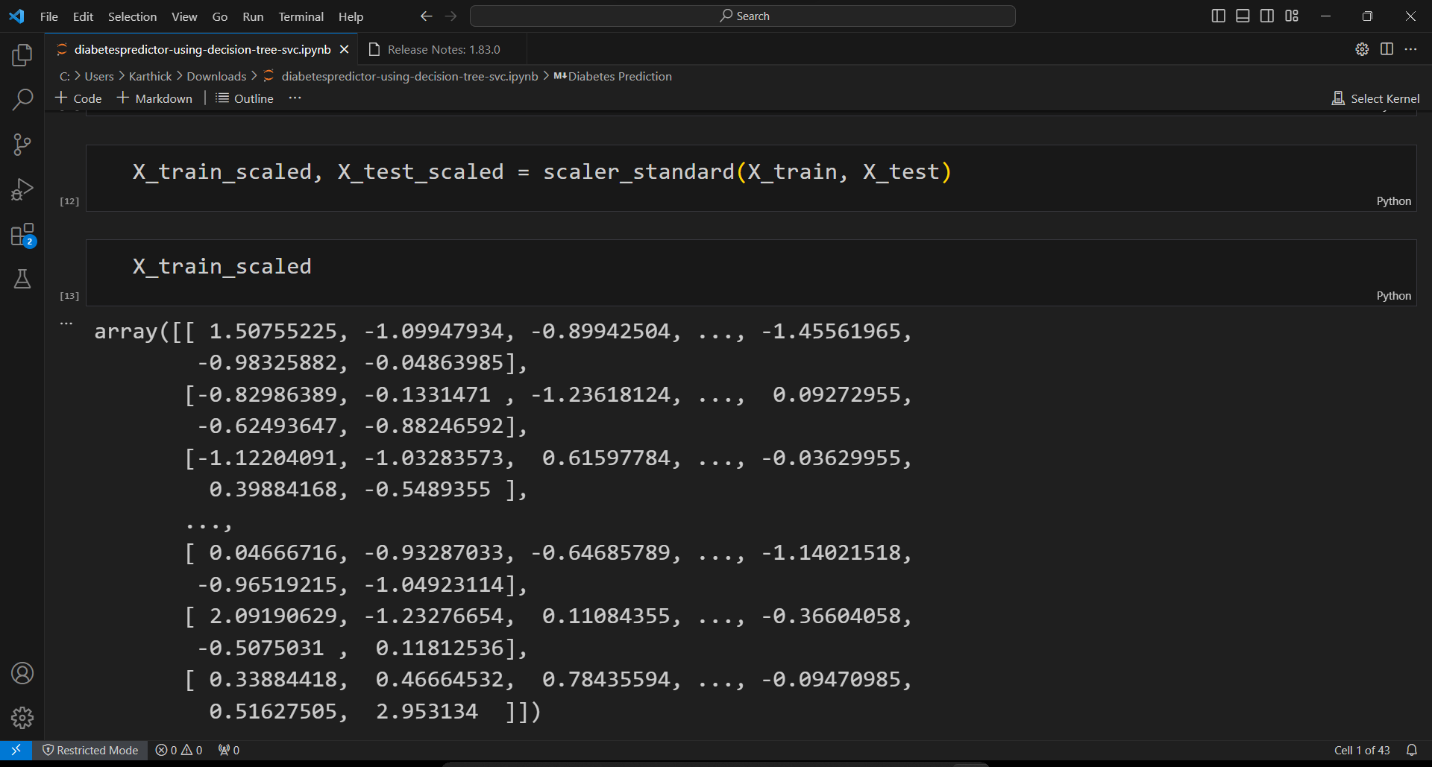
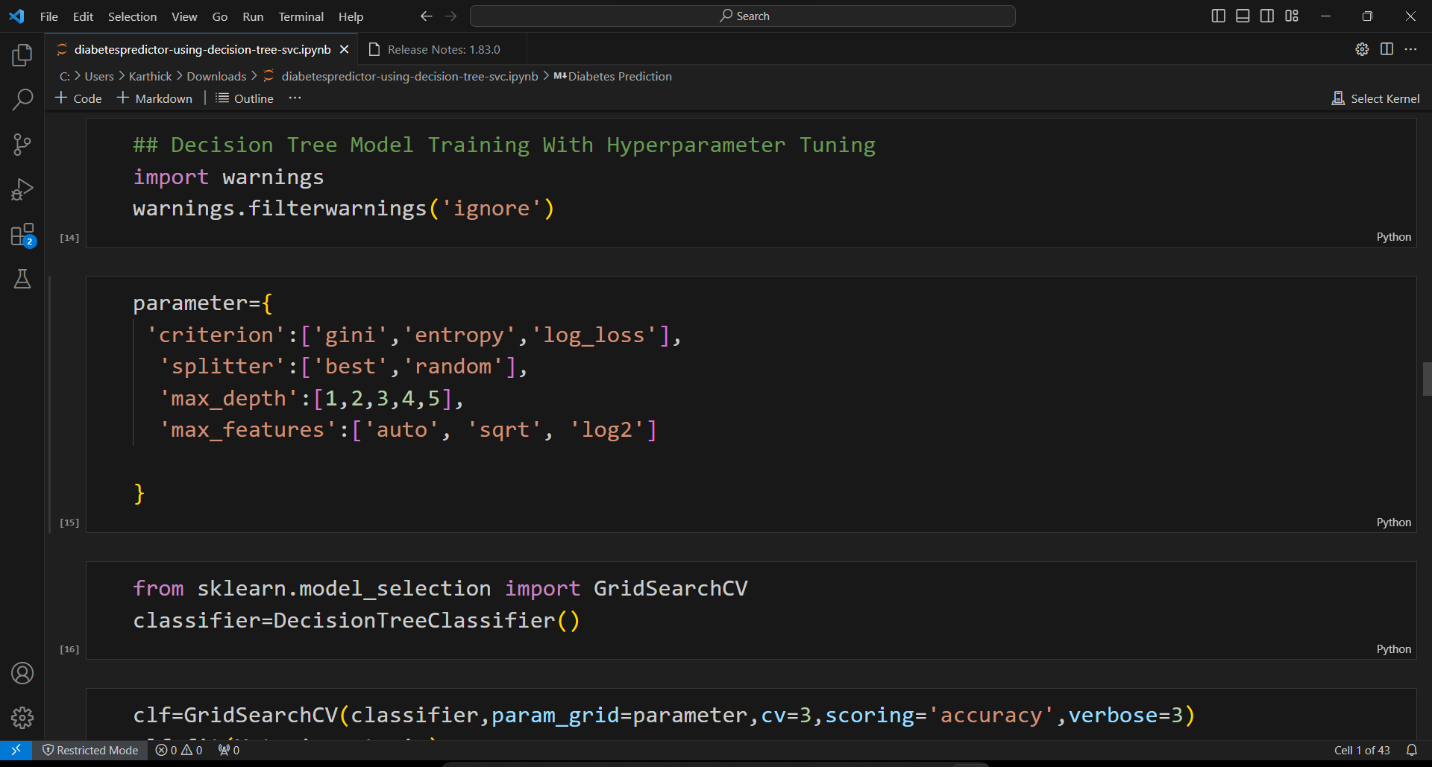
****

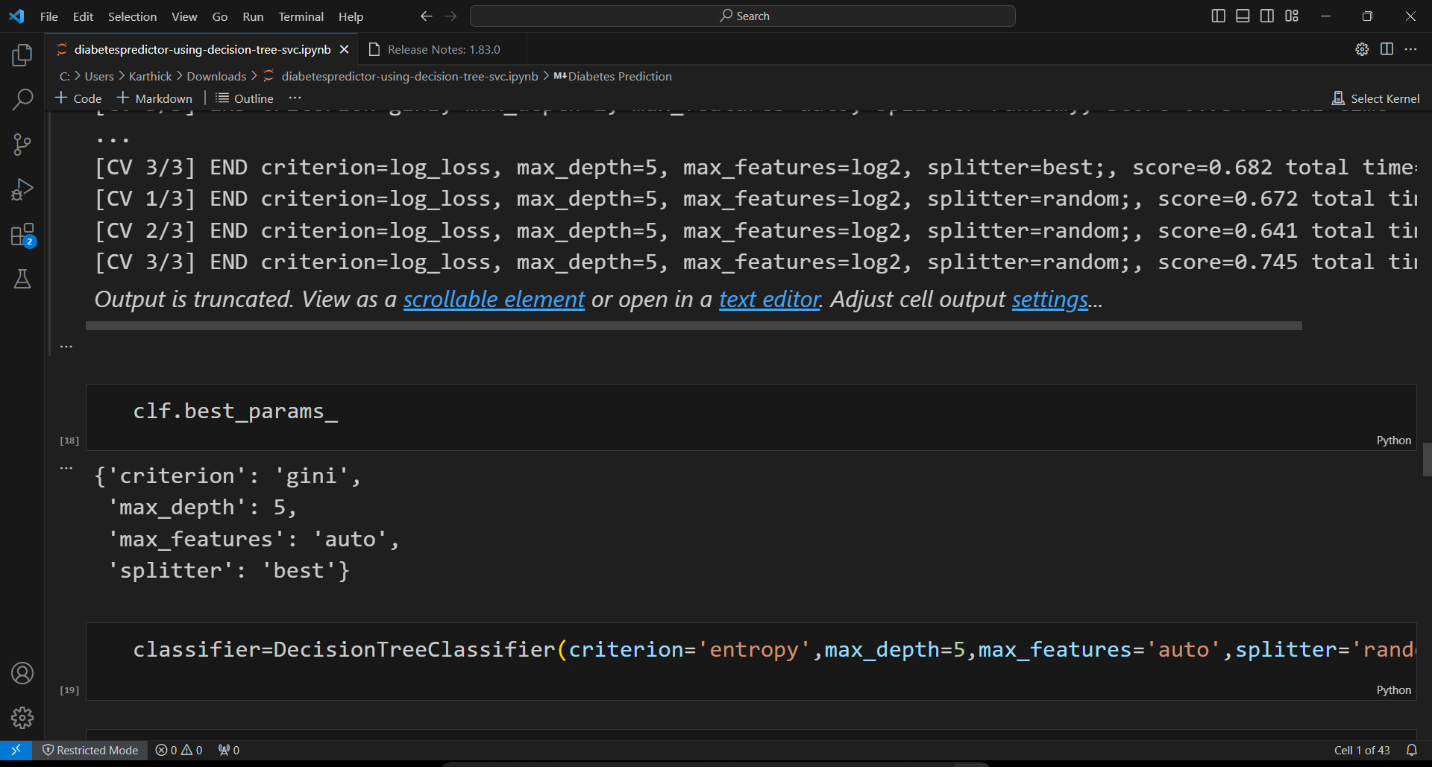
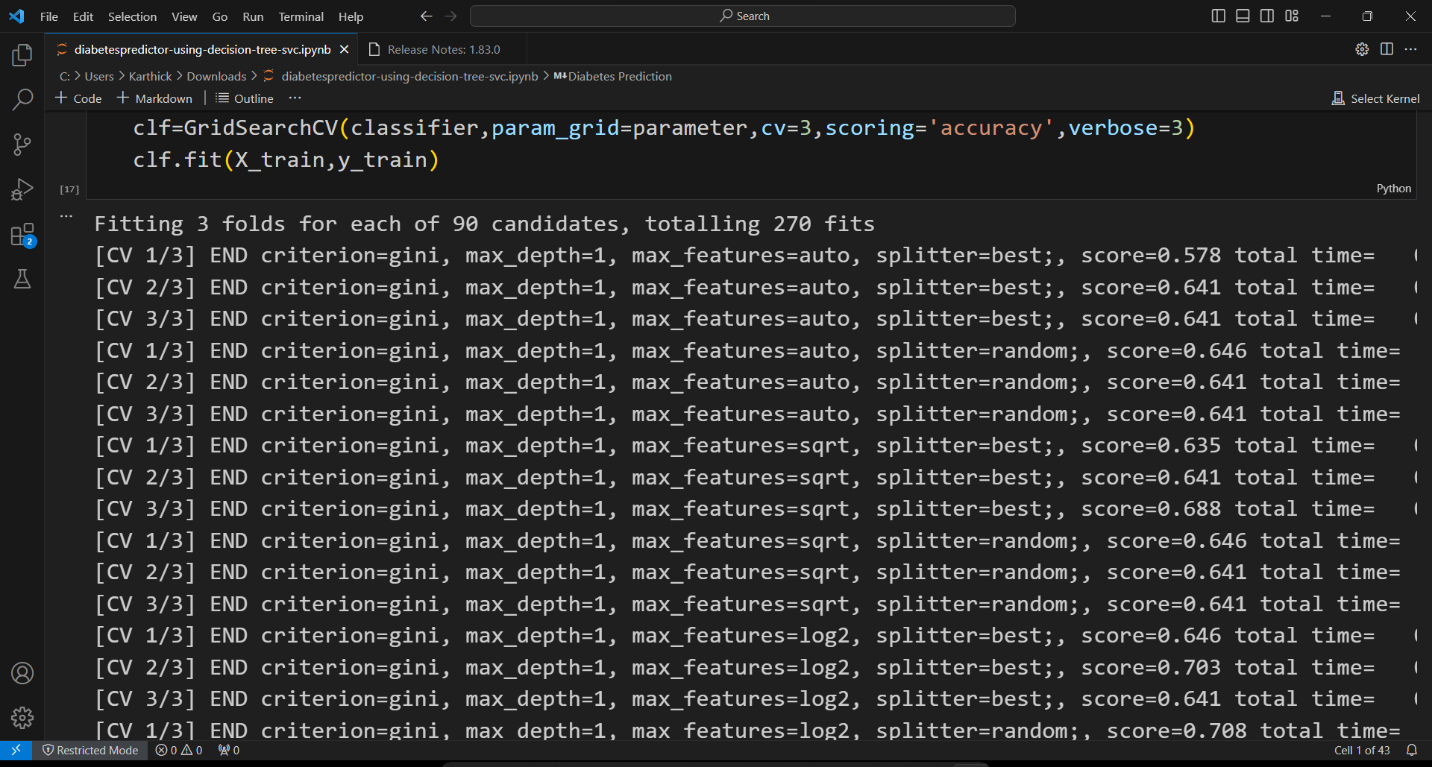
****

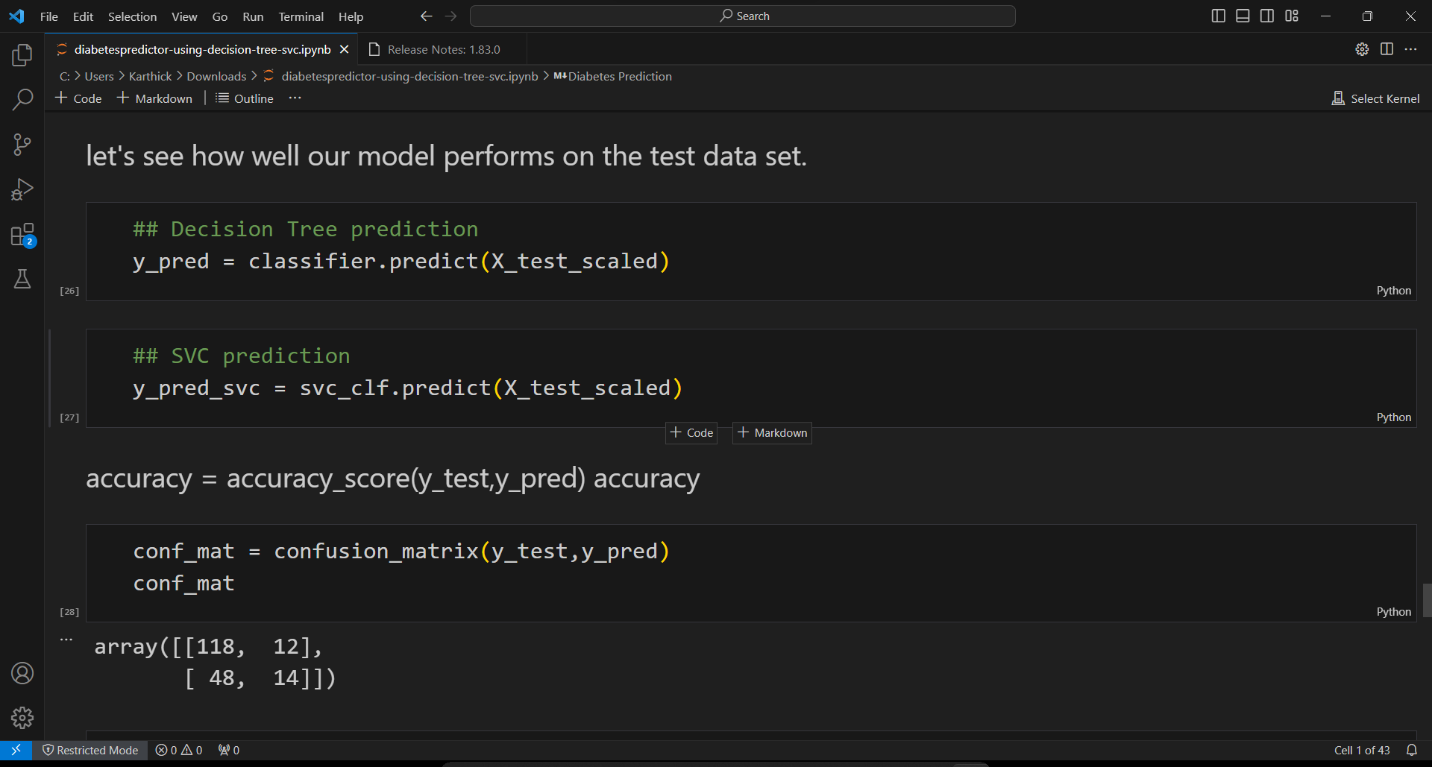
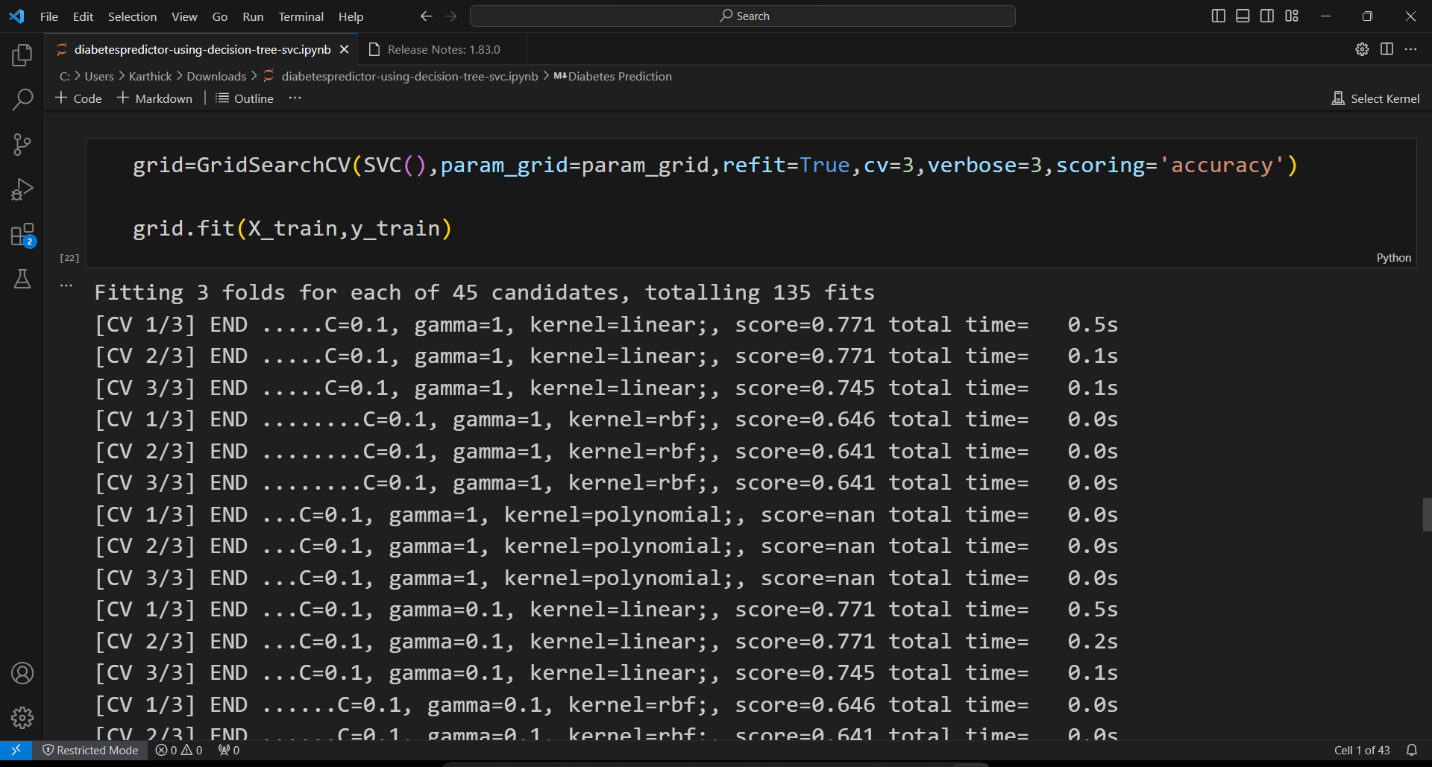
****

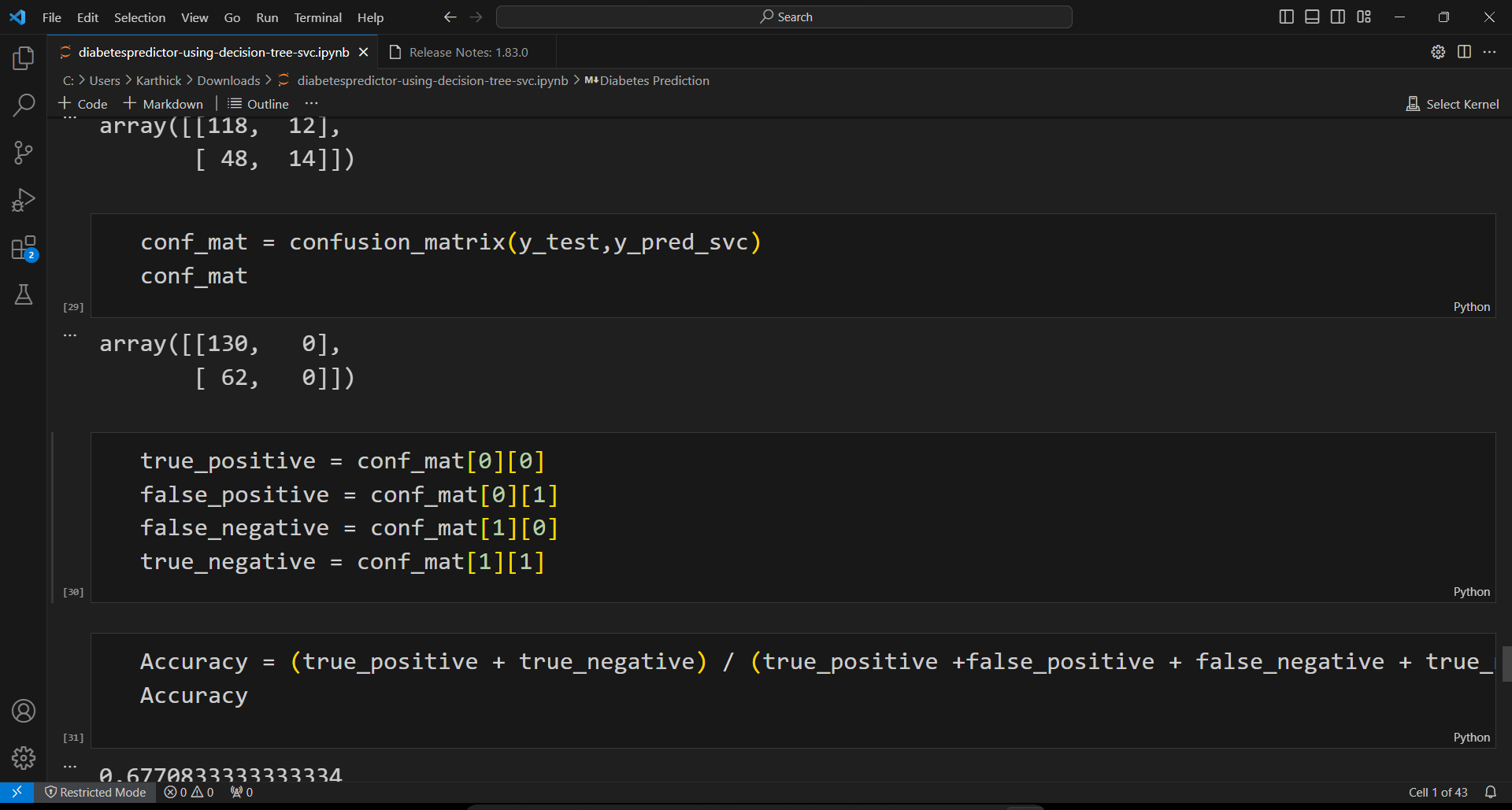
****

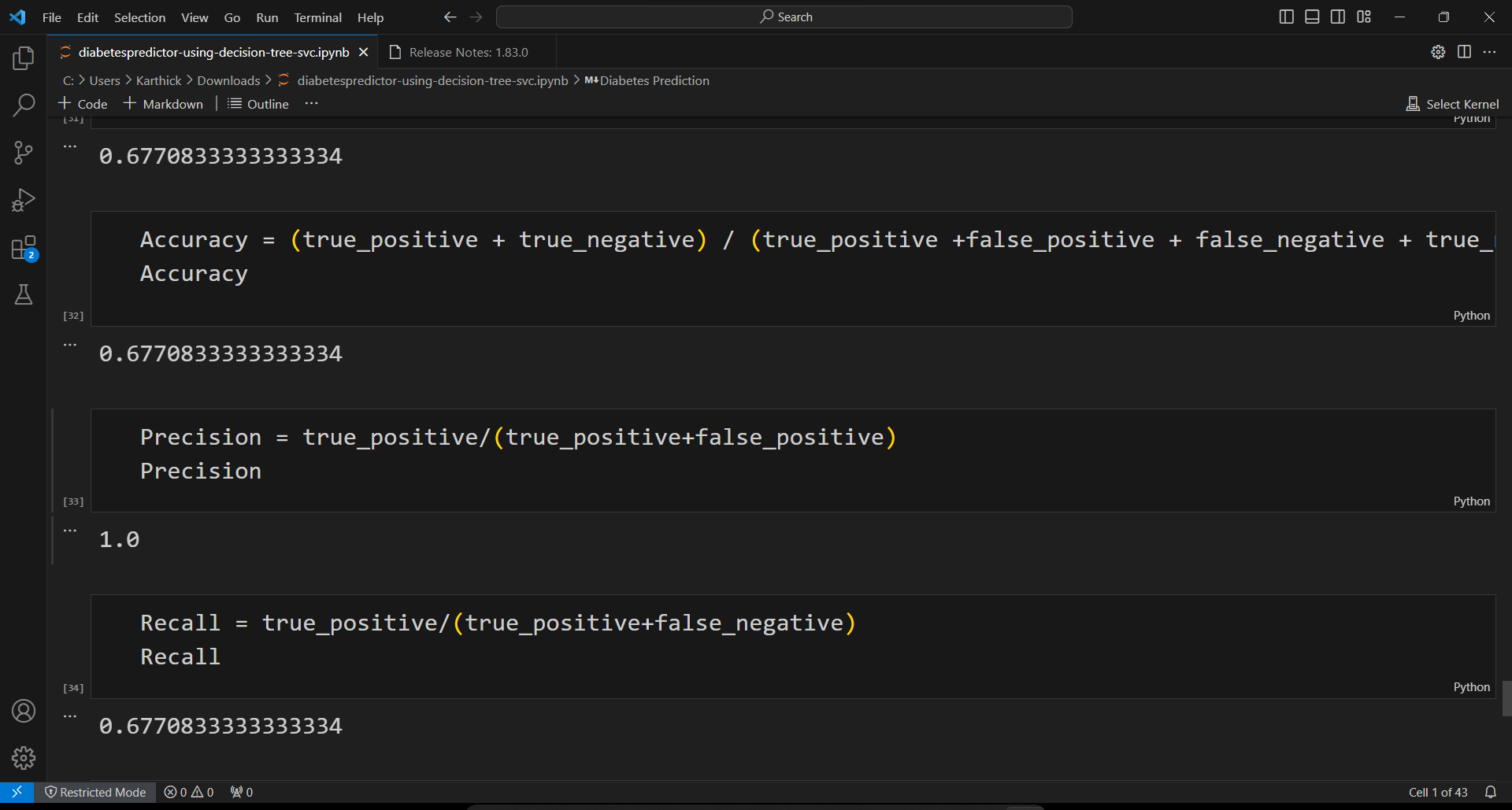
****

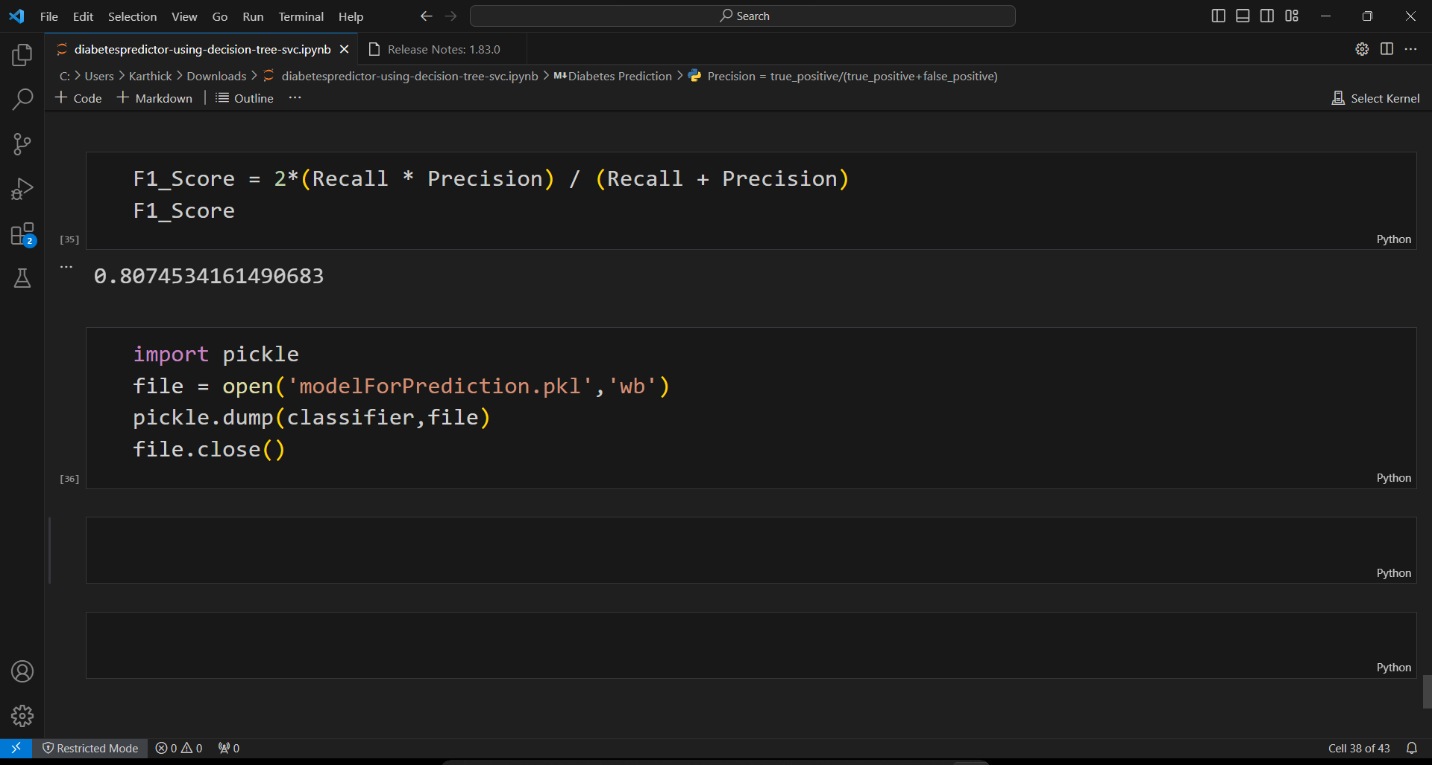
****

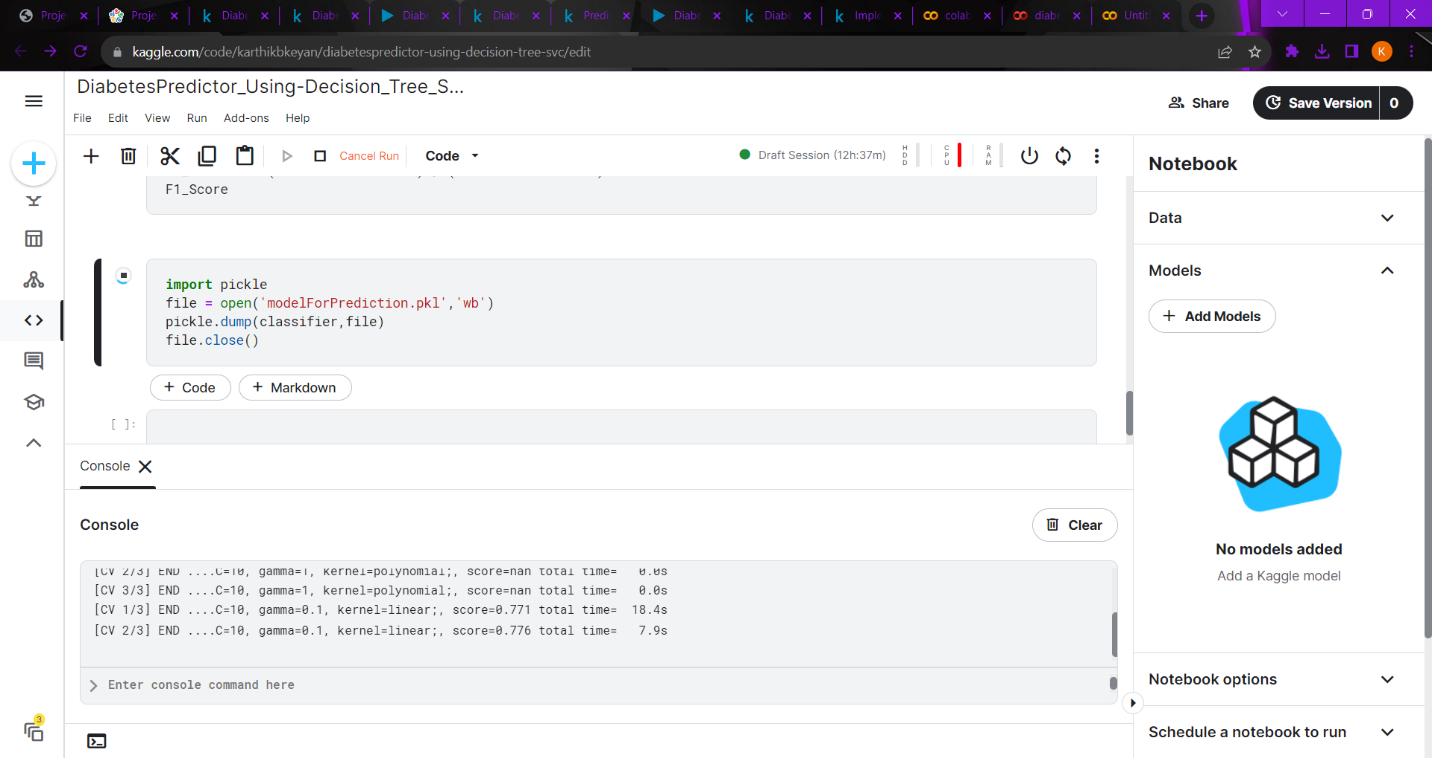
****

****

****

****

****

****

**OUTPUT:**

[CV 2/3] END ......C=10, gamma=1, kernel=linear;, score=0.776 total time= 7.9s

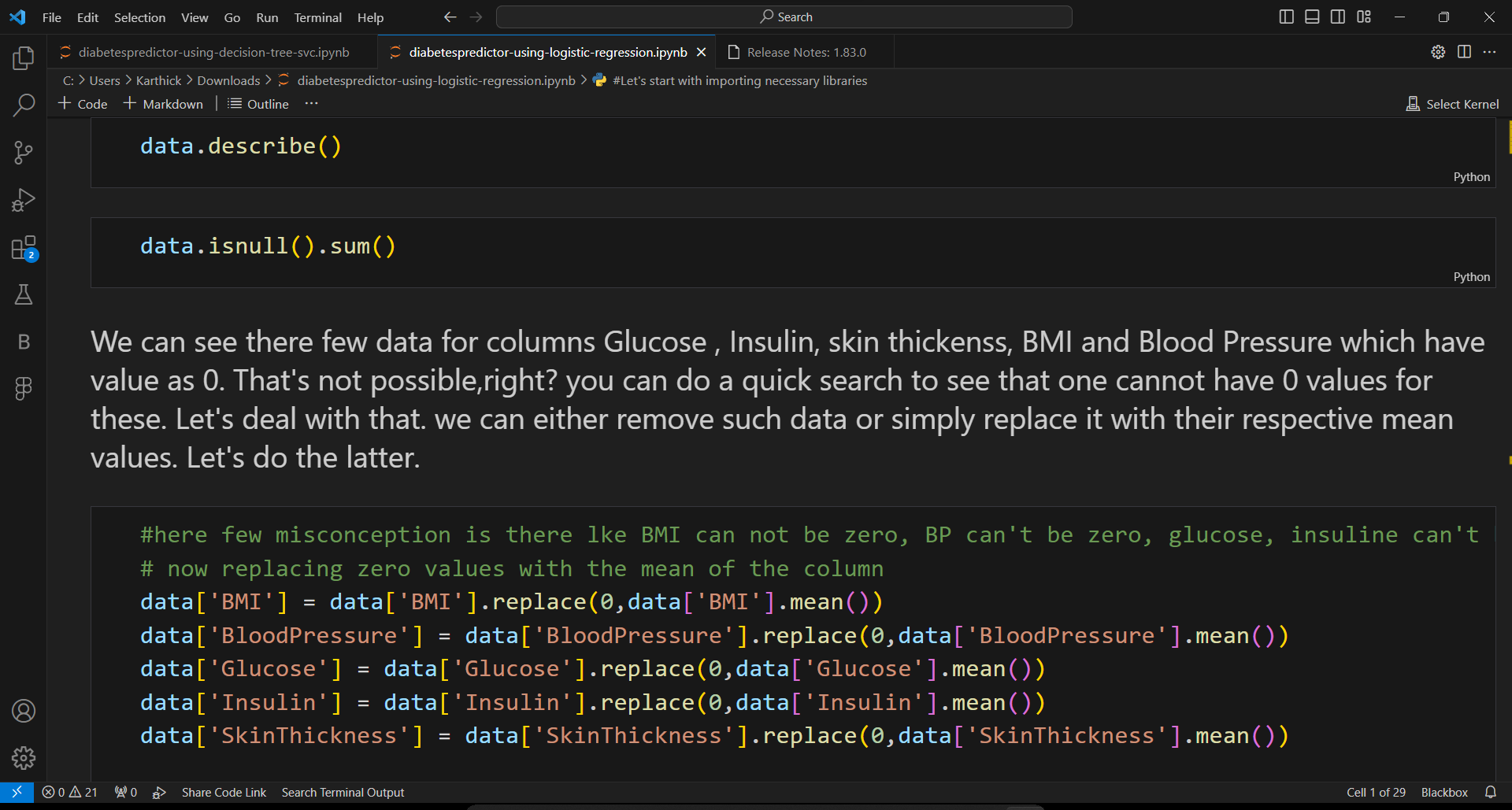
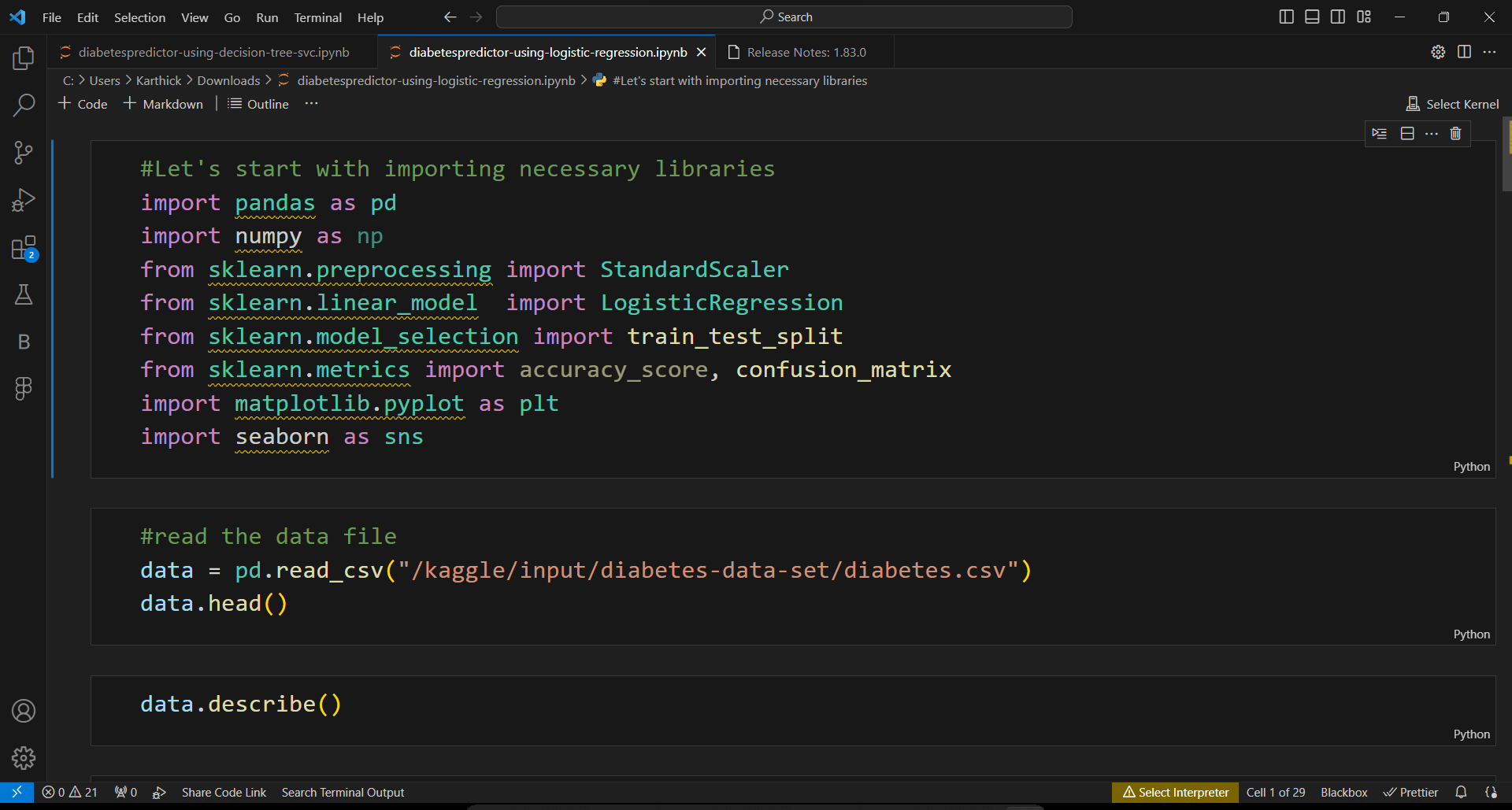
[CV 3/3] END ......C=10, gamma=1, kernel=linear;, score=0.740 total time= 5.0s

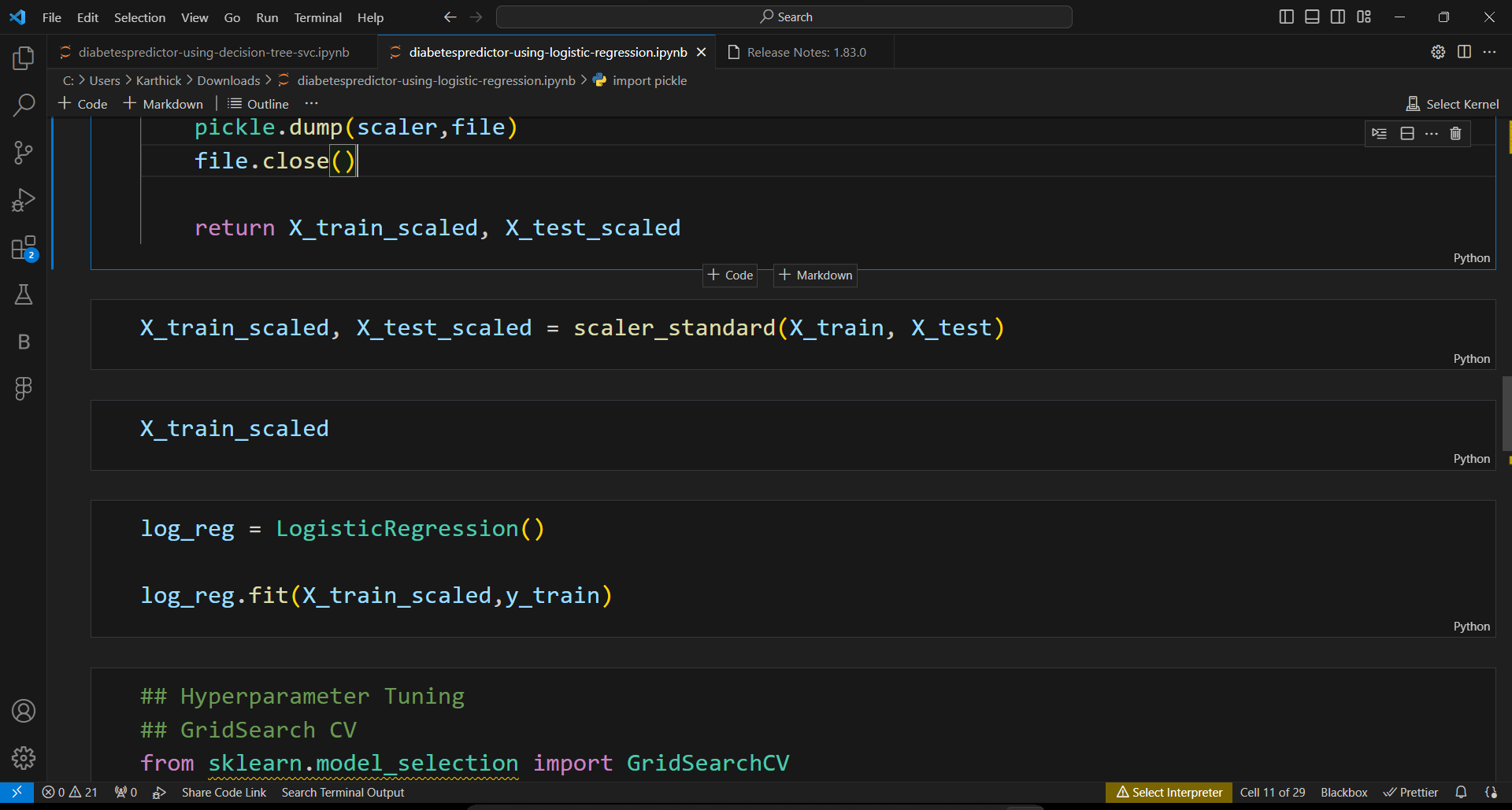
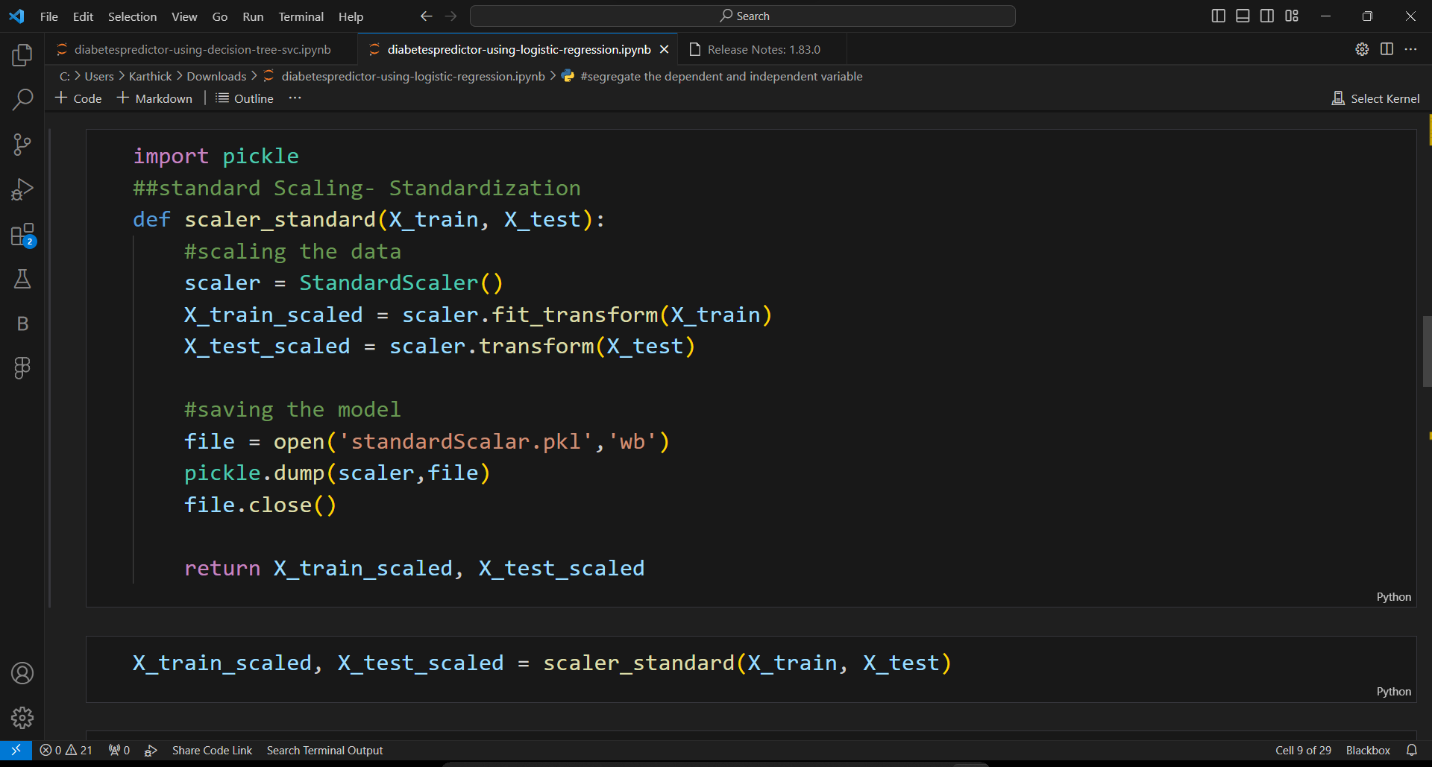
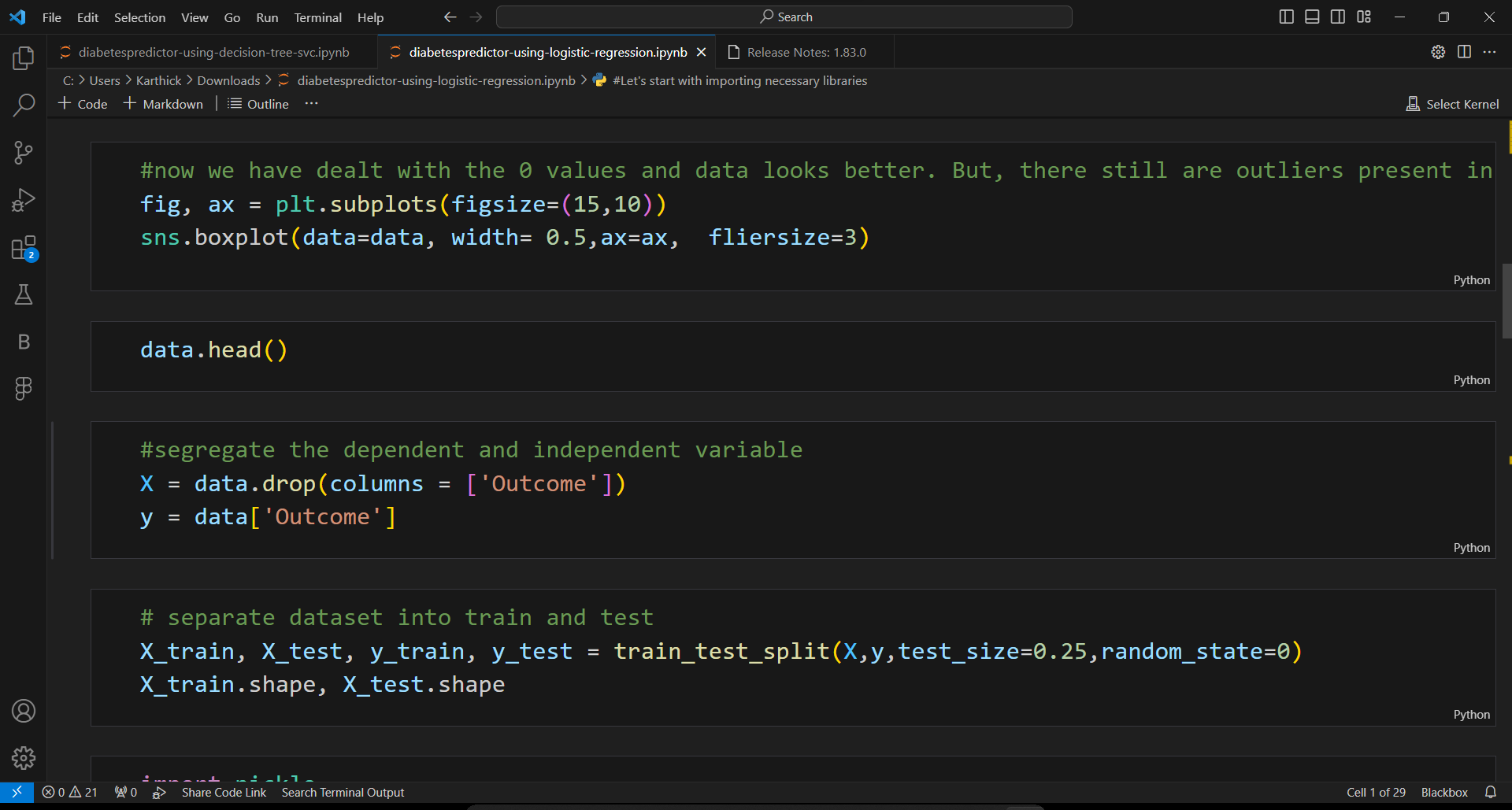
[CV 1/3] END .........C=10, gamma=1, kernel=rbf;, score=0.646 total time= 0.0s

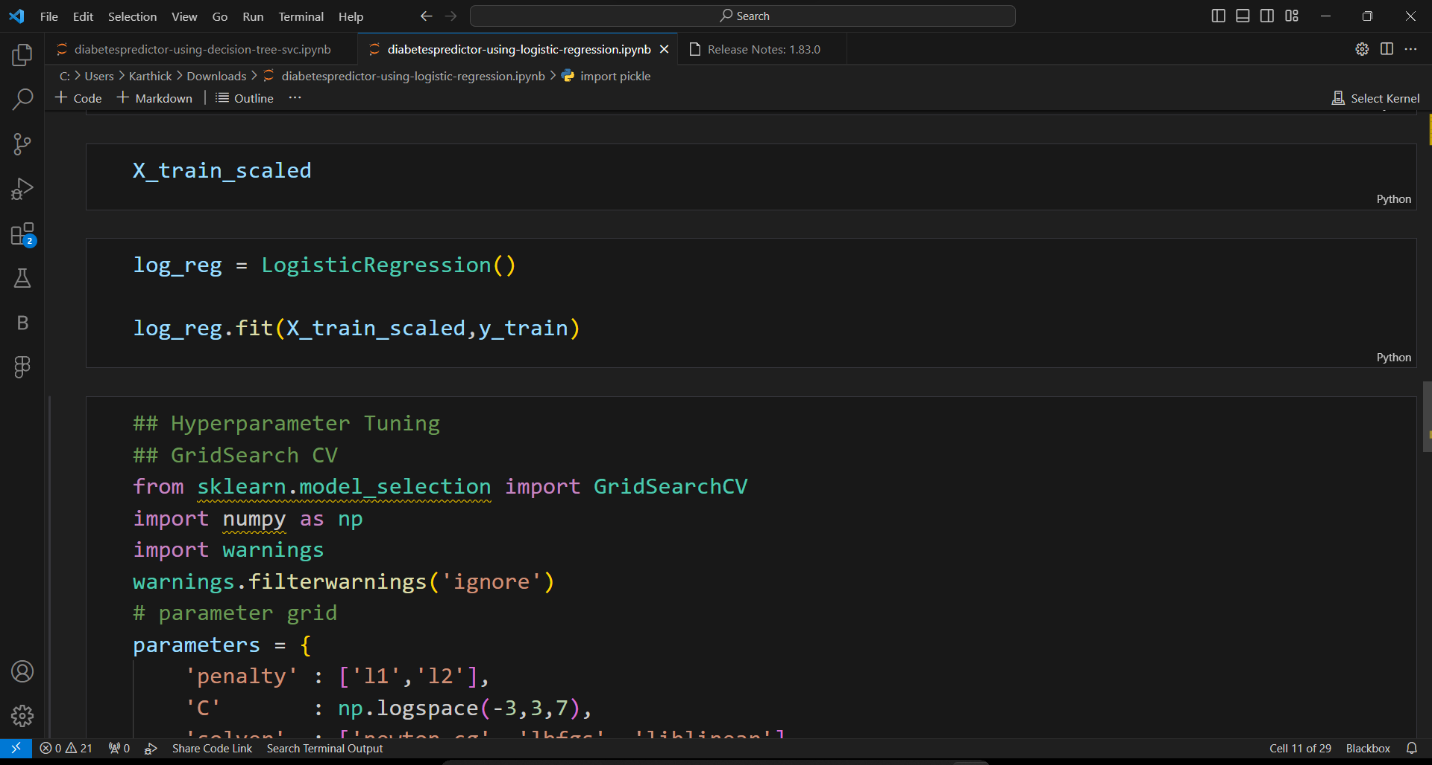
[CV 2/3] END .........C=10, gamma=1, kernel=rbf;, score=0.641 total time= 0.0s

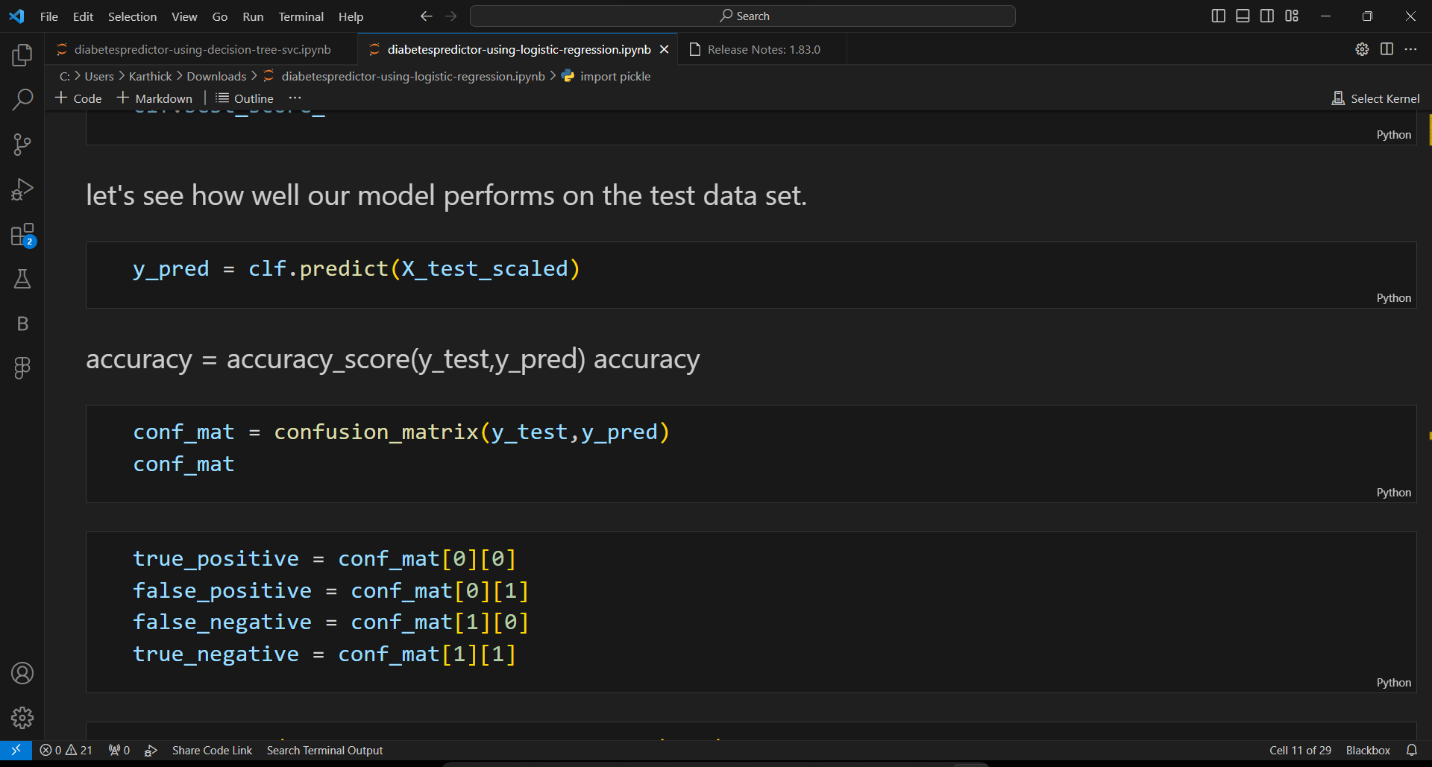
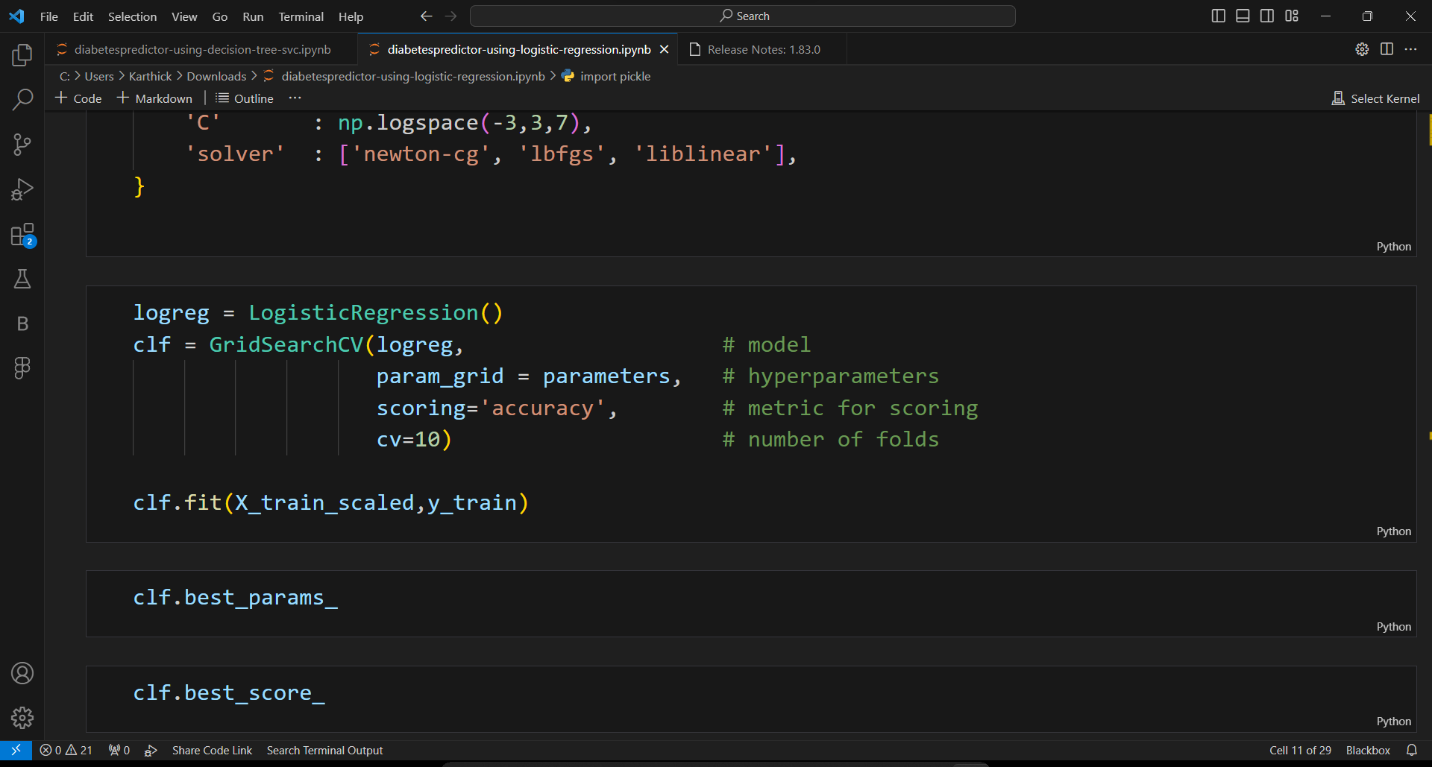
[CV 3/3] END .........C=10, gamma=1, kernel=rbf;, score=0.641 total time= 0.0s

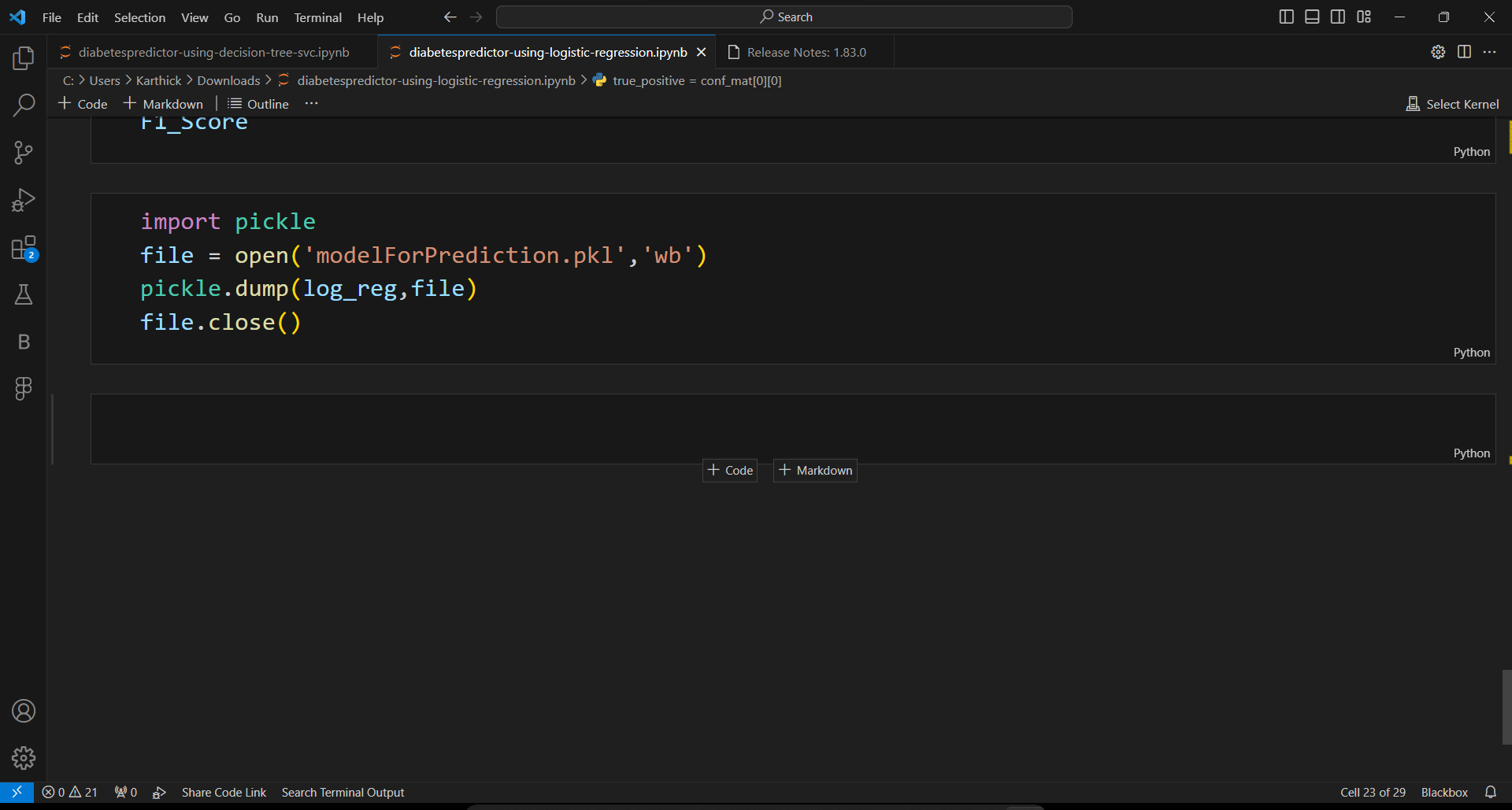
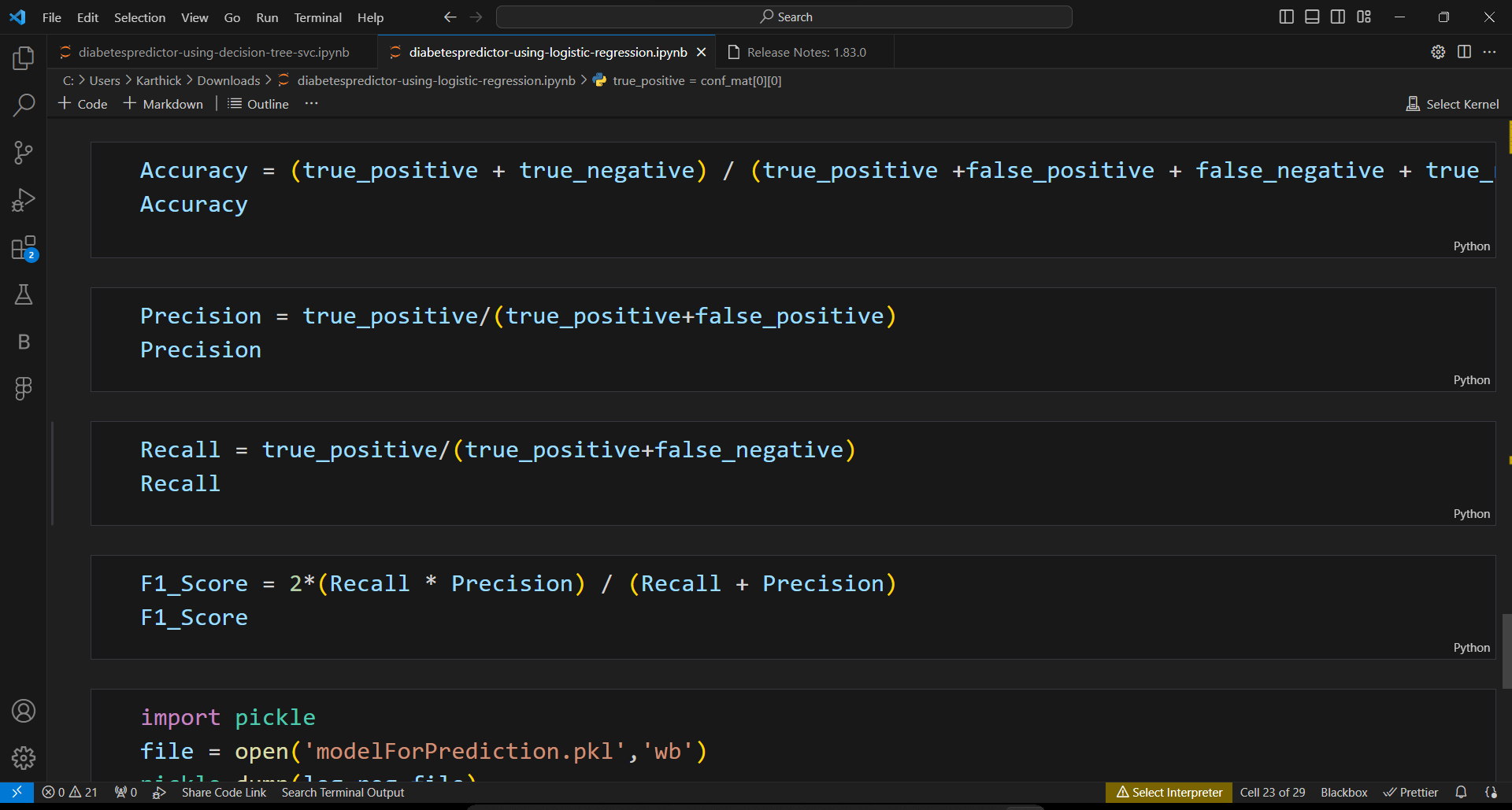
**DIABETES PREDICTION USING LOGISTIC REGRESSION:**

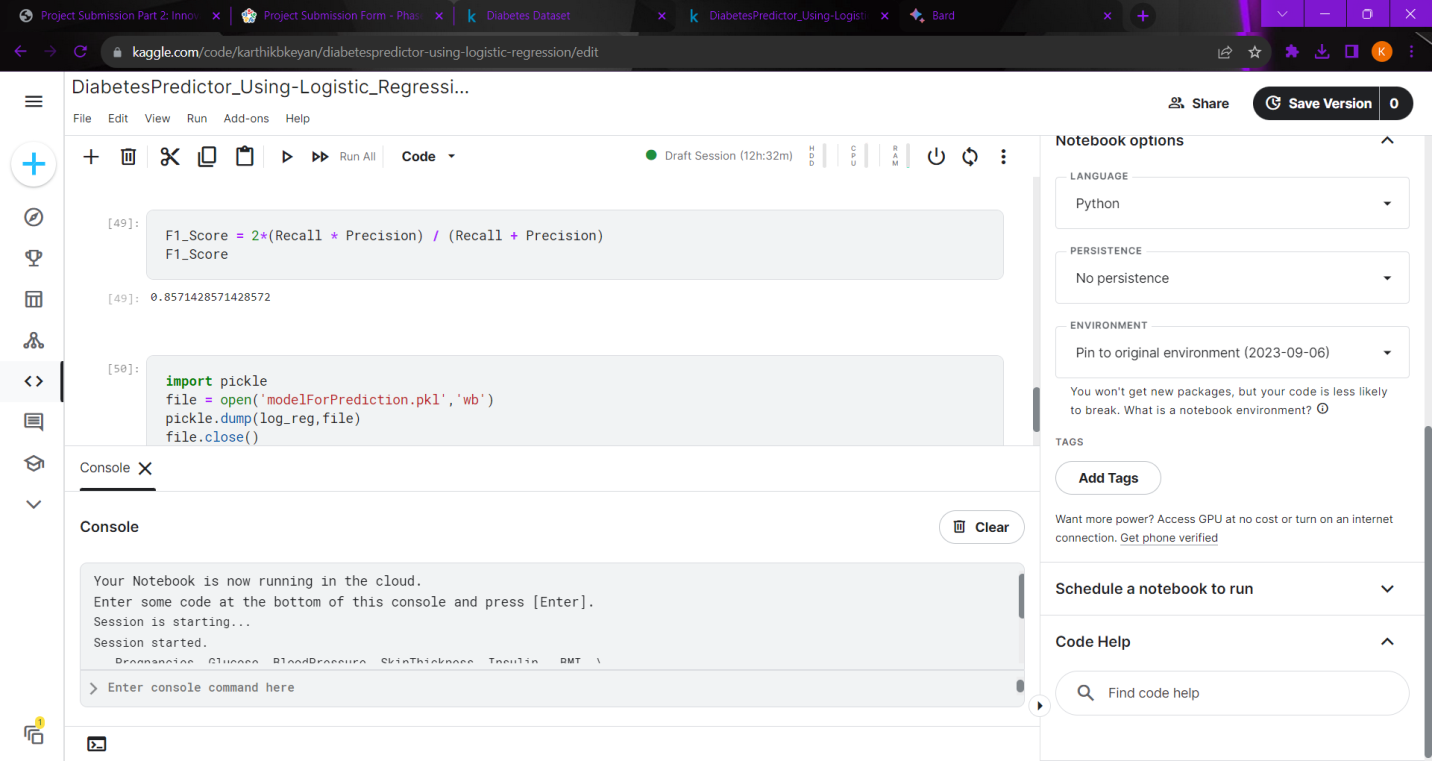
****

****

****

****

****

****

**OUTPUT:**

0.796875

0.9

0.8181818181818182

0.8571428571428572

**CONCLUSION:**

In phase 2 conclusion we will summarize the key findings and insights from the advanced logistic regression technique and decision technique. We will reiterate the impact of these techniques on improving the accuracy and robustness of diabetes prediction.