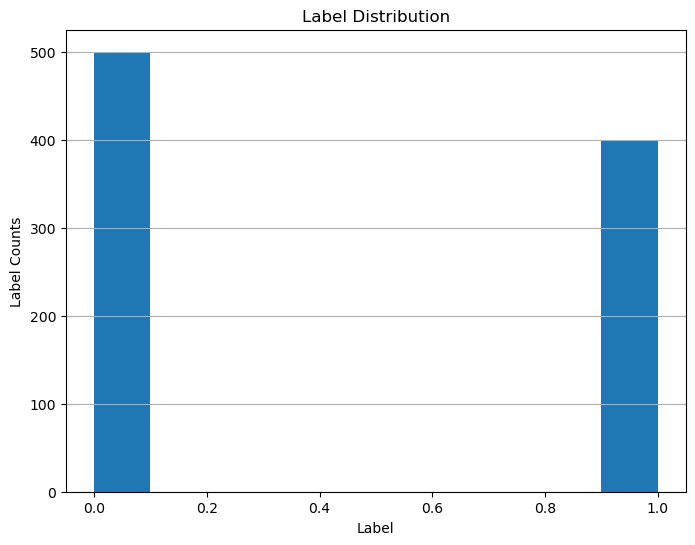
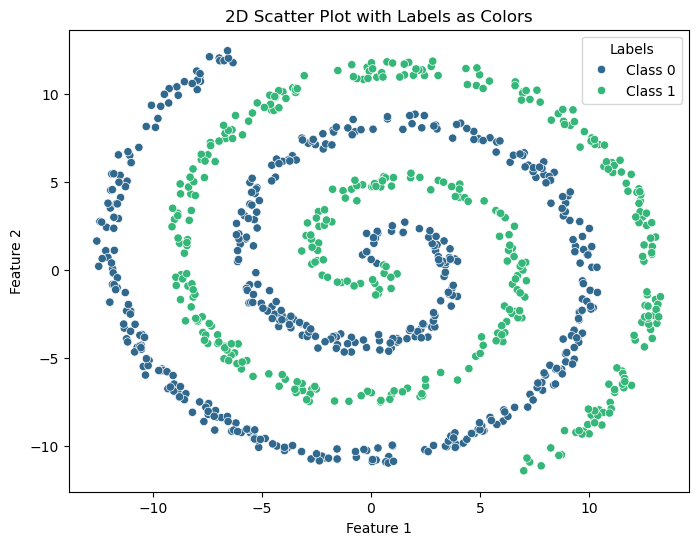
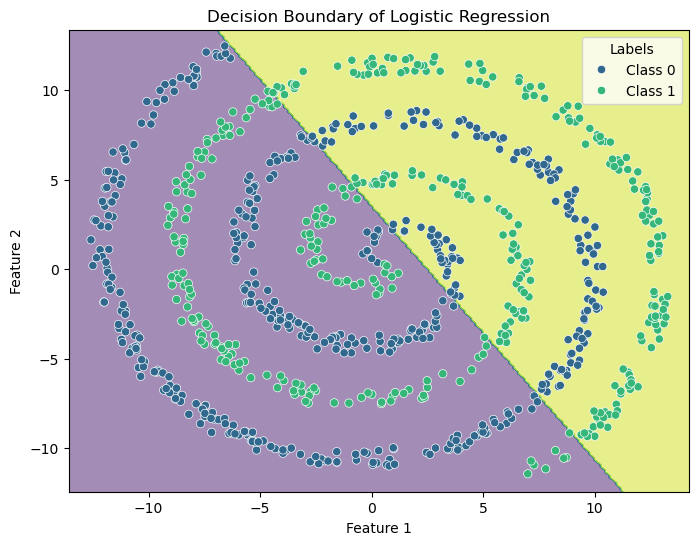
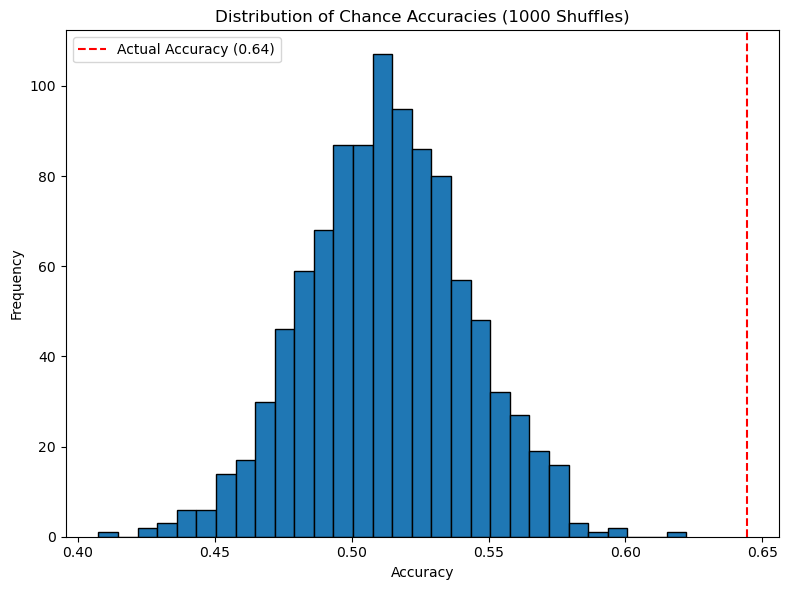
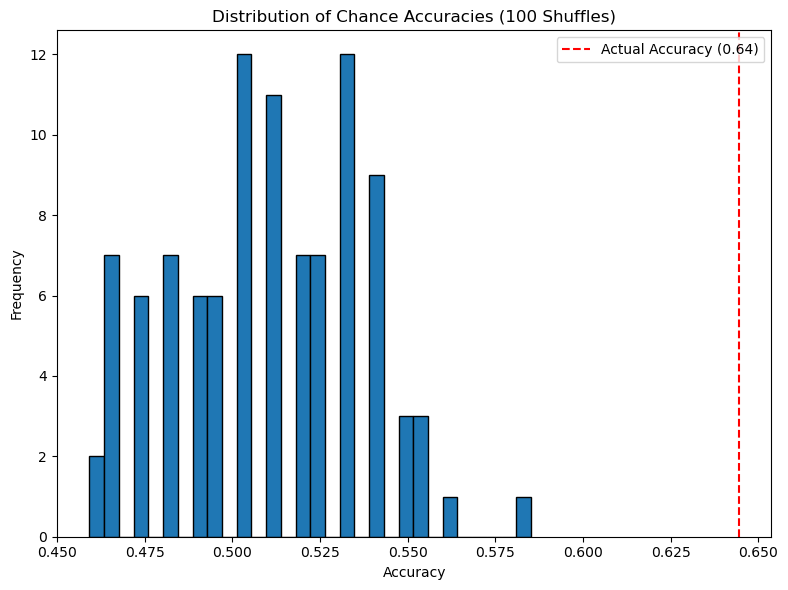
**Report Foundation of Data Analysis**

*Part I*

1. **Understanding the Data**
2. There are 900 samples and 2 features.
3. binary classification, because there are just two unique labels (0 and 1)
4. ****
5. **Data visualisation**
7. No, it’s not. The data is not linearly separable at first glance, because plot of the data is obviously spiral-shaped and it’s not possible here to separate it by a straight line.
8. No, because again the shape of the data is spiral-shaped, since linear model tries always to separate data by a straight line, in our case here it’s not possible. So, for this reason it’s not possible here to achieve zero empirical risk.
9. **Fitting the first classifier**
10. The split is necessary, because we want to make sure the model can handle the new data, and not just the data it was already trained with.
11. I’ve chosen **Logistic Regression,** as it’s the best algorithm used for binary classifications. So, exactly what we need for our example data.
12. “Logistic Regression model accuracy: 64.44%”.

In our case, where we’ve got 900 samples and as we can see from histogram, we have 500 samples for class 0 and 400 samples for class 1. This means for us, that always predicting class 0 gives us 55,56%. Since our Logistic Regression is 64.44%, it’s better than the baseline, but it’s still not perfect.

1. ****
2. **Does the model perform better than chance?**
3. Accuracy after shuffling the data one time: 51.85%
4. As we see from histogram, the most chance accuracy lies within 51-54% and an actual accuracy is 64.44%, which is obviously higher than any chance accuracy.
5. This plot suggests us, that our model’s actual accuracy is much better than what we’ve got after 1000 shuffles. So, none of these random guesses could exceed or at least reach the actual accuracy, which means, our model’s accuracy hasn’t occurred by chance.
6. As was suggested at the task description, I repeated the shuffling 1000 times, because it provides more stable estimation of the chance distribution, and a fewer number of shuffling provides less trusted output. For comparison, see histogram below, where we obviously can see many free spaces between each bar chart.



1. The most frequent class accuracy: 58.15%

No, it’s not the same as chance accuracy, because baseline always picks the majority class, and chance accuracy is based on random guesses. So, baseline (dummy classifier) compares the performance of more sophisticated models.

1. After testing chance accuracy with both 100 and 1000 shuffles, and comparing it to a baseline classifier, we clearly can see, that none of them came close to the model’s actual accuracy of 64.44%. It means for us, the performance of our model isn’t that perfect, as we could wish, but still it learned something meaningful from the data. It’s clearly better than random guessing or dummy classifier.
2. **Neural nets**
3. Created a neural network with 2 input layers, 20 hidden layers + ReLU, 1 output layer, “BCEWithLogitsLoss” loss, Adam optimizer with 0.01 learning rate, trained over 100 epochs.

Source:

1. [numpy.histogram — NumPy v2.2 Manual](https://numpy.org/doc/2.2/reference/generated/numpy.histogram.html)
2. [pandas.Series.map — pandas 2.2.3 documentation](https://pandas.pydata.org/docs/reference/api/pandas.Series.map.html)
3. [numpy.load — NumPy v2.2 Manual](https://numpy.org/doc/2.2/reference/generated/numpy.load.html)
4. [Histograms in Python](https://plotly.com/python/histograms/)
5. [Data Analysis with Python: Zero to Pandas | Jovian](https://jovian.com/learn/data-analysis-with-python-zero-to-pandas#C49)
6. [seaborn.scatterplot — seaborn 0.13.2 documentation](https://seaborn.pydata.org/generated/seaborn.scatterplot.html)
7. [Choosing color palettes — seaborn 0.13.2 documentation](https://seaborn.pydata.org/tutorial/color_palettes.html)
8. [matplotlib.pyplot.legend — Matplotlib 3.10.3 documentation](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.legend.html)
9. [Logistic Regression in Machine Learning | GeeksforGeeks](https://www.geeksforgeeks.org/understanding-logistic-regression/)
10. [Plot Decision Boundary in Logistic Regression: Python Example](https://vitalflux.com/plot-decision-boundary-using-logistic-regression-python-example/)
11. [How to Shuffle a Pandas DataFrame? | by Hey Amit | Medium](https://medium.com/@heyamit10/how-to-shuffle-a-pandas-dataframe-60678a3fb94b)
12. [python - What does numpy.random.seed(0) do? - Stack Overflow](https://stackoverflow.com/questions/21494489/what-does-numpy-random-seed0-do)
13. [python - Randomly shuffle items in each row of numpy array - Stack Overflow](https://stackoverflow.com/questions/50554272/randomly-shuffle-items-in-each-row-of-numpy-array)
14. [A Dummy Classifier, A Baseline Classifier, or a Null Model | by Sahel Eskandar | Medium](https://medium.com/@eskandar.sahel/a-dummy-classifier-a-baseline-classifier-or-a-null-model-71df50fd8947)