Person 5. Manager (and more). By Sagar Darnal

a. For classification and regression tasks, kNN (k-Nearest Neighbors) and decision trees are two common machine learning methods.

kNN categorizes new data based on its nearest neighbors from the training set. The algorithm determines the separation between each point in the training group and the new data point. The new data point is categorized based on the most prevalent class among the k nearest neighbors, who are chosen.
Decision trees, on the other hand, operate by breaking the data up into smaller subsets based on a set of if-then rules. Each branch of the tree reflects the results of tests conducted on each internal node, which each represents a test conducted on a particular feature. The tree's branches stand in for the ultimate classification or prediction.

The outcome from Regression:
All three algorithms performed poorly on the given dataset. Linear regression was ineffective due to the simplicity of the data. The kNN model had the highest correlation but struggled due to the challenge of finding similar data points. Decision trees performed the worst because of the lack of patterns or correlations between the input features and the target variable. These algorithms each have unique strategies and are useful in different situations. Careful consideration in data set selection is necessary for optimal performance.

The outcome from Classification:
It appears that no computer could accurately predict if an aircraft will be delayed. If this were easy to forecast using the data, as we would expect it to be, the airlines would be aware of it and would have issues to address. Most likely, a weather field would include the data required to obtain an accuracy greater than roughly 60%, but that wasn't provided in the data. Therefore, as these algorithms cycled through the data, logistic regression was unable to establish a linear classification separator, and decision trees would have encountered the same problem. The failure of kNN to find a trend reflects both the data's relative lack of one and the odd way its inner workings interact with factors, as they have to be arbitrarily ordered and kNN assigns their values a mathematical meaning where there is none. Both the relative lack of a pattern in the data and the peculiar manner in that kNN's internal workings interact with factors—which must be arbitrarily sorted and have their values given mathematical meaning when there is none—are reflected in the method's failure to detect a trend.

b. K-means, hierarchical clustering, and density-based clustering are the three clustering techniques used in phase 3.

The iterative algorithm k-means clustering divides the data into k groups. Each data point is given the closest centroid by arbitrarily choosing k centroids. As the process progresses toward convergence, the centroids are updated to reflect the average of the points in each of their individual clusters.

To create a single cluster or a predetermined number of clusters, hierarchical clustering repeatedly merges the closest clusters. You can approach this from the top down or the bottom up. All the data points begin in the same cluster in top-down clustering, and the algorithm recursively breaks the data down into smaller clusters until the desired number is attained. Each data point begins in its own cluster in bottom-up clustering, and the algorithm merges the closest cluster pairs until the desired number is attained.
When using density-based clustering, high-density areas in the data are found and each location is assigned to a cluster based on its density. The algorithm begins by randomly picking a spot and locating all its neighbors within a predetermined radius. When no more points can be added, it expands the cluster by adding nearby points that satisfy certain density requirements.

The outcomes from our test:
The three methods of clustering mentioned in our test are kMeans, Hierarchical, and Model Based. Although effective for examining data that naturally has a hierarchy since it allows for the presentation of all cluster branches, hierarchical clustering wasn't a good fit for the dataset. This occurs because of the lack of a hierarchical structure in the data used. The dataset was successfully clustered by kMeans, which was able to cluster the data instances substantially better than random assignment. Model-based clustering, which may group data according to flexible frameworks, is perhaps the most reliable. An ellipsoidal, equal orientation clustering was determined to be the best fit for the data by the algorithm.

c. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are techniques used for dimensionality reduction in machine learning.

In order for PCA to function, the data must be projected onto the dimensions in which the data vary most. By doing this, the number of data variables is decreased while most of the crucial details are kept. PCA is helpful for both visualizing high-dimensional data and lowering the computational burden of machine learning models.
LDA is a supervised dimensionality reduction method that reduces the number of classes by identifying the directions that optimize class separation. As a result, LDA searches for the projection that best distinguishes the data points according to their class names. By eliminating duplicate or irrelevant features, LDA is helpful for feature extraction and can enhance the performance of machine learning models.
In conclusion, the most well-known machine learning algorithms for classification and regression problems are kNN and decision trees. K-means, hierarchical clustering, and density-based clustering are the three clustering techniques employed in phase 3. Dimensionality reduction methods used in machine learning, such as PCA and LDA, can enhance the effectiveness and performance of machine learning models.

The outcome of the test:
It appears that this has improved phase 1's accuracy. We are not sure why any of the improved accuracies worked.