Iris

Write Code

(Clustering – K Means)

What’s in this session?

1. Question and Data
2. Create project
3. Add NuGet packages
4. Add using name space
5. Create data set input/output scheme
6. Set data set path
7. Load data

8. Add algorithm
9. Train the model
10. Save model
11. Evaluate the model and show accuracy stats
12. Predict single item

Question and Data

Question: Which group is this flower belong to?

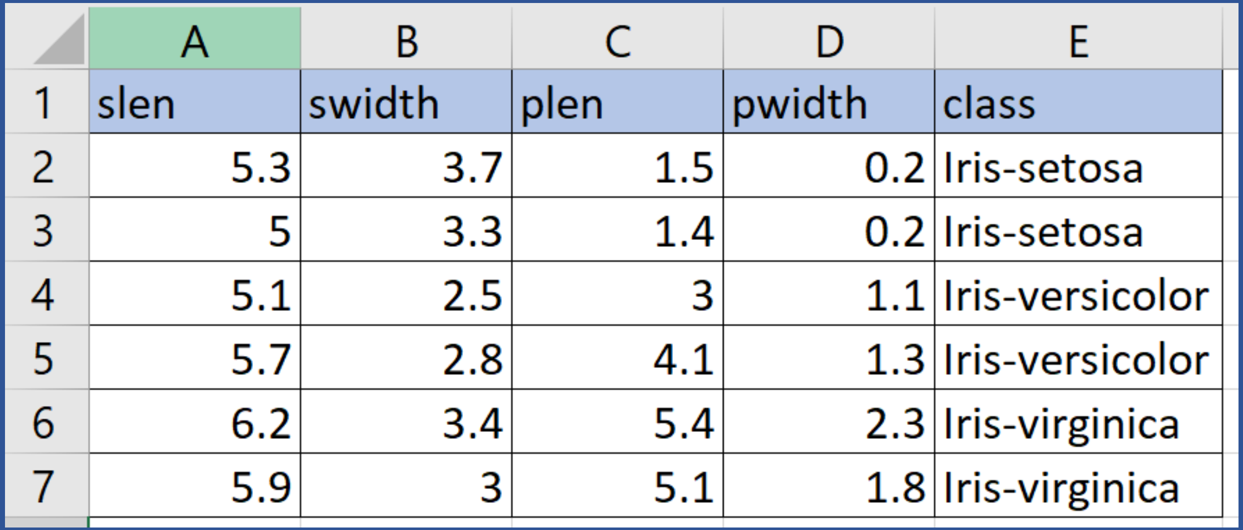
Dataset:

Train
https://raw.githubusercontent.com/laploy/ML.NET/master/Iris/iris-data-train.csv

Test
https://raw.githubusercontent.com/laploy/ML.NET/master/Taxi-fare/taxi-fare-score.csv

Dataset description

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
   -- Iris Setosa
   -- Iris Versicolour
   -- Iris Virginica



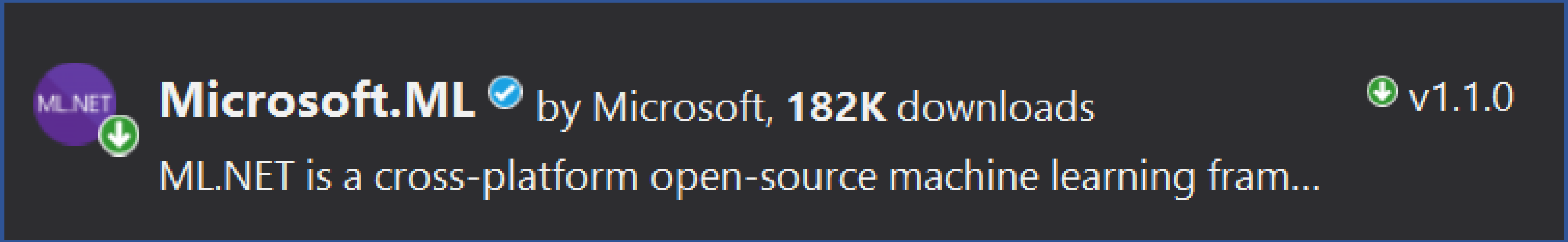
	A	B	C	D	E
1	slen	swidth	plen	pwidth	class
2	5.3	3.7	1.5	0.2	Iris-setosa
3	5	3.3	1.4	0.2	Iris-setosa
4	5.1	2.5	3	1.1	Iris-versicolor
5	5.7	2.8	4.1	1.3	Iris-versicolor
6	6.2	3.4	5.4	2.3	Iris-virginica
7	5.9	3	5.1	1.8	Iris-virginica

Create

Create new .NET CORE console app project name = “Iris”

Add NuGet Package

- Microsoft.ML

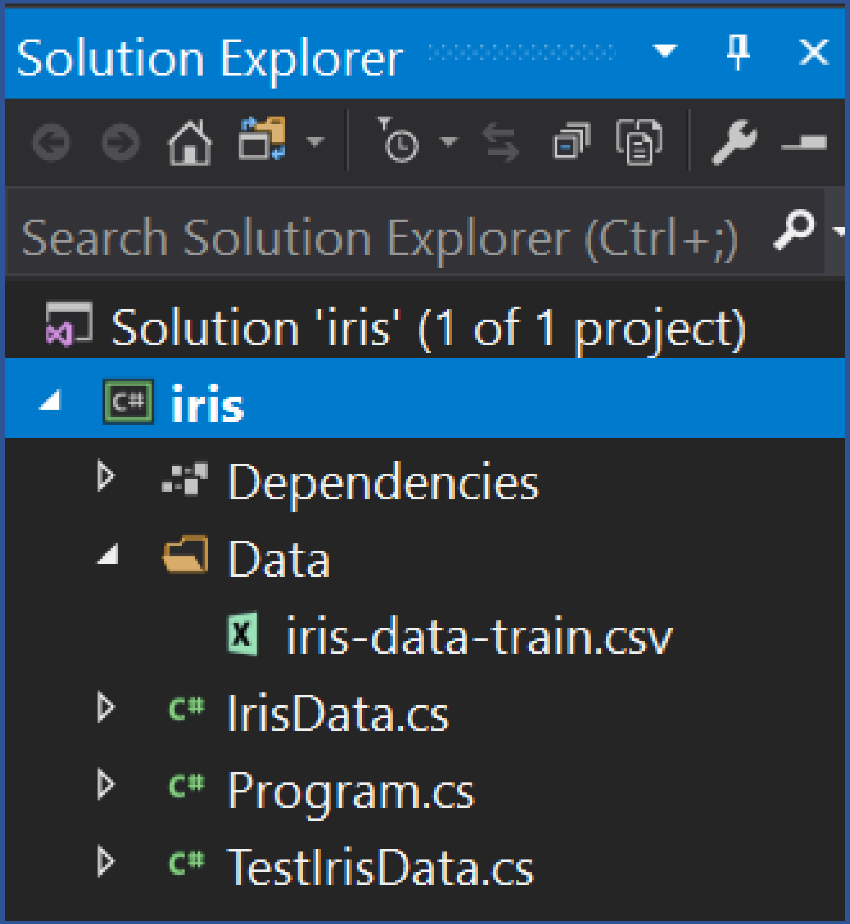


Microsoft.ML by Microsoft, **182K** downloads

v1.1.0

ML.NET is a cross-platform open-source machine learning fram...

Add new folder "Data"
Copy datasets to this folder

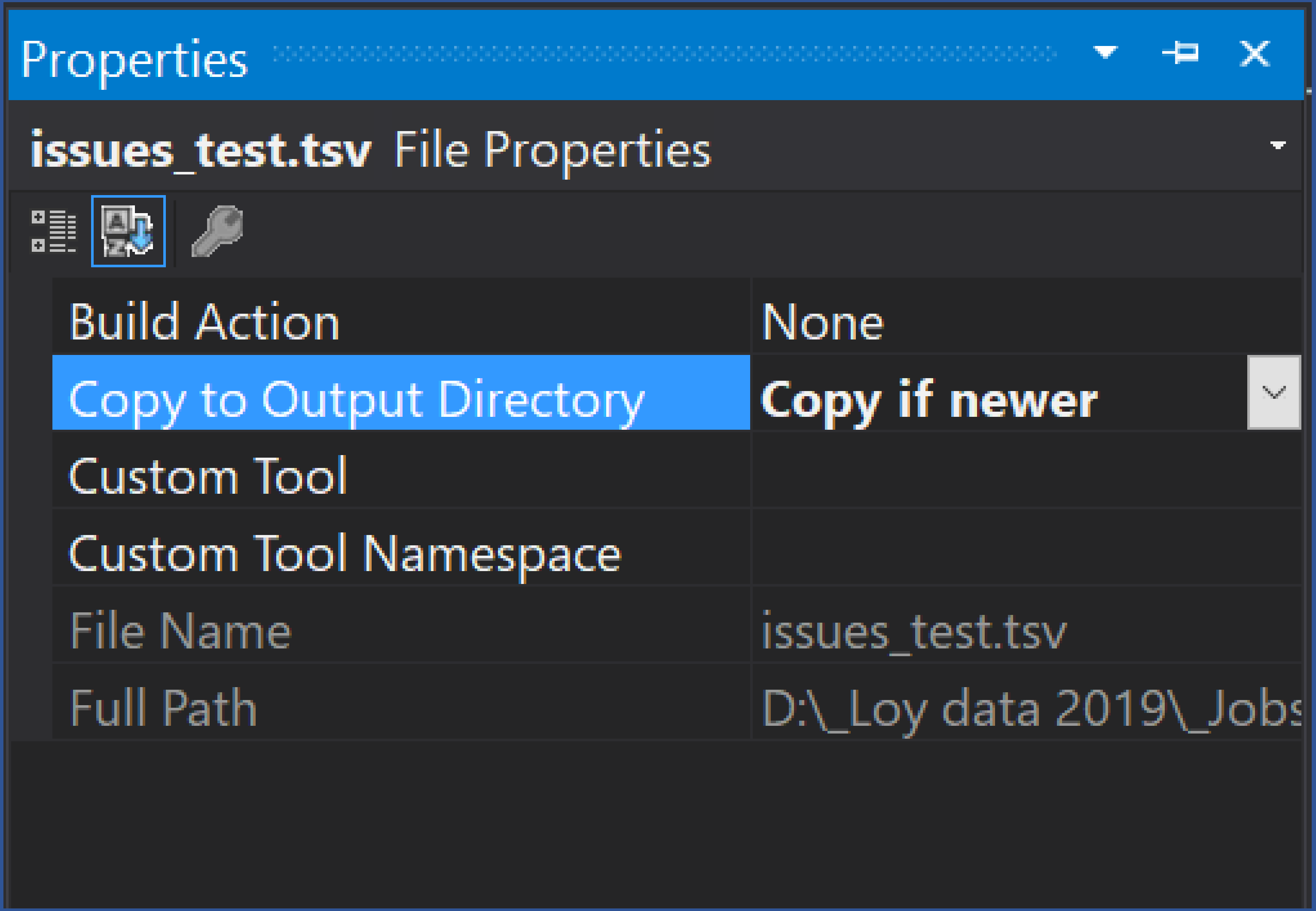


Solution Explorer

Search Solution Explorer (Ctrl+;)

Solution 'iris' (1 of 1 project)
- iris
  - Dependencies
  - Data
    - iris-data-train.csv
  - IrisData.cs
  - Program.cs
  - TestIrisData.cs

Set property of each datasets to “Copy if newer”

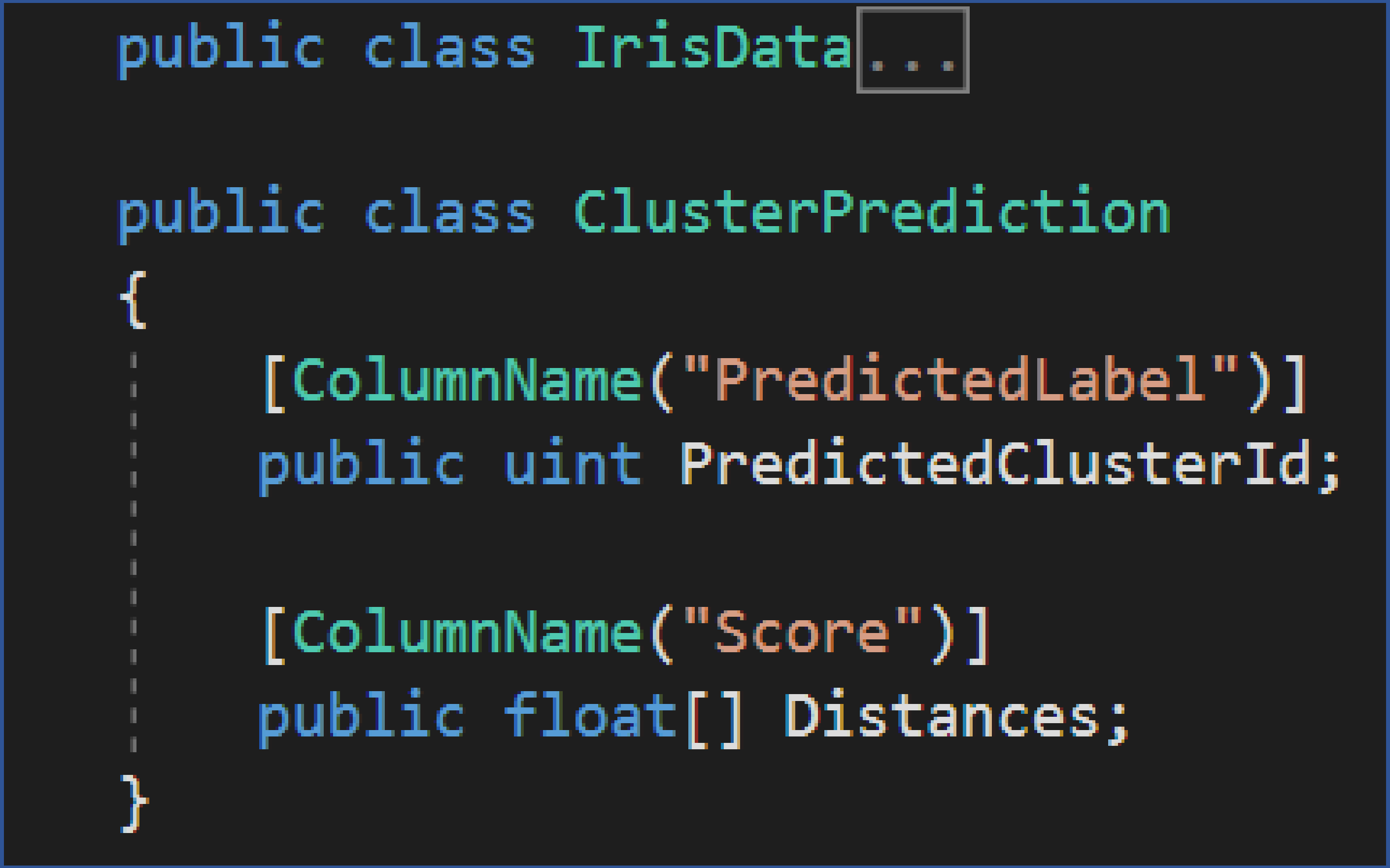


Properties

issues_test.tsv File Properties

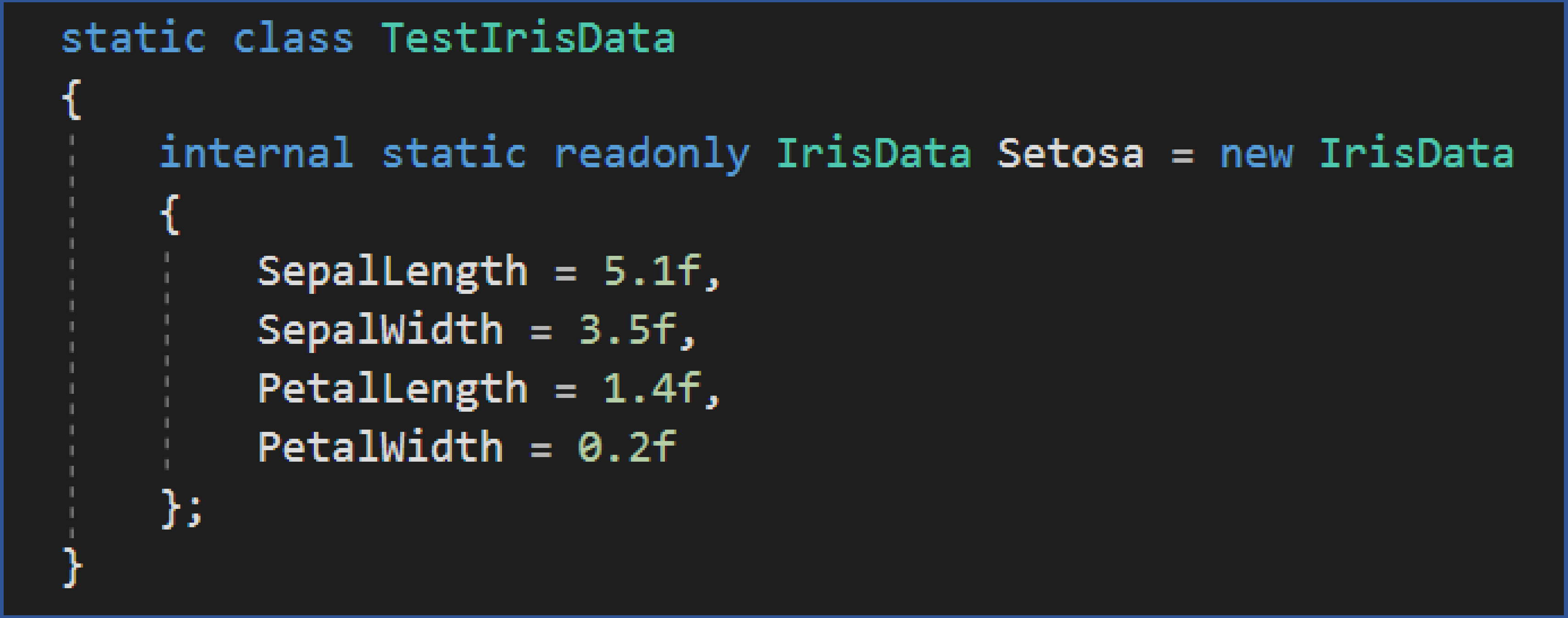
Build Action	None
Copy to Output Directory	**Copy if newer**
Custom Tool	
Custom Tool Namespace	
File Name	issues_test.tsv
Full Path	D:_Loy data 2019_Jobs

Create data set input/output scheme



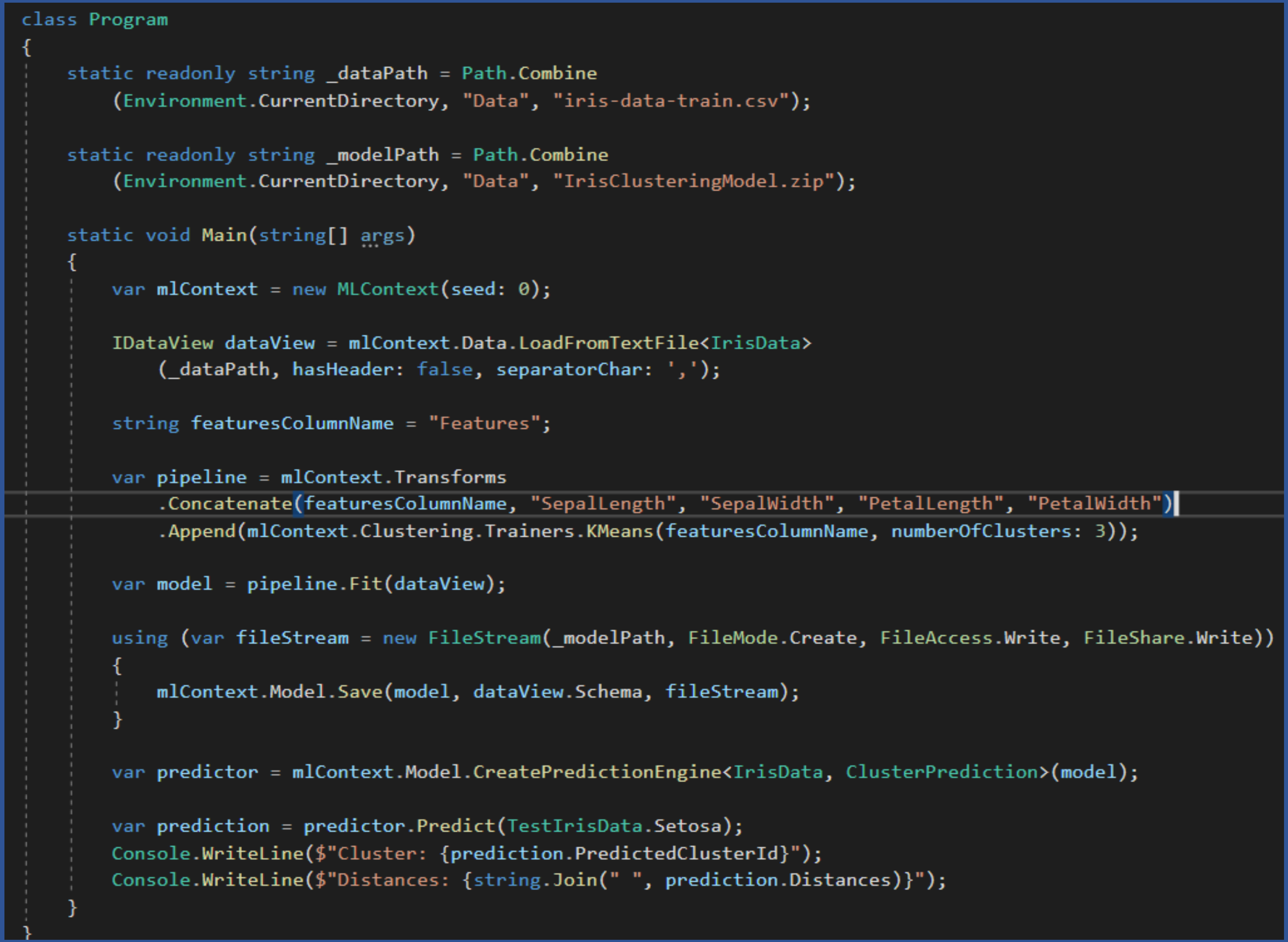
```
public class IrisData...

public class ClusterPrediction
{
    [ColumnName("PredictedLabel")]
    public uint PredictedClusterId;

    [ColumnName("Score")]
    public float[] Distances;
}
```

Create class for test data

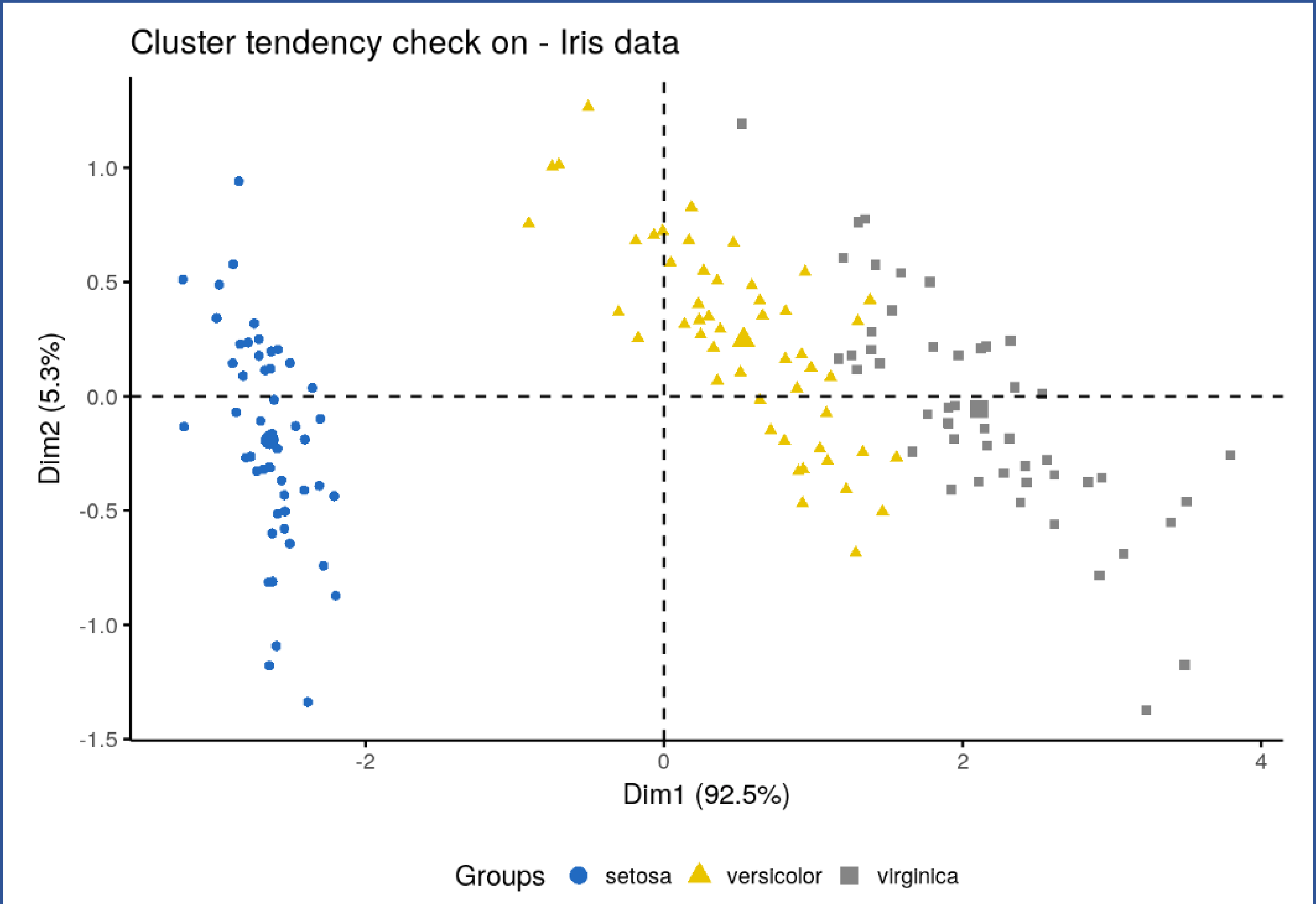


```
static class TestIrisData
{
    internal static readonly IrisData Setosa = new IrisData
    {
        SepalLength = 5.1f,
        SepalWidth = 3.5f,
        PetalLength = 1.4f,
        PetalWidth = 0.2f
    };
}
```

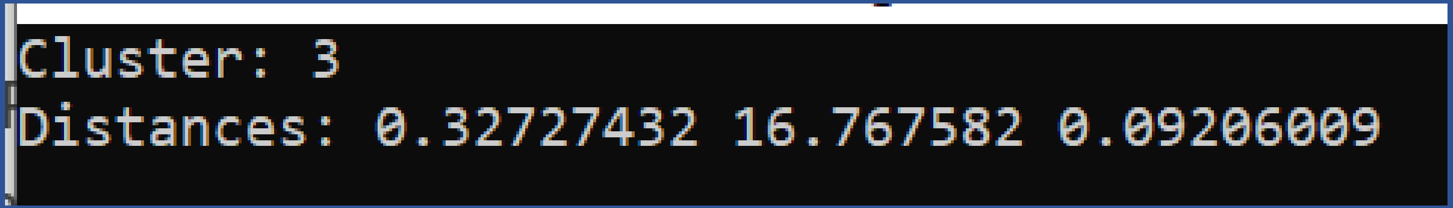
Write code in class Program



```
class Program
{
    static readonly string _dataPath = Path.Combine
        (Environment.CurrentDirectory, "Data", "iris-data-train.csv");

    static readonly string _modelPath = Path.Combine
        (Environment.CurrentDirectory, "Data", "IrisClusteringModel.zip");

    static void Main(string[] args)
    {
        var mlContext = new MLContext(seed: 0);

        IDataView dataView = mlContext.Data.LoadFromTextFile<IrisData>
            (_dataPath, hasHeader: false, separatorChar: ',');

        string featuresColumnName = "Features";

        var pipeline = mlContext.Transforms
            .Concatenate(featuresColumnName, "SepalLength", "SepalWidth", "PetalLength", "PetalWidth")
            .Append(mlContext.Clustering.Trainers.KMeans(featuresColumnName, numberOfClusters: 3));

        var model = pipeline.Fit(dataView);

        using (var fileStream = new FileStream(_modelPath, FileMode.Create, FileAccess.Write, FileShare.Write))
        {
            mlContext.Model.Save(model, dataView.Schema, fileStream);
        }

        var predictor = mlContext.Model.CreatePredictionEngine<IrisData, ClusterPrediction>(model);

        var prediction = predictor.Predict(TestIrisData.Setosa);
        Console.WriteLine($"Cluster: {prediction.PredictedClusterId}");
        Console.WriteLine($"Distances: {string.Join(" ", prediction.Distances)}");
    }
}
```


Cluster tendency check on - Iris data
1.0
0.5
0.0
-0.5
-1.0
-1.5
Dim2 (5.3%)
-2
0
2
4
Dim1 (92.5%)
Groups
setosa
versicolor
virginica

The program output result



```
Cluster: 3
Distances: 0.32727432 16.767582 0.09206009
```

Understand the result

- Distance value smaller is better because it close to that cluster
- In this case Cluster = 3 because 0.09 is the nearest cluster

What’s next?