**Assignment 4**

**Spark Machine Learning**

DINH HOANG VIET PHUONG – 301123263

1. **Exercise 1: Supervised Learning Decision Trees**
2. **Research and investigate the LIBSVM format, in your analysis report define the format and show  
   an example with explanation.**

LIBSVM is a file format used primarily for Support Vector Machine (SVM) models. The format is favored for its simplicity and efficiency, particularly in handling sparse datasets.

Definition of LIBSVM Format

* Each Line Represents One Data Point: In LIBSVM, each line in the file corresponds to one training example or data point.
* Format of a Line: A line in a LIBSVM file has the following structure:
* Code:

<label> <index1>:<value1> <index2>:<value2> ...

* <label>: This is the target value for the data point. In classification tasks, this is the class label. In regression tasks, it's the value to be predicted.
* <indexN>: This is an integer representing the feature index. Indices must be in ascending order.
* <valueN>: This is the feature value corresponding to <indexN>.
* Sparse Data Representation: One of the key strengths of the LIBSVM format is its ability to efficiently represent sparse data. If a feature has a value of 0, it can be omitted from the line.

Example with Explanation

Consider a dataset with three features (e.g., age, income, and years of education) for a binary classification task (e.g., loan approval with labels 1 for approved and 0 for not approved).

Suppose we have a data point where:

* The loan is approved (label = 1).
* The person is 30 years old (feature 1).
* The income is $50,000 (feature 2).
* Years of education is not available (or feature 3 is 0).

In LIBSVM format, this data point would be represented as:

1 1:30 2:50000

Explanation:

* 1: This is the label indicating loan approval.
* 1:30: The first feature (age) is 30.
* 2:50000: The second feature (income) is 50,000.
* The third feature (years of education) is not mentioned because it's either not available or its value is 0, showcasing the sparse data handling capability.
* Conclusion
* The LIBSVM format is widely used for its efficiency in handling sparse datasets and its simplicity, making it easy to process and use in various machine learning frameworks, especially for SVM models.

1. **Load the data stored in the file “sample\_libsvm\_data.txt” from the data available on the VMware image under the directory /home/centos/data/ into a data frame and name it df\_x where x is your firstname.**

A screen shot of a computer program

Description automatically generated

+-----+--------------------+

|label| features|

+-----+--------------------+

| 0.0|(692,[127,128,129...|

| 1.0|(692,[158,159,160...|

| 1.0|(692,[124,125,126...|

| 1.0|(692,[152,153,154...|

| 1.0|(692,[151,152,153...|

| 0.0|(692,[129,130,131...|

| 1.0|(692,[158,159,160...|

| 1.0|(692,[99,100,101,...|

| 0.0|(692,[154,155,156...|

| 0.0|(692,[127,128,129...|

| 1.0|(692,[154,155,156...|

| 0.0|(692,[153,154,155...|

| 0.0|(692,[151,152,153...|

| 1.0|(692,[129,130,131...|

| 0.0|(692,[154,155,156...|

| 1.0|(692,[150,151,152...|

| 0.0|(692,[124,125,126...|

| 0.0|(692,[152,153,154...|

| 1.0|(692,[97,98,99,12...|

| 1.0|(692,[124,125,126...|

+-----+--------------------+

only showing top 20 rows

1. **Carry out some basic investigation: count the number of records, count the number of columns print the inferred schema.**

A screen shot of a computer code

Description automatically generated

Number of Records: 100

Number of Columns: 2

Inferred Schema:

root

|-- label: double (nullable = true)

|-- features: vector (nullable = true)

Data Frame df\_phuong contains two columns:

label: This column is of type double and appears to contain numeric labels for each record. This column typically represents the target variable or the class label for each record. The values in this column are used to identify the category or numerical value that each record is supposed to represent. For instance, in a binary classification problem, these labels could be 0 or 1, indicating two different classes.

features: This column is described as a vector and is likely a dense or sparse vector containing the feature data for each record. In LIBSVM format, features are often stored in a sparse format where each entry in the vector represents a feature and its value. The vector's size corresponds to the number of features in the dataset. Each record's feature vector provides the input data that is used by machine learning models for training or prediction. The notation (692,[127,128,129... in your data sample suggests that there are 692 features, and the values in the square brackets indicate the indices of the non-zero features in the sparse vector.

1. **Use the StringIndexer to index labels, in other words you will add metadata to the label column. Name the output column "indexedLabel\_phuong”. Store the result in a variable named labelIndexer\_phuong**

A screenshot of a computer code

Description automatically generated

1. **Use the VectorIndexer to automatically identify categorical features and index them. Set the maxCategories to 4. Name the output column " indexedFeatures\_phuong. Store the result in a variable named featureIndexer\_phuong.**

A screen shot of a computer code

Description automatically generated

+-----+--------------------+----------------------+

|label| features|indexedFeatures\_phuong|

+-----+--------------------+----------------------+

| 0.0|(692,[127,128,129...| (692,[127,128,129...|

| 1.0|(692,[158,159,160...| (692,[158,159,160...|

| 1.0|(692,[124,125,126...| (692,[124,125,126...|

| 1.0|(692,[152,153,154...| (692,[152,153,154...|

| 1.0|(692,[151,152,153...| (692,[151,152,153...|

| 0.0|(692,[129,130,131...| (692,[129,130,131...|

| 1.0|(692,[158,159,160...| (692,[158,159,160...|

| 1.0|(692,[99,100,101,...| (692,[99,100,101,...|

| 0.0|(692,[154,155,156...| (692,[154,155,156...|

| 0.0|(692,[127,128,129...| (692,[127,128,129...|

| 1.0|(692,[154,155,156...| (692,[154,155,156...|

| 0.0|(692,[153,154,155...| (692,[153,154,155...|

| 0.0|(692,[151,152,153...| (692,[151,152,153...|

| 1.0|(692,[129,130,131...| (692,[129,130,131...|

| 0.0|(692,[154,155,156...| (692,[154,155,156...|

| 1.0|(692,[150,151,152...| (692,[150,151,152...|

| 0.0|(692,[124,125,126...| (692,[124,125,126...|

| 0.0|(692,[152,153,154...| (692,[152,153,154...|

| 1.0|(692,[97,98,99,12...| (692,[97,98,99,12...|

| 1.0|(692,[124,125,126...| (692,[124,125,126...|

+-----+--------------------+----------------------+

only showing top 20 rows

1. **Printout the following:**

A screenshot of a computer program

Description automatically generated

Name of Input Column: features

Name of Output Column: indexedFeatures\_phuong

Number of Features: 692

Map of Categories:

{645: {0.0: 0}, 69: {0.0: 0}, 365: {0.0: 0}, 138: {0.0: 0}, 479: {0.0: 0}, 333: {0.0: 0}, 249: {0.0: 0}, 0: {0.0: 0}, 666: {0.0: 0, 10.0: 1}, 88: {0.0: 0}, 170: {0.0: 0}, 115: {0.0: 0}, 276: {0.0: 0, 3.0: 1, 153.0: 2, 252.0: 3}, 308: {0.0: 0}, 5: {0.0: 0}, 449: {0.0: 0}, 120: {0.0: 0, 253.0: 1}, 614: {0.0: 0, 140.0: 1}, 677: {0.0: 0}, 202: {0.0: 0, 13.0: 1, 44.0: 2, 87.0: 3}, 10: {0.0: 0}, 56: {0.0: 0}, 533: {0.0: 0}, 142: {0.0: 0}, 340: {0.0: 0}, 670: {0.0: 0}, 174: {0.0: 0, 175.0: 1}, 42: {0.0: 0}, 417: {0.0: 0}, 24: {0.0: 0}, 37: {0.0: 0}, 25: {0.0: 0}, 257: {0.0: 0, 73.0: 1, 120.0: 2}, 389: {0.0: 0}, 52: {0.0: 0}, 14: {0.0: 0}, 504: {0.0: 0}, 110: {0.0: 0}, 587: {0.0: 0}, 619: {0.0: 0}, 196: {0.0: 0}, 559: {0.0: 0}, 638: {0.0: 0, 1.0: 1, 29.0: 2, 137.0: 3}, 20: {0.0: 0}, 421: {0.0: 0}, 46: {0.0: 0}, 93: {0.0: 0}, 284: {0.0: 0}, 228: {0.0: 0}, 448: {0.0: 0}, 57: {0.0: 0}, 78: {0.0: 0}, 29: {0.0: 0}, 475: {0.0: 0}, 164: {0.0: 0, 14.0: 1}, 591: {0.0: 0}, 646: {0.0: 0}, 253: {0.0: 0}, 106: {0.0: 0}, 121: {0.0: 0, 63.0: 1, 132.0: 2}, 84: {0.0: 0}, 147: {0.0: 0, 241.0: 1}, 280: {0.0: 0}, 61: {0.0: 0}, 221: {0.0: 0}, 396: {0.0: 0, 19.0: 1}, 89: {0.0: 0}, 133: {0.0: 0, 9.0: 1, 18.0: 2, 52.0: 3}, 116: {0.0: 0}, 1: {0.0: 0}, 507: {0.0: 0}, 312: {0.0: 0}, 74: {0.0: 0}, 307: {0.0: 0}, 452: {0.0: 0, 24.0: 1, 29.0: 2}, 6: {0.0: 0}, 248: {0.0: 0, 13.0: 1, 250.0: 2}, 60: {0.0: 0}, 117: {0.0: 0}, 678: {0.0: 0, 37.0: 1, 40.0: 2}, 529: {0.0: 0}, 85: {0.0: 0}, 201: {0.0: 0}, 220: {0.0: 0, 250.0: 1}, 366: {0.0: 0}, 534: {0.0: 0}, 102: {0.0: 0, 5.0: 1, 72.0: 2}, 334: {0.0: 0}, 28: {0.0: 0}, 38: {0.0: 0}, 561: {0.0: 0}, 392: {0.0: 0}, 70: {0.0: 0}, 424: {0.0: 0, 5.0: 1, 29.0: 2}, 192: {0.0: 0, 146.0: 1}, 21: {0.0: 0}, 137: {0.0: 0}, 165: {0.0: 0}, 33: {0.0: 0}, 92: {0.0: 0}, 229: {0.0: 0, 23.0: 1}, 252: {0.0: 0}, 197: {0.0: 0}, 361: {0.0: 0}, 65: {0.0: 0}, 97: {0.0: 0, 64.0: 1, 121.0: 2}, 665: {0.0: 0, 25.0: 1, 71.0: 2, 173.0: 3}, 224: {0.0: 0}, 615: {0.0: 0}, 9: {0.0: 0}, 53: {0.0: 0}, 169: {0.0: 0}, 141: {0.0: 0}, 420: {0.0: 0}, 109: {0.0: 0}, 256: {0.0: 0}, 225: {0.0: 0}, 339: {0.0: 0}, 77: {0.0: 0}, 193: {0.0: 0}, 669: {0.0: 0}, 476: {0.0: 0}, 642: {0.0: 0}, 590: {0.0: 0}, 679: {0.0: 0, 239.0: 1, 251.0: 2}, 96: {0.0: 0, 247.0: 1}, 393: {0.0: 0}, 647: {0.0: 0}, 173: {0.0: 0}, 13: {0.0: 0}, 41: {0.0: 0}, 503: {0.0: 0}, 134: {0.0: 0}, 73: {0.0: 0}, 105: {0.0: 0}, 2: {0.0: 0}, 311: {0.0: 0}, 558: {0.0: 0}, 674: {0.0: 0}, 530: {0.0: 0}, 586: {0.0: 0}, 618: {0.0: 0}, 166: {0.0: 0}, 32: {0.0: 0}, 34: {0.0: 0}, 148: {0.0: 0, 71.0: 1, 251.0: 2}, 45: {0.0: 0}, 279: {0.0: 0}, 64: {0.0: 0}, 17: {0.0: 0}, 584: {0.0: 0}, 562: {0.0: 0}, 423: {0.0: 0}, 191: {0.0: 0, 250.0: 1}, 22: {0.0: 0}, 44: {0.0: 0}, 59: {0.0: 0}, 118: {0.0: 0}, 281: {0.0: 0}, 27: {0.0: 0}, 641: {0.0: 0}, 71: {0.0: 0}, 391: {0.0: 0}, 12: {0.0: 0}, 445: {0.0: 0}, 54: {0.0: 0}, 611: {0.0: 0, 19.0: 1, 20.0: 2, 29.0: 3}, 144: {0.0: 0}, 49: {0.0: 0}, 335: {0.0: 0}, 86: {0.0: 0}, 672: {0.0: 0}, 172: {0.0: 0}, 113: {0.0: 0}, 219: {0.0: 0, 18.0: 1, 20.0: 2, 250.0: 3}, 419: {0.0: 0}, 81: {0.0: 0}, 362: {0.0: 0}, 451: {0.0: 0}, 76: {0.0: 0}, 7: {0.0: 0}, 39: {0.0: 0}, 649: {0.0: 0, 83.0: 1}, 98: {0.0: 0, 70.0: 1, 191.0: 2}, 616: {0.0: 0}, 477: {0.0: 0}, 367: {0.0: 0}, 535: {0.0: 0}, 103: {0.0: 0}, 140: {0.0: 0}, 621: {0.0: 0, 82.0: 1, 236.0: 2}, 91: {0.0: 0}, 66: {0.0: 0}, 251: {0.0: 0}, 668: {0.0: 0}, 198: {0.0: 0}, 108: {0.0: 0}, 278: {0.0: 0}, 223: {0.0: 0}, 394: {0.0: 0}, 306: {0.0: 0}, 135: {0.0: 0}, 563: {0.0: 0}, 226: {0.0: 0}, 3: {0.0: 0}, 505: {0.0: 0}, 80: {0.0: 0}, 167: {0.0: 0}, 35: {0.0: 0}, 473: {0.0: 0}, 675: {0.0: 0}, 589: {0.0: 0}, 531: {0.0: 0}, 255: {0.0: 0}, 648: {0.0: 0}, 112: {0.0: 0}, 617: {0.0: 0}, 194: {0.0: 0}, 145: {0.0: 0}, 48: {0.0: 0}, 557: {0.0: 0}, 63: {0.0: 0}, 640: {0.0: 0}, 18: {0.0: 0}, 282: {0.0: 0}, 95: {0.0: 0, 56.0: 1}, 310: {0.0: 0}, 50: {0.0: 0}, 67: {0.0: 0}, 199: {0.0: 0}, 673: {0.0: 0}, 16: {0.0: 0}, 585: {0.0: 0}, 502: {0.0: 0}, 338: {0.0: 0}, 643: {0.0: 0}, 31: {0.0: 0}, 336: {0.0: 0}, 613: {0.0: 0}, 11: {0.0: 0}, 72: {0.0: 0}, 446: {0.0: 0}, 612: {0.0: 0}, 143: {0.0: 0}, 43: {0.0: 0}, 250: {0.0: 0}, 450: {0.0: 0}, 99: {0.0: 0, 70.0: 1, 166.0: 2, 255.0: 3}, 363: {0.0: 0}, 87: {0.0: 0}, 671: {0.0: 0}, 104: {0.0: 0}, 368: {0.0: 0}, 588: {0.0: 0}, 40: {0.0: 0}, 26: {0.0: 0}, 390: {0.0: 0}, 55: {0.0: 0}, 114: {0.0: 0}, 171: {0.0: 0}, 139: {0.0: 0}, 418: {0.0: 0}, 23: {0.0: 0}, 8: {0.0: 0}, 75: {0.0: 0}, 119: {0.0: 0, 85.0: 1}, 58: {0.0: 0}, 667: {0.0: 0}, 478: {0.0: 0}, 82: {0.0: 0}, 620: {0.0: 0, 62.0: 1}, 447: {0.0: 0}, 36: {0.0: 0}, 168: {0.0: 0}, 146: {0.0: 0, 82.0: 1}, 30: {0.0: 0}, 51: {0.0: 0}, 19: {0.0: 0}, 422: {0.0: 0}, 564: {0.0: 0, 9.0: 1, 20.0: 2, 73.0: 3}, 305: {0.0: 0}, 107: {0.0: 0}, 4: {0.0: 0}, 136: {0.0: 0}, 506: {0.0: 0}, 79: {0.0: 0}, 195: {0.0: 0}, 474: {0.0: 0}, 532: {0.0: 0}, 94: {0.0: 0}, 283: {0.0: 0}, 395: {0.0: 0}, 644: {0.0: 0}, 47: {0.0: 0}, 15: {0.0: 0}, 163: {0.0: 0, 85.0: 1}, 200: {0.0: 0}, 68: {0.0: 0}, 62: {0.0: 0}, 277: {0.0: 0}, 691: {0.0: 0, 36.0: 1, 73.0: 2}, 501: {0.0: 0}, 90: {0.0: 0}, 111: {0.0: 0}, 254: {0.0: 0}, 227: {0.0: 0}, 337: {0.0: 0}, 83: {0.0: 0}, 309: {0.0: 0}, 560: {0.0: 0}, 639: {0.0: 0}, 676: {0.0: 0}, 222: {0.0: 0}, 592: {0.0: 0, 73.0: 1}, 364: {0.0: 0}}

1. **Split your original data into 65% for training and 35% for testing and store the training data into a data frame named training\_phuong and testing\_phuong respectively.**

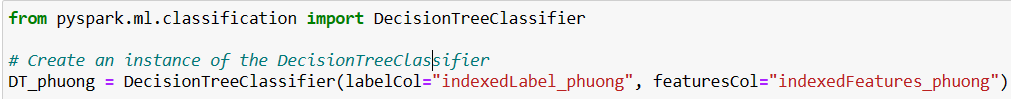
A close-up of a computer code

Description automatically generated

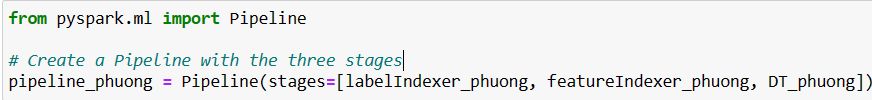
Number of records in training set: 61

Number of records in testing set: 39

1. **Create an estimator object that contains a decision tree classifier make sure to set the correct input and output columns you created during the transformation steps 4 & 5 above. Name the estimator DT\_phuong.**



1. **Create a pipeline object with three stages the first two are the transformers you defined in steps 4 & 5 and the third is the decision tree estimator you defined in step 8. Name the pipeline object pipeline\_phuong**



1. **Fit the training data to the pipeline. Store the results into an object named model\_phuong.**

A close up of a text

Description automatically generated

1. **Using the model\_phuong predict the testing data. Store the results into a data frame named predictions\_phuong.**

A close-up of a word

Description automatically generated

1. **Print the schema of the predictions.**

A close up of a text

Description automatically generated

root

|-- label: double (nullable = true)

|-- features: vector (nullable = true)

|-- indexedLabel\_phuong: double (nullable = false)

|-- indexedFeatures\_phuong: vector (nullable = true)

|-- rawPrediction: vector (nullable = true)

|-- probability: vector (nullable = true)

|-- prediction: double (nullable = false)

1. **Print the accuracy of your model and the test error.**

A computer screen shot of a program

Description automatically generated

Accuracy: 1.0

Test Error: 0.0

1. Show the first 10 predictions with the actual labels and features.



+-----+----------+--------------------+

|label|prediction| features|

+-----+----------+--------------------+

| 0.0| 1.0|(692,[124,125,126...|

| 0.0| 1.0|(692,[126,127,128...|

| 0.0| 1.0|(692,[126,127,128...|

| 0.0| 1.0|(692,[126,127,128...|

| 0.0| 1.0|(692,[127,128,129...|

| 1.0| 0.0|(692,[123,124,125...|

| 1.0| 0.0|(692,[123,124,125...|

| 1.0| 0.0|(692,[124,125,126...|

| 1.0| 0.0|(692,[124,125,126...|

| 1.0| 0.0|(692,[126,127,128...|

+-----+----------+--------------------+

only showing top 10 rows

1. **Exercise 2: Un-supervised Learning Clustering**
2. **Load the wine dataset into a data frame named wine\_x1.**

A close-up of a computer code

Description automatically generated

Row(fixed acidity=7.4, volatile acidity=0.7, citric acid=0.0, residual sugar=1.9, chlorides=0.076, free sulfur dioxide=11.0, total sulfur dioxide=34.0, density=0.9978, pH=3.51, sulphates=0.56, alcohol=9.4, quality=5)

1. **Using spark high level API functions (i.e. not pandas), carry out some initial investigation and record the results in your analysis, at minimum provide the following:**

A screen shot of a computer code

Description automatically generated

Column Names: ['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'quality']

A close-up of a text

Description automatically generated

root

|-- fixed acidity: double (nullable = true)

|-- volatile acidity: double (nullable = true)

|-- citric acid: double (nullable = true)

|-- residual sugar: double (nullable = true)

|-- chlorides: double (nullable = true)

|-- free sulfur dioxide: double (nullable = true)

|-- total sulfur dioxide: double (nullable = true)

|-- density: double (nullable = true)

|-- pH: double (nullable = true)

|-- sulphates: double (nullable = true)

|-- alcohol: double (nullable = true)

|-- quality: integer (nullable = true)

A close-up of a white background

Description automatically generated

+-------+------------------+-------------------+-------------------+------------------+--------------------+-------------------+--------------------+--------------------+-------------------+------------------+------------------+------------------+

|summary| fixed acidity| volatile acidity| citric acid| residual sugar| chlorides|free sulfur dioxide|total sulfur dioxide| density| pH| sulphates| alcohol| quality|

+-------+------------------+-------------------+-------------------+------------------+--------------------+-------------------+--------------------+--------------------+-------------------+------------------+------------------+------------------+

| count| 1599| 1599| 1599| 1599| 1599| 1599| 1599| 1599| 1599| 1599| 1599| 1599|

| mean| 8.319637273295838| 0.5278205128205131| 0.2709756097560964|2.5388055034396517| 0.08746654158849257| 15.874921826141339| 46.46779237023139| 0.9967466791744831| 3.311113195747343|0.6581488430268921|10.422983114446502|5.6360225140712945|

| stddev|1.7410963181276948|0.17905970415353525|0.19480113740531824| 1.40992805950728|0.047065302010090085| 10.46015696980971| 32.89532447829907|0.001887333953842...|0.15438646490354271|0.1695069795901101|1.0656675818473935|0.8075694397347051|

| min| 4.6| 0.12| 0.0| 0.9| 0.012| 1.0| 6.0| 0.99007| 2.74| 0.33| 8.4| 3|

| max| 15.9| 1.58| 1.0| 15.5| 0.611| 72.0| 289.0| 1.00369| 4.01| 2.0| 14.9| 8|

+-------+------------------+-------------------+-------------------+------------------+--------------------+-------------------+--------------------+--------------------+-------------------+------------------+------------------+------------------+

A close-up of a text

Description automatically generated

+-------+-------------+

|summary|fixed acidity|

+-------+-------------+

| min| 4.6|

| max| 15.9|

+-------+-------------+

+-------+----------------+

|summary|volatile acidity|

+-------+----------------+

| min| 0.12|

| max| 1.58|

+-------+----------------+

+-------+-----------+

|summary|citric acid|

+-------+-----------+

| min| 0.0|

| max| 1.0|

+-------+-----------+

+-------+--------------+

|summary|residual sugar|

+-------+--------------+

| min| 0.9|

| max| 15.5|

+-------+--------------+

+-------+---------+

|summary|chlorides|

+-------+---------+

| min| 0.012|

| max| 0.611|

+-------+---------+

+-------+-------------------+

|summary|free sulfur dioxide|

+-------+-------------------+

| min| 1.0|

| max| 72.0|

+-------+-------------------+

+-------+--------------------+

|summary|total sulfur dioxide|

+-------+--------------------+

| min| 6.0|

| max| 289.0|

+-------+--------------------+

+-------+-------+

|summary|density|

+-------+-------+

| min|0.99007|

| max|1.00369|

+-------+-------+

+-------+----+

|summary| pH|

+-------+----+

| min|2.74|

| max|4.01|

+-------+----+

+-------+---------+

|summary|sulphates|

+-------+---------+

| min| 0.33|

| max| 2.0|

+-------+---------+

+-------+-------+

|summary|alcohol|

+-------+-------+

| min| 8.4|

| max| 14.9|

+-------+-------+

+-------+-------+

|summary|quality|

+-------+-------+

| min| 3|

| max| 8|

+-------+-------+



+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+---+---------+-------+-------+

|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density| pH|sulphates|alcohol|quality|

+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+---+---------+-------+-------+

| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0|

+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+---+---------+-------+-------+

1. **Show all the distinct values in the “quality” column.**

A white background with black text

Description automatically generated

+-------+

|quality|

+-------+

| 6|

| 3|

| 5|

| 4|

| 8|

| 7|

+-------+

1. **Show the mean of the various chemical compositions across samples for the different groups of the wine quality.**

A screenshot of a computer code

Description automatically generated

+-------+------------------+---------------------+-------------------+-------------------+-------------------+------------------------+-------------------------+------------------+------------------+------------------+------------------+

|quality|mean\_fixed\_acidity|mean\_volatile\_acidity| mean\_citric\_acid|mean\_residual\_sugar| mean\_chlorides|mean\_free\_sulfur\_dioxide|mean\_total\_sulfur\_dioxide| mean\_density| mean\_pH| mean\_sulphates| mean\_alcohol|

+-------+------------------+---------------------+-------------------+-------------------+-------------------+------------------------+-------------------------+------------------+------------------+------------------+------------------+

| 6| 8.347178683385575| 0.49748432601880965| 0.2738244514106587| 2.477194357366772|0.08495611285266458| 15.711598746081505| 40.86990595611285|0.9966150626959255|3.3180721003134837|0.6753291536050158|10.629519331243463|

| 3| 8.36| 0.8845000000000001|0.17099999999999999| 2.6350000000000002|0.12250000000000001| 11.0| 24.9|0.9974640000000001|3.3979999999999997|0.5700000000000001| 9.955000000000002|

| 5| 8.167254038179149| 0.5770411160058732|0.24368575624082198| 2.528854625550658|0.09273568281938328| 16.983847283406753| 56.51395007342144|0.9971036270190888|3.3049486049926546|0.6209691629955947| 9.899706314243753|

| 4| 7.779245283018868| 0.6939622641509429| 0.1741509433962264| 2.69433962264151|0.09067924528301884| 12.264150943396226| 36.24528301886792|0.9965424528301886| 3.381509433962264|0.5964150943396227|10.265094339622639|

| 8| 8.566666666666665| 0.4233333333333334|0.39111111111111113| 2.5777777777777775|0.06844444444444445| 13.277777777777779| 33.44444444444444|0.9952122222222223|3.2672222222222214|0.7677777777777778|12.094444444444443|

| 7| 8.872361809045225| 0.4039195979899498|0.37517587939698493| 2.7206030150753793|0.07658793969849244| 14.045226130653266| 35.02010050251256|0.9961042713567828| 3.290753768844219|0.7412562814070353|11.465912897822443|

+-------+------------------+---------------------+-------------------+-------------------+-------------------+------------------------+-------------------------+------------------+------------------+------------------+------------------+

1. **Re-load the wine dataset into a data frame named wine\_x as you load add a new column named feature\_x of vector type that contains four columns as follows: “citric acid", "volatile acidity", "chlorides", "sulphates". Spread the data frame across 3 RDD partitions.**

A screenshot of a computer code

Description automatically generated

+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+----+---------+-------+-------+--------------------+

|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density| pH|sulphates|alcohol|quality| feature\_x|

+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+----+---------+-------+-------+--------------------+

| 7.4| 0.7| 0.0| 1.9| 0.076| 11.0| 34.0| 0.9978|3.51| 0.56| 9.4| 5|[0.0,0.7,0.076,0.56]|

| 7.8| 0.88| 0.0| 2.6| 0.098| 25.0| 67.0| 0.9968| 3.2| 0.68| 9.8| 5|[0.0,0.88,0.098,0...|

| 7.8| 0.76| 0.04| 2.3| 0.092| 15.0| 54.0| 0.997|3.26| 0.65| 9.8| 5|[0.04,0.76,0.092,...|

| 11.2| 0.28| 0.56| 1.9| 0.075| 17.0| 60.0| 0.998|3.16| 0.58| 9.8| 6|[0.56,0.28,0.075,...|

| 7.4| 0.7| 0.0| 1.9| 0.076| 11.0| 34.0| 0.9978|3.51| 0.56| 9.4| 5|[0.0,0.7,0.076,0.56]|

| 7.4| 0.66| 0.0| 1.8| 0.075| 13.0| 40.0| 0.9978|3.51| 0.56| 9.4| 5|[0.0,0.66,0.075,0...|

| 7.9| 0.6| 0.06| 1.6| 0.069| 15.0| 59.0| 0.9964| 3.3| 0.46| 9.4| 5|[0.06,0.6,0.069,0...|

| 7.3| 0.65| 0.0| 1.2| 0.065| 15.0| 21.0| 0.9946|3.39| 0.47| 10.0| 7|[0.0,0.65,0.065,0...|

| 7.8| 0.58| 0.02| 2.0| 0.073| 9.0| 18.0| 0.9968|3.36| 0.57| 9.5| 7|[0.02,0.58,0.073,...|

| 7.5| 0.5| 0.36| 6.1| 0.071| 17.0| 102.0| 0.9978|3.35| 0.8| 10.5| 5|[0.36,0.5,0.071,0.8]|

| 6.7| 0.58| 0.08| 1.8| 0.097| 15.0| 65.0| 0.9959|3.28| 0.54| 9.2| 5|[0.08,0.58,0.097,...|

| 7.5| 0.5| 0.36| 6.1| 0.071| 17.0| 102.0| 0.9978|3.35| 0.8| 10.5| 5|[0.36,0.5,0.071,0.8]|

| 5.6| 0.615| 0.0| 1.6| 0.089| 16.0| 59.0| 0.9943|3.58| 0.52| 9.9| 5|[0.0,0.615,0.089,...|

| 7.8| 0.61| 0.29| 1.6| 0.114| 9.0| 29.0| 0.9974|3.26| 1.56| 9.1| 5|[0.29,0.61,0.114,...|

| 8.9| 0.62| 0.18| 3.8| 0.176| 52.0| 145.0| 0.9986|3.16| 0.88| 9.2| 5|[0.18,0.62,0.176,...|

| 8.9| 0.62| 0.19| 3.9| 0.17| 51.0| 148.0| 0.9986|3.17| 0.93| 9.2| 5|[0.19,0.62,0.17,0...|

| 8.5| 0.28| 0.56| 1.8| 0.092| 35.0| 103.0| 0.9969| 3.3| 0.75| 10.5| 7|[0.56,0.28,0.092,...|

| 8.1| 0.56| 0.28| 1.7| 0.368| 16.0| 56.0| 0.9968|3.11| 1.28| 9.3| 5|[0.28,0.56,0.368,...|

| 7.4| 0.59| 0.08| 4.4| 0.086| 6.0| 29.0| 0.9974|3.38| 0.5| 9.0| 4|[0.08,0.59,0.086,...|

| 7.9| 0.32| 0.51| 1.8| 0.341| 17.0| 56.0| 0.9969|3.04| 1.08| 9.2| 6|[0.51,0.32,0.341,...|

+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+----+---------+-------+-------+--------------------+

only showing top 20 rows

1. **Cache the data frame**

A screenshot of a computer code

Description automatically generated

1599

1. **Define an estimator that uses K-means clustering to cluster all the wine instances into 6 clusters using the new feature\_x vector column you added in step 6.**

A screen shot of a computer code

Description automatically generated

1. **Print the cluster sizes and the cluster centroids.**

A screenshot of a computer program

Description automatically generated

Cluster Sizes for 6 Clusters:

+----------+-----+

|prediction|count|

+----------+-----+

| 0| 400|

| 1| 252|

| 2| 1|

| 3| 605|

| 4| 49|

| 5| 292|

+----------+-----+

Cluster Sizes for 6 Clusters:

Cluster 0: [0.434875 0.42475 0.08378 0.5968 ]

Cluster 1: [0.10079365 0.81123016 0.08349603 0.55968254]

Cluster 2: [1. 0.52 0.61 2. ]

Cluster 3: [0.1308595 0.56433884 0.08129587 0.61791736]

Cluster 4: [0.44795918 0.51755102 0.21959184 1.24142857]

Cluster 5: [0.45143836 0.3505137 0.08476712 0.80804795]

1. **Repeat steps 8 & 9 but set the number of k to 4.**

A screenshot of a computer program

Description automatically generated

Cluster Sizes for 4 Clusters:

+----------+-----+

|prediction|count|

+----------+-----+

| 0| 619|

| 1| 55|

| 2| 470|

| 3| 455|

+----------+-----+

Cluster Centroids for 4 Clusters:

Cluster 0: [0.08555732 0.68045382 0.08275637 0.59882166]

Cluster 1: [0.46037037 0.51185185 0.21912963 1.2387037 ]

Cluster 2: [0.2926569 0.48105649 0.08201255 0.60267782]

Cluster 3: [0.48931663 0.36235763 0.08394761 0.73200456]