Camarines Sur Polytechnic Colleges
College of Computer Studies
Nabua, Camarines Sur

SEAN XANDER B. AQUINO | BSCS-2A

The program accesses wine quality data from the UCI Machine Learning Repository under ID: 186. The dataset includes comprehensive details regarding physical as well as sensory measurements about both white and red versions of Portuguese wine named "Vinho Verde." A total of 11 numerical input characteristics comprise the dataset which represents various chemical attributes such as:

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

The "alcohol" target column contains integer ratings from 3 to 9 that represent human reviewers' wine quality evaluation. The predefined structure and smooth organization of this dataset allow users to conduct supervised and unsupervised learning through classification, regression and anomaly detection methods as shown in this program.

The unsupervised anomaly detection pipeline in the program seeks to find outliers of selected features ("alcohol" or "residual sugar") through Gaussian (Normal) distribution fitting followed by Probability Density Function (PDF) value analysis.

- The following details illustrate how everything unfolded according to the results:
- The selected feature received an applied Gaussian fit that served to generate its distribution model.
- The PDF calculations were performed on each data point through the use of the fitted Gaussian distribution.
- The KMeans clustering algorithm with a default cluster count of two divided the points according to their PDF measurement values.

- The set of points with the minimum average PDF received the anomaly flag because it represented the least likely distribution under the Gaussian assumption.
- Multiple visualizations of anomalies included scatter plots and box plots and PDF histograms that illustrated their distribution as well as separation.
- Several silhouette scores spanning from 2 to 150 cluster counts validated that the meaningful yet non-random anomaly cluster structure existed.

The output dataframe contains an "Anomaly" column which marks anomalies through True values throughout corresponding records. The anomalies related to alcohol measurements would consist of wines whose alcohol content significantly deviates from the typical Gaussian distribution.