Question#1

Consider a bird classification task. Suppose your chosen algorithm performs as follows:

- Training error = 2%
- Validation error = 12%

1. 1 Identify whether your algorithm suffers from a high bias (underfitting) or high variance (overfitting) problem.

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Variance is the variability of model prediction for a given data point.
Because in this task, the model prediction error on validation dataset (12%) is much higher than the error on training dataset (2%), which means the model prediction has high variance and it has **overfitting** problem.

1. 2 Suggest methods to address the same.

There are some methods to limit the overfitting problem

1. Train with more data
   Training with more data can help algorithms get better result on more dataset, which can reduce the overfitting problem on many cases

2. Reduce features
   Try to improve the generalizability of the model by removing irrelevant input features. Normally the more generalized model has less variance.

3. Regularization
   Regularization means using different methods to make the model simpler. The methods will depend on the model. E.g. Use lower order hypermeters in the polynomial fit.

Question#2

Assume that there is a total of 80 Machine learning textbooks in a library of 1000 textbooks. Suppose that a search engine retrieves 10 textbooks after a user enters 'Machine Learning' as a query, of which 2 are not Machine Learning textbooks. What is the precision and recall of the search engine model?

Below is the Confusion Matrix for Imbalanced Classification

| | Positive Prediction | Negative Prediction |
|---|---|---|
| Positive Class | True Positive (TP) | False Negative (FN) |
| Negative Class | False Positive (FP) | True Negative (TN) |

Precision = True Positives / (True Positives + False Positives)

Recall = True Positives / (True Positives + False Negatives)

In this case, "**True Positives**" is the number the search engine gives the correct classification in the result, which is **8**. **False Positives** is the number the search engine give the incorrect classification in the result, which is **2**. **False Negative** is the number of the rest of ML textbooks that the search engine does not put into the result, which is **72**

So,

**precision = 8 / (8+2) = 80%**

**recall = 8 / (8+72) = 10%**