cUNIT 3 ASSIGNMENT

Understanding the Mechanics of   
ML Algorithms

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the unit. You’ve practiced these concepts in the coding activities, exercises, and coding portion of the assignment; let’s now formulate your programming into well-reasoned responses.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator   
as part of your assignment completion. You may also have additional supporting documents that   
you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or   
shared outside the course, you should take care to obscure any information you feel might be   
of a sensitive or confidential nature.

*Begin your assignment by completing the questions below. Directions to submit your work can be found on the assignment page. Information about the grading rubric is available on any of the course assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

Unit 3 Written Portion

# Building and Evaluating a Model

Answer the questions below about building and evaluating your models using algorithms such as decision trees and k-nearest neighbors.

## Questions:

1. What are the advantages and disadvantages of decision trees?

|  |
| --- |
| Advantages: Compared to other machine learning algorithms, decision trees are easily interpretable by humans as long as the tree isn’t too big. Additionally, they are computationally efficient. They can also handle both numerical and categorical data.  Disadvantages: Decision trees can become overly complex if the proper hyperparameters are not specified. Decision trees are easily susceptible to variations in data; thus, they are quite unstable. |

1. What are the advantages and disadvantages of k-nearest neighbors?

|  |
| --- |
| Advantages: K-nearest neighbors does not require a training period. It learns from the data at the time of prediction. This results in two main benefits: 1) it is quite fast compared to other machine learning algorithms and 2) new data can be added at any point.  Disadvantages: The k-nearest neighbors algorithm relies on a distance metric, so it requires feature scaling to ensure that each feature is weighted equally. Additionally, it is difficult to calculate distance in high dimensions. KNN is also susceptible to outliers. |

1. Explain the difference between k-nearest neighbors and decision trees. When would you decide to use one over the other?

|  |
| --- |
| If I have categorical data, I would use decision trees. When I have numerical data, I can choose either KNN or decision trees. If there are too many dimensions in my data, I would probably use decision trees. |

1. What are hyperparameters? List some hyperparameters in k-nearest neighbors and decision trees.

|  |
| --- |
| Hyperparameters are parameters that we can tune in a machine learning algorithm to achieve the desired learning result.  Decision tree hyperparameters: entropy, maximum tree depth, minimum samples per leaf  K-nearest neighbors: number of neighbors, distance metric |

1. What is overfitting? How can you avoid overfitting? Give examples using a model discussed   
   so far.

|  |
| --- |
| Overfitting is the occurrence of a machine learning model learning too strictly from its training dataset, making it hard to generalize the model on other datasets. Overfitting can be avoided by tuning the hyperparameters accordingly and testing on validation data. You can avoid overfitting a decision tree model by training the model with a list of different maximum depths on a training dataset and then testing the performance of each of them on a validation dataset. Pick the maximum depth with the best performance on the validation set for your model. |

1. What is the purpose of splitting data into different sets?

|  |
| --- |
| Splitting data into different sets helps us avoid overfitting and underfitting and helps us gauge how well our model is performing on datasets besides our training dataset. |

*To submit this assignment, please refer to the instructions in the course*.