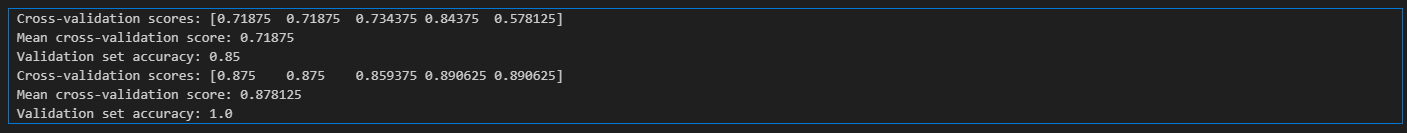
Diego Armando Sarmiento (301379640)

***Analysis report***



Evaluating each of the models, we can observe that KNeighborsClassifier has lower accuracy than the model using RandomForestClassifier, however when reviewing the accuracy, we can observe that the second one identified all the validations correctly, this could give us an indication that there is an overfitting in the training of the models, possibly due to the number of clusters used, in this case 5.

*Agglomerative hierarchical clustering*

***Euclidean distance***

A diagram of a cluster of data

Description automatically generated with medium confidence

***Manhattan distance***

A diagram of a cluster of labels

Description automatically generated

***Cosine distance***

A diagram of a cluster of labels

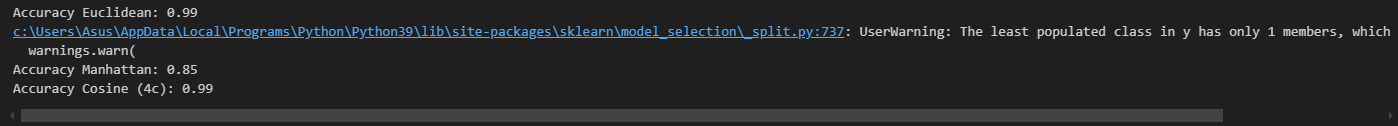
Description automatically generated

The three dendrograms—Euclidean, Manhattan, and Cosine—showcase various clustering patterns according to the selected distance metrics. A wider degree of dissimilarity between the data points is shown by the more varied and distinct cluster formations seen in the Euclidean dendrogram. The Manhattan distance results in greater overall distances, indicating wider separations between clusters, with closer groups in the lower levels and more substantial gaps at the top. In contrast, the Cosine dendrogram shows more compact and clearly defined clusters with smaller overall distances, indicating a higher degree of similarity among the data points.

A graph with different colored lines

Description automatically generated

The optimal number of clusters appears to be 2, as all three distance metrics (Euclidean, Manhattan, and Cosine) show the highest Silhouette Score at this point. The scores generally decrease as the number of clusters increases, indicating that the data is best separated into two distinct groups. For this reason, we could say that a good approach could be between 2 – 4 clusters.



The Euclidean and Cosine distance metrics perform equally well with 99% accuracy, while the Manhattan distance lags at 85%. This suggests that for this dataset and model, Euclidean and Cosine distances are more effective measures of similarity or dissimilarity between data points. However, we could be getting an over-fit model when using Euclidean or Cosine.