Shivam Patel – Machine Learning Project Part 1 (Proposal)

Course: CSCI 5371 – Machine Learning
Instructor: Dr. Zachary Stine (Assistant Professor)
Project Title: Titanic Survival Prediction
Dataset Source: https://www.tensorflow.org/datasets/catalog/titanic

1. Name of the dataset and link

- Dataset: Titanic passenger survival dataset
  Link: https://www.tensorflow.org/datasets/catalog/titanic

2. What do observations in this dataset represent?

- Each observation represents about one passenger abroad the Titanic – including their personal details, ticket class, and survival status.

3. How many observations are there?

- There is total 1309 observation.

4. What do attributes in this dataset represent?

- Each attribute describes characteristics of a passenger. Examples include:
  - pclass: Ticket class (1^{st}, 2^{nd}, 3^{rd})
  - sex: Gender
  - age: Age in years
  - sibsp: Number of siblings/spouses aboard
  - parch: number of parents/children aboard
  - fare: Ticket fare
  - embarked: Port of embarkation (C, Q, or S)
  - survived: whether the passenger survived (1) or not (0)

5. What is a reasonable target attribute for the dataset?

- The target attribute is “survived”, which indicates whether a passenger survived or not.

6. Given the target identified, is it regression or classification?

- This is a binary classification problem.

7. How would you use validation to choose between two modeling setups?

- I would split the dataset into training, validation, and test sets (for example, 70%/15%/15%). Then I would train two different models and evaluate them using validation accuracy, precision, recall, or F1-score

8. If using an artificial neural network, what would the input layer consist of and how many nodes would it have?

- The layer would include all the numerical and encoded categorical features. After one-hot encoding, there would be approx. 10-12 input nodes, representing attributes such as class, age, gender, fare, and embarkation port.

9. What would the output layer look like and how many nodes would it have?

- The output layer would have 1 node with a sigmoid activation function to predict the probability of survival.

10. What modeling decisions would you be interested in comparing in the final project?

- I want to compare different models like Logistic Regression, Decision Tree, and Neural Network. I will also try different learning rates, activation functions, and data preprocessing methods to see which one gives better accuracy.