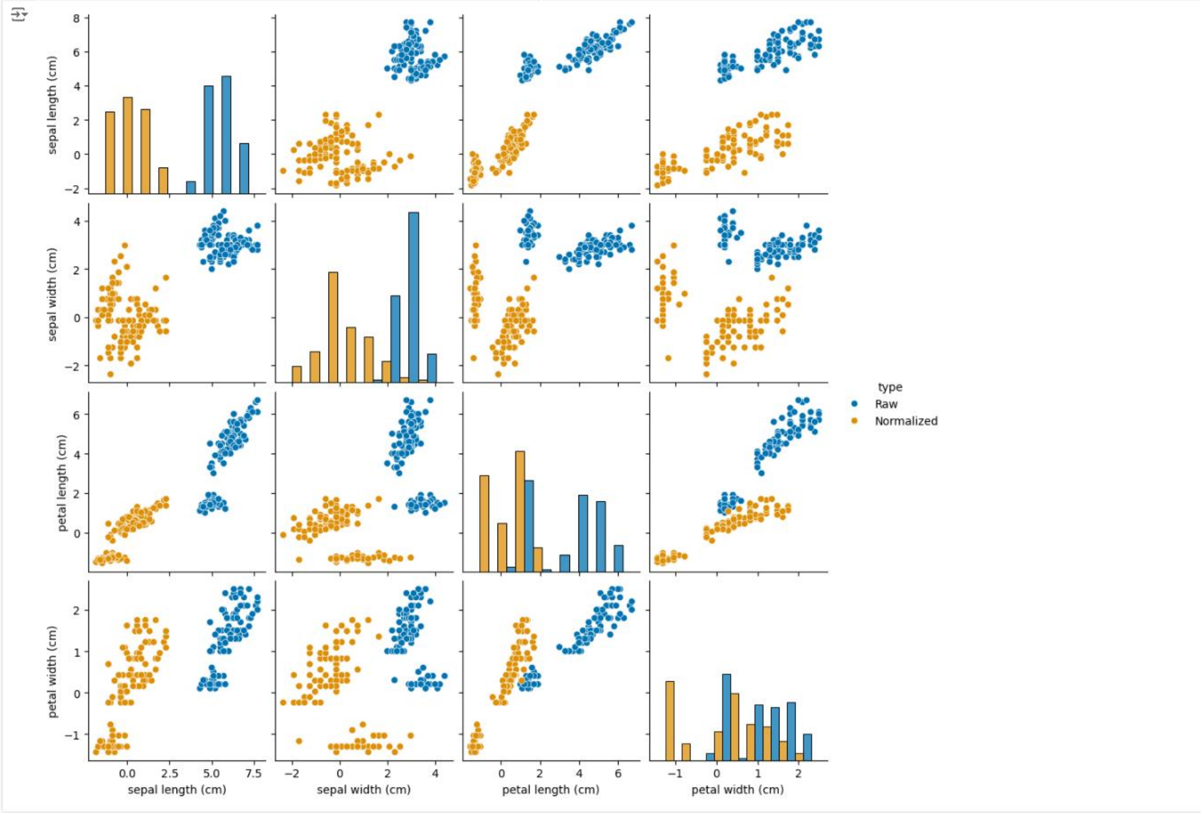
Name: Fahad Iqbal **ID:220708016** **ML Assessment 1**

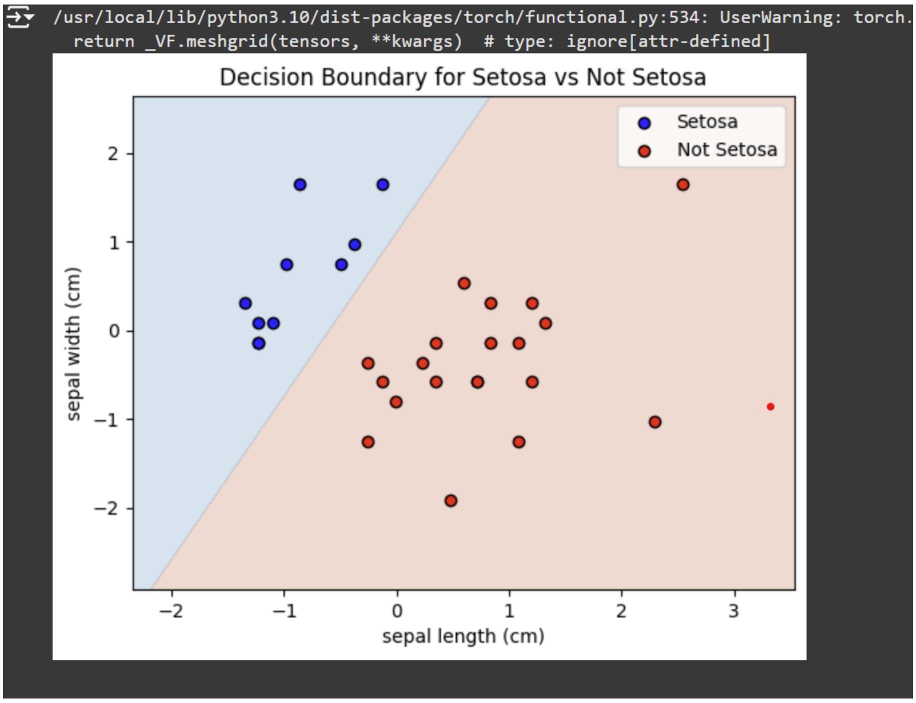
Assignment 1 Part 2(Classification)

Q1. We again notice that the attributes are on different scales. Use the normalisation method from last lab, to standardize the scales of each attribute on both sets. Plot the normalized and raw training sets; what do you observe? [2 marks]



In the figure above we can observe the pairplot of the features from the iris dataset, we have the raw and normalised dataset color coded in blue and orange respectively.The pieces of data have somewhat similar shapes due to maintaing the inherent relationships but are adjusted for scale due to normalization. This can be observed where sepal length and petal length show a positive linear relationship in both the raw and normalized data .Interestingly, the normalized data is centered more arround zero (which is expected due to us normalizing) .The reason for this is because this process shifts the attributes into having a standard deviation of 1 and mean of zero and allowing scaling across all our features.Normalization of our data was important because it prevents larger numeric ranges like sepal and petal features from disproportionately affecting our model.We can view this in the histograms where blue(raw data) are in a data range much larger than its orange counterparts, petal width ranges from 0-25 and sepal length range from 4-8. This can be seen as there are some very small blue bins compared to the larger ones.Normalised data the range for many pieces of the data is from -2 to 2 which shows that its centered arround the mean zero.

Q5. Draw the decision boundary on the test set using the learned parameters. Is this decision boundary separating the classes? Does this match our expectations? [2 marks]

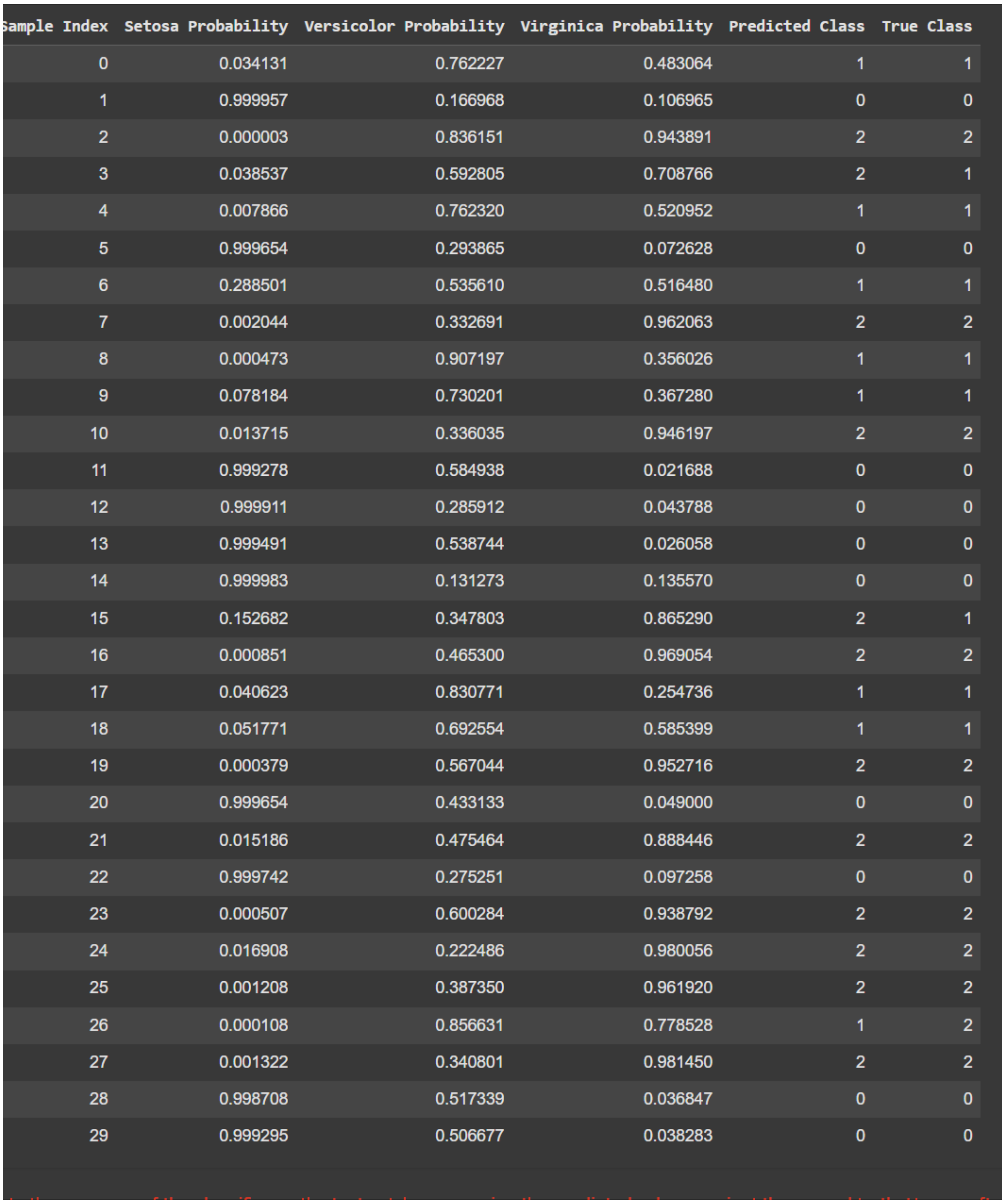


```
/usr/local/lib/python3.10/dist-packages/torch/functional.py:534: UserWarning: torch.
  return _VF.meshgrid(tensors, **kwargs)  # type: ignore[attr-defined]
```

The above figure represents a graph with a decision boundary separating the classes of setosa and not setosa based on logisitic regression parameters .Blue region to the left of the linear boundary is where the model predicts Setosa and the larger red region is where it predicts not setosa,The decision boundary is linear which is what we expect as we want a linear seperation between the class setosa and not setosa. The Setosa samples are red and not setosa samples are blue , they are both in their correct regions as we see no crossover with blue in red region or vice versa indicating our boundary is correct. This matches with my expectations as previously setosa was linearly separable from the other classes in terms of the features sepal length and sepal width and we can see this in the figure.

Q7. Using the 3 classifiers, predict the classes of the samples in the test set and show the predictions in a table. Do you observe anything interesting? [4 marks]

Name: Fahad Iqbal **ID:220708016** **ML Assessment 1**



Sample Index	Setosa Probability	Versicolor Probability	Virginica Probability	Predicted Class	True Class
0	0.034131	0.762227	0.483064	1	1
1	0.999957	0.166968	0.106965	0	0
2	0.000003	0.836151	0.943891	2	2
3	0.038537	0.592805	0.708766	2	1
4	0.007866	0.762320	0.520952	1	1
5	0.999654	0.293865	0.072628	0	0
6	0.288501	0.535610	0.516480	1	1
7	0.002044	0.332691	0.962063	2	2
8	0.000473	0.907197	0.356026	1	1
9	0.078184	0.730201	0.367280	1	1
10	0.013715	0.336035	0.946197	2	2
11	0.999278	0.584938	0.021688	0	0
12	0.999911	0.285912	0.043788	0	0
13	0.999491	0.538744	0.026058	0	0
14	0.999983	0.131273	0.135570	0	0
15	0.152682	0.347803	0.865290	2	1
16	0.000851	0.465300	0.969054	2	2
17	0.040623	0.830771	0.254736	1	1
18	0.051771	0.692554	0.585399	1	1
19	0.000379	0.567044	0.952716	2	2
20	0.999654	0.433133	0.049000	0	0
21	0.015186	0.475464	0.888446	2	2
22	0.999742	0.275251	0.097258	0	0
23	0.000507	0.600284	0.938792	2	2
24	0.016908	0.222486	0.980056	2	2
25	0.001208	0.387350	0.961920	2	2
26	0.000108	0.856631	0.778528	1	2
27	0.001322	0.340801	0.981450	2	2
28	0.998708	0.517339	0.036847	0	0
29	0.999295	0.506677	0.038283	0	0

Analyzing the predictions table from Q7, here are some interesting observations:

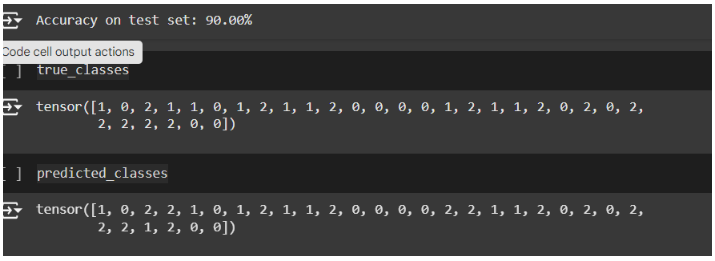
Looking at the table above we can observe that the true class and predictions class match each other consistently which means that the models performing well on the test set.We can see the models is very consistent in this regard, looking at sample index 1 the setosa probability is 0.999957 which matches the true class compared to the other 2 classes.

Speaking on seperability which was an aspect we discussed previously , class 0/setosa where it is correctly classified has especially high probabilities such as sample index 0 or sample index 1. These values were much higher than the other 2 classes on their rows which is a good indication that Setosa is distinctly separable from the other two classes.

- In sample 1 and 2 we can observe Versicolor and Virginica classes (classes 1 and 2) occasionally have closer probabilities compared to Setosa, this shows that sometimes the classes values are ovarlapping and are difficult to seperate.

The reason our model was mostly confident and chose a correct class was the use of the torch.argmax method which picks the highest probability of the three classes from from our classifier . The only downside is if there is a case where classes have similar values for probability.

Q8) Calculate the accuracy of the classifier on the test set, by comparing the predicted values against the ground truth. Use a softmax for the classifier outputs. [1 mark]



```
Accuracy on test set: 90.00%
```

```
true_classes
```

```
tensor([1, 0, 2, 1, 1, 0, 1, 2, 1, 1, 2, 0, 0, 0, 0, 1, 2, 1, 1, 2, 0, 2, 0, 2,
        2, 2, 2, 2, 0, 0])
```

```
predicted_classes
```

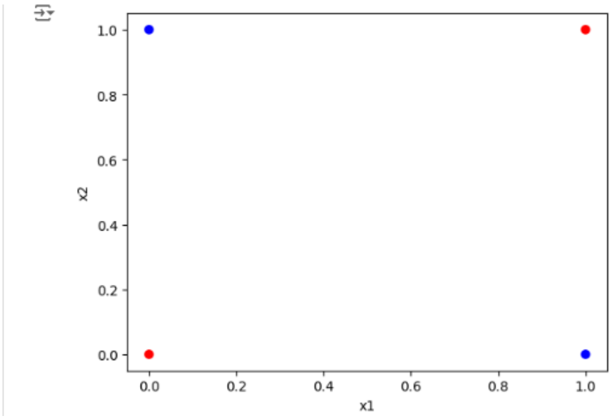
```
tensor([1, 0, 2, 2, 1, 0, 1, 2, 1, 1, 2, 0, 0, 0, 0, 2, 2, 1, 1, 2, 0, 2, 0, 2,
        2, 2, 1, 2, 0, 0])
```

In order to calculate accuracy on the test set we used a softmax function for multiclass classifcations and then subsequently used an argmax function to select the class with the highest probabilityfor each sample which was the final prediction.

We can view above in the figure that our model has an accuracy of 90.00% indicating that the model is working quite well . This is a good value for our logisitic regression model and shows that the "1 vs all approach" we used is efficient at distinguishing between the 3 classes.

The 90 percent accuracy is further proven by the comparison of the "true classes" tensor to "predicted classes" .The reason that some of the predictions don't match the ground truth can be attributed to the overlap and the difficulty in seperation of virginica and versicolor.

Q9)



The figure above of the XOR shows why a decision boundary cannot be drawn for logistic regression. The positioning of the points in each corner are such that its not possible for a linear decision boundary to sepearate all four of these classes. This is what is meant by linearly inseperable data as its impossible for a “straight line” to separate these four classes. This is a major problem for logistic regression as it is a linear classifier which relies on finding straight line(hyperplane for higher dimensions) and is not equipped for non linear problems such as this.

This specific problem cannot be solved using Logistic Regression but a non-linear model, such as a neural network with hidden layers could solve the problem.