Automating Sub-Classification for Enhanced Supply Chain Optimization and Decision-Making

By: Rohith Yelisetty, Aarav Gupta

10/21/2024

Part 1 - Statement and Project Goal:

This dataset includes information on shipments of antiretroviral drugs and HIV lab supplies to various countries, along with their prices. It also covers supply chain costs involved in delivering these items. When used with other global procurement data, this dataset provides insights into global spending on health commodities. It's especially helpful for analyzing pricing trends and the quantities delivered to each country, but the cost and lead time data should be viewed carefully, considering the broader context.

Categorization of key attributes in the dataset can go a long way in enhancing the speed and accuracy of the supply chain operation. For example, proper classification of various products leads to automatic changes in the logistics workflows, thereby reducing manual errors, which could lead to delays in the process or complete disruption. This ensures the deployment of the right resources with apt urgency so that the flow of goods is optimally maintained right from storage to distribution.

The product sub-classifications include:

- HIV Test
- HIV Test - Ancillary
- Pediatric
- Adult
- ACT
- Malaria

Automation of the classification process also avails opportunities for improved financial forecasting, where, taking into consideration factors such as transportation cost and unit price, the system would be in a better position to forecast financial needs. This, on the other hand, enables organizations to reduce risks and sudden losses. Automation can also reveal a lot of stuff about underlying product trends that help identify supply gaps and shifting demands. This may provide strategies for the improvement of resource allocation, channeling goods where they are most needed, thus enabling the support for effective and responsive public policies.

Part 2 - Dataset Description:

Attributes and Descriptions:

Attribute	**Description**
id	(Number) Shipment Identification Number
project_code	Project code

pq	Price quote (PQ) number
po_so	Purchase order
asn_dn	Shipment number: Advanced Shipment Note (ASN) for Direct Drop deliveries, or Delivery Note (DN) for from RDC deliveries
country	Destination country
managed_by	SCMS managing office: either the Program Management Office (PMO) in the U.S. or the relevant SCMS field office
fulfill_via	Method through which the shipment was fulfilled: via Direct Drop from vendor or from stock available in the RDCs
vendor_inco_term	The vendor INCO term (also known as International Commercial Terms) for Direct Drop deliveries
shipment_mode	Method by which commodities are shipped
pq_first_sent_to_client_date	Date the PQ is first sent to the client
po_sent_to_vendor_date	Date the PO is first sent to the vendor
scheduled_delivery_date	Current anticipated delivery date
delivered_to_client_date	Date of delivery to client
delivery_recorded_date	Date on which delivery to client was recorded in SCMS information systems
product_group	Product group for item, i.e. ARV, HRDT
sub_classification *(class)*	Identifies relevant product sub classifications, such as whether ARVs are pediatric or adult, whether a malaria product is an artemisinin-based combination therapy (ACT), etc.
vendor	Vendor name
item_description	Product name and formulation from Partnership for Supply Chain Management

	(PFSCM) Item Master
molecule_test_type	Active drug(s) or test kit type
brand	Generic or branded name for the item
dosage	Item dosage and unit
dosage_form	Dosage form for the item (tablet, oral solution, injection, etc.).
unit_of_measure_per_pack	Pack quantity (pills or test kits) used to compute unit price
line_item_quantity	Total quantity (packs) of commodity per line item
line_item_value	Total value of commodity per line item
pack_price	Cost per pack (i.e. month's supply of ARVs, pack of 60 test kits)
unit_price	Cost per pill (for drugs) or per test (for test kits)
manufacturing_site	Identifies manufacturing site for the line item for direct drop and from RDC deliveries
first_line_designation	Designates if the line in question shows the aggregated freight costs and weight associated with all items on the ASN DN
weight_kilograms	Weight for all lines on an ASN DN
freight_cost_usd	Freight charges associated with all lines on the respective ASN DN
line_item_insurance_usd	Line item cost of insurance, created by applying an annual flat rate ( ) to commodity cost

The initial dataset consists of 10,325 shipments with 33 attributes, including key details like the country of delivery, managing company, price, delivery date, vendor, dosage, and brand. Using these and all of the other 32 attributes mentioned in the table above, our target variable to predict is the ‘sub_classification’ attribute, which categorizes products into six possible

sub-classifications: HIV test, Pediatric, Adult, HIV test - Ancillary, ACT (artemisinin-based combination therapy), and Malaria. The dataset is imbalanced, as the majority of shipments (6,595 instances) belong to the 'Adult' sub-classification, followed by 'Pediatric' (1,955) and 'HIV test' (1,567). The remaining categories - 'HIV test - Ancillary' (161), 'ACT' (16), and 'Malaria' (30) - are significantly underrepresented. Additionally, there are missing values in attributes such as shipment mode (360), dosage (1,736), and line item insurance (287). The imbalance across the sub-classifications presents a challenge for accurate classification of the underrepresented categories, as it is not uniform and is heavily skewed.

Part 3 - Preprocessing:

In order to allow for the dataset to be effectively trained on, classified upon, and allow for proper attribute selection, the various preprocessing steps we took are described below.

3.1 Weka Preparation

In order to allow for the dataset to open on Weka, several changes were made to the dataset. This country's name was held as "C√¥te d'Ivoire," which further caused processing errors in Weka. Due to Côte d'Ivoire's poor representation in the data, this error caused an indexing issue in Weka because the special characters used to represent the accented "o" and the apostrophe in the name of the country were misencoded. To fix this, we removed the special characters, inserted the proper "o" in its place, and replaced the apostrophe with a space to ensure that the data would be properly represented and could interface appropriately with the tools being utilized for further analysis.

In addition to correcting the country name, some of the attributes had commas in values including 'country', 'vendor', 'item description', 'molecule/test type', and 'manufacturing site' which created a misalignment of data after parsing it into certain software tools. For instance, there is an issue with the attribute value in the field of the 'vendor' "Aurobindo Unit III, India", where a comma acts like a delimiter between two different attributes. The script below was written to systematically remove all commas within attribute values for the attributes listed above.

```
import pandas as pd

df = pd.read_csv('Supply_Chain_Shipment_Pricing_Dataset_New.csv', quotechar="'")

columns = ["country", "vendor", "item description", "molecule/test type", "manufacturing site"]

for column in columns:
    for val in df[column]:
        print(val)
        if "," in val:
            newVal = val.replace(",", "")
            df[column] = df[column].replace(val, newVal)

df.to_csv('Supply_Chain_Shipment_Pricing_Dataset_New_WithoutCommas.csv', index=False)
```

3.2 Missing Value Correction

The three attributes of ‘shipment mode’, ‘dosage’, and ‘line item insurance field (usd)’ each had missing value counts of 360, 1,736, and 287, respectively. To correct this issue, we employed the ReplaceMissingValues filter in Weka which replaced the values with the mode value for nominal data and the mean value for numeric data. Therefore, the missing values in ‘shipment mode’ and ‘dosage’ were replaced with the mode values of “Air” and “300mg” while the missing values in ‘line item insurance field (usd)’ was replaced with the mean value of 240.117626. This correction allows for consistency in the dataset and allows for smoother classification.

3.3 Hidden Value Correction

The dataset had numerous attributes in which there were no missing values; however, default or hidden missing values were present. The first was the recurring problem with the placeholder “Date Not Captured” in the attribute ‘pq first sent to client date’, which cropped up 205 times. These entries were a variety of missing values that needed proper treatment not to distort the results. In this case, we used the mode value for this column “Pre-PQ Process” and replaced these placeholders with that so any missing values have been consistently replaced without biasedly affecting the dataset in any way. Similarly in the ‘po sent to vendor date’ column, “Date Not Captured” is a hidden missing value that occurs 328 times, leading us to replace it with the mode value of the attribute of “N/A - From RDC”. The script below makes this correction and replaces the default values in these columns.

```
import pandas as pd

df = pd.read_csv('Supply_Chain_Shipment_Pricing_Dataset_New_WithoutCommas.csv', quotechar="'")

column1 = "pq first sent to client date"
column2 = "po sent to vendor date"

for val in column1:
    if val == "Date Not Captured":
        df[column1] = df[column1].replace(val, "Pre-PQ Process")

for val in column1:
    if val == "Date Not Captured":
        df[column1] = df[column1].replace(val, "N/A - From RDC")

df.to_csv('Supply_Chain_Shipment_Pricing_Dataset_New_WithoutCommas_DateNotCapFixed.csv', index=False)
```

One of the more complex data cleanup tasks involved cleaning the ‘weight (kilograms)’ and ‘freight cost (usd)’ fields. Some values were considered to be missing values, but only because they cross-referenced another instance already present in the dataset. For example, an entry might say, “See DN-426 (ID#:10633)” for weight or freight cost information and would store correct values under another record. Since these were missing values, the script below was written to identify the correct ID (10633 in the example above) and replace the missing values of that referring record with the correct value from the original record. This way, all linked records would be full and accurate.

```
import pandas as pd

df = pd.read_csv('Supply_Chain_Shipment_Pricing_Dataset_New_WithoutCommas_DateNotCapFixed.csv', quotechar="'")

columns = ["weight (kilograms)", "freight cost (usd)"]

for column in columns:
    for val in df[column]:
        if "See" in val:
            id_look_at = int(val.split("ID#:")[1][:-1])
            result_value = df.loc[df.index[df['id'] == id_look_at].tolist()[0], column]
            df[column] = df[column].replace(val, result_value)

df.to_csv('Supply_Chain_Shipment_Pricing_Dataset_New_WithoutCommas_DateFixed_WeightFreightFixed.csv', index=False)
```

A similar issue of missing values was faced in the ‘weight (kilograms)’ and ‘freight cost (usd)’ attributes. In this case, the weight attribute had hidden missing values of "Weight Captured Separately," whereas the freight cost column had placeholder information such as "Freight Included in Commodity Cost" and "Invoiced Separately." We used a similar approach as before for both columns and replaced the missing values with the median of the column. It maintains the balance of the dataset and does not allow analytical bias or distortion caused by outliers that might facilitate better accuracy. We wrote a script for this, as shown below, rather than utilizing Weka, as the application only replaces missing values with mean or mode values.

```
import pandas as pd

df = pd.read_csv('Supply_Chain_Shipment_Pricing_Dataset_New_WithoutCommas_DateFixed_WeightFreightFixed.csv', quotechar="'")

columns = ["weight (kilograms)", "freight cost (usd)"]

for column in columns:
    listVal = []
    for val in df[column]:
        try:
            listVal.append(float(val))
        except:
            continue
    medVal = sorted(listVal)[int(len(listVal)/2)]
    print(medVal)
    for val in df[column]:
        if "e" in val:
            df[column] = df[column].replace(val, medVal)

df.to_csv('Supply_Chain_Shipment_Pricing_Dataset_FinalCleanup.csv', index=False)
```

3.4 Removing Redundant and Derived Columns

This part of the pre-processing was done at the end rather than at the beginning because some useful columns were dependent on redundant columns for handling missing values. We removed columns such as ‘id’, ‘po / so #’, and ‘asn/dn #’ as they are unique for each shipment and will not be helpful in the identification of the sub-classification of shipment. Also, we removed ‘vendor’ and ‘item description’ as these are derived values. ‘vendor’ contains information from the pre-existing columns ‘manufacturing site’, while ‘item description’ contains information from ‘molecule/test type’ and ‘dosage form’. In summary, since the two attributes contained information from other attributes, we could remove them. In addition to this, we removed ‘product group’ because it is a derived metric from the class variable ‘sub classification’ and would not allow for real testing of model and attribute selection performance.

3.5 Normalization by Scaling

Some of our values from the dataset had really large values, so we wanted to normalize them on a scale from 1 to 1000 (the reason for this larger scale is due to certain outlier values keeping the mean on the lower side for numerical categories). We used the Weka "Normalize" Filter to do this. The attributes that were normalized are the following:

- 'unit of measure (per pack)'
- 'line item quantity'
- 'line item value'
- 'pack price'
- 'unit price'
- 'weight (kilograms)'
- 'freight cost (usd)'
- 'line item insurance (usd)'

3.6 Stratified Sampling

Since the dataset is large, we wanted to take a stratified sample of the dataset to reduce the size and time complexity of classification and training. The dataset went from 10,324 instances to 4,129 instances; we took a 40% stratified sample. Below are the number of initial instances per class to the new number of instances per class after sampling:

- Adult (6,595) ⟶ 2,638
- Pediatric (1,955) ⟶391
- 'HIV test' (1,567) ⟶ 627
- 'HIV test - Ancillary' (161) ⟶ 64
- 'ACT' (16) ⟶6
- 'Malaria' (30) ⟶12

The script to sample the dataset is below:

```
import pandas as pd

df = pd.read_csv('Supply_Chain_Shipment_Pricing_Dataset_FinalPreprocessing.csv')

df = df.groupby("'sub classification'", group_keys=False).apply(lambda x: x.sample(frac=0.4))

df.to_csv('Supply_Chain_Shipment_Pricing_Dataset_FinalSampled.csv', index=False)
```

3.7 Final Preprocessing Dataset

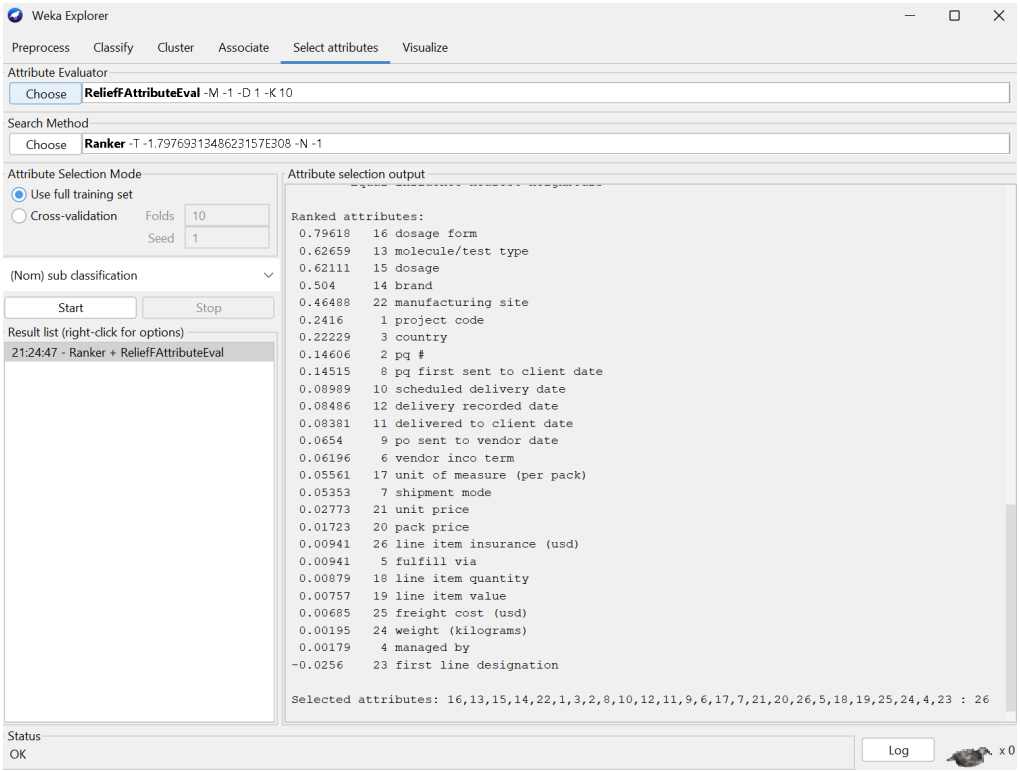
The final preprocessing dataset includes two versions: one before (linked here) and one after (linked here) stratified sampling. The class attribute used is 'sub classification.' Below are listed the features of the revised dataset.

Class Attribute: ‘sub classification’
Features: 'project code', 'pq #', ‘country’, 'managed by', 'fulfill via', 'vendor inco term', 'shipment mode', 'pq first sent to client date', 'po sent to vendor date', 'scheduled delivery date', 'delivered to client date', 'delivery recorded date', 'molecule/test type', ‘brand’, ‘dosage’, ‘dosage form’, ‘unit of measure (per pack)', 'line item quantity', 'line item value', 'pack price', 'unit price', 'manufacturing site', 'first line designation', 'weight (kilograms)', 'freight cost (usd)', 'line item insurance (usd)'

Part 4: Attribute Selection and Data Splitting

4.1 ReliefFAttributeEval

The evaluation method assesses the importance of an attribute by repeatedly sampling instances and comparing the attribute's value with the nearest instance from both the same class and a different class. This approach helps determine how effectively the attribute distinguishes between classes. The Ranker search method is used to rank the attributes based on their performance in this process.



Using an arbitrary cutoff value of ≤ 0.1, the following attributes were removed:

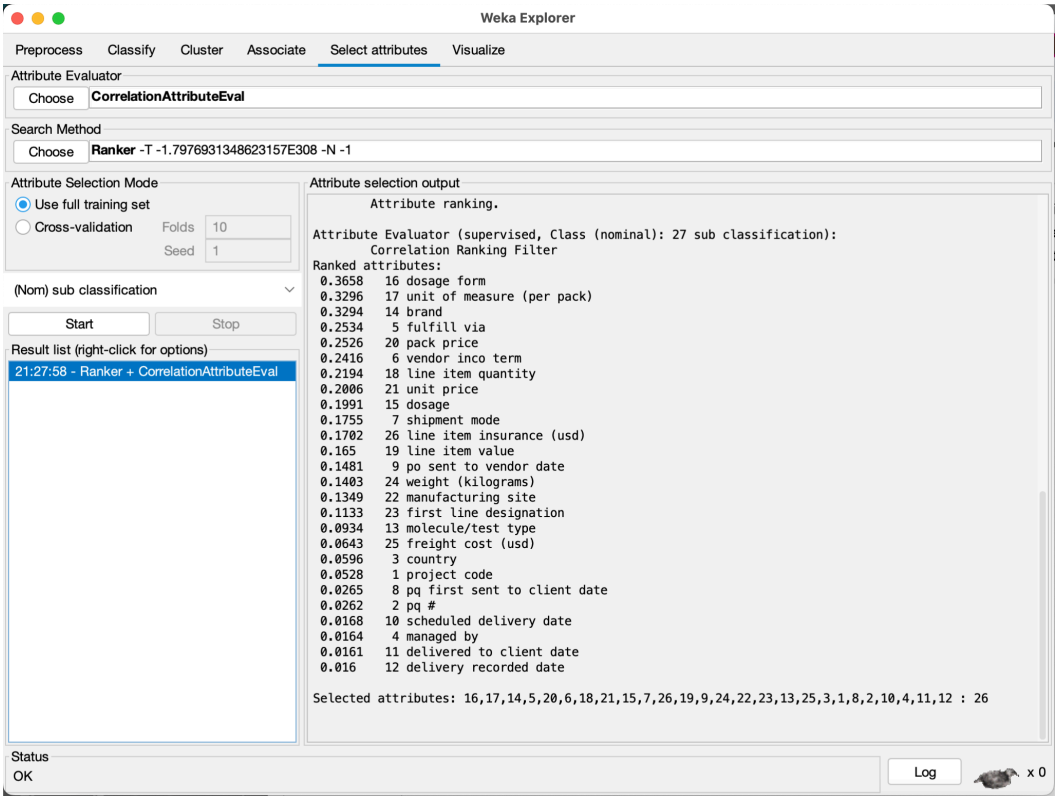
- ‘scheduled delivery date’
- ‘delivery recorded date’

- ‘delivered to client date’
- ‘po sent to vendor date’
- ‘vendor inco term’
- ‘unit of measure (per pack)’
- ‘shipment mode’
- ‘unit price’
- ‘pack price’
- ‘line item insurance (usd)‘
- ‘fulfill via’
- ‘line item quantity’
- ‘line item value‘
- ‘freight cost (usd)’
- ‘managed by’
- ‘weight (kilograms)’
- ‘first line designation’

The dataset link for ReliefFAttributeEval is provided below:
https://drive.google.com/file/d/1MSNK52eYNSQSJd61hqt5tKcII9cqu6vu/view?usp=drive_link

4.2 CorrelationAttributeEval

The evaluation method assesses the importance of an attribute by calculating the Pearson correlation between the attribute and the class. For nominal attributes, each value is treated as an indicator, with correlations calculated on a value-by-value basis. The overall correlation for a nominal attribute is then determined through a weighted average. This approach utilizes the Ranker method to rank the attributes based on their correlation to the class, and this is shown in the snippet below from Weka:



Using an arbitrary cutoff value of ≤ 0.1, the following attributes were removed:

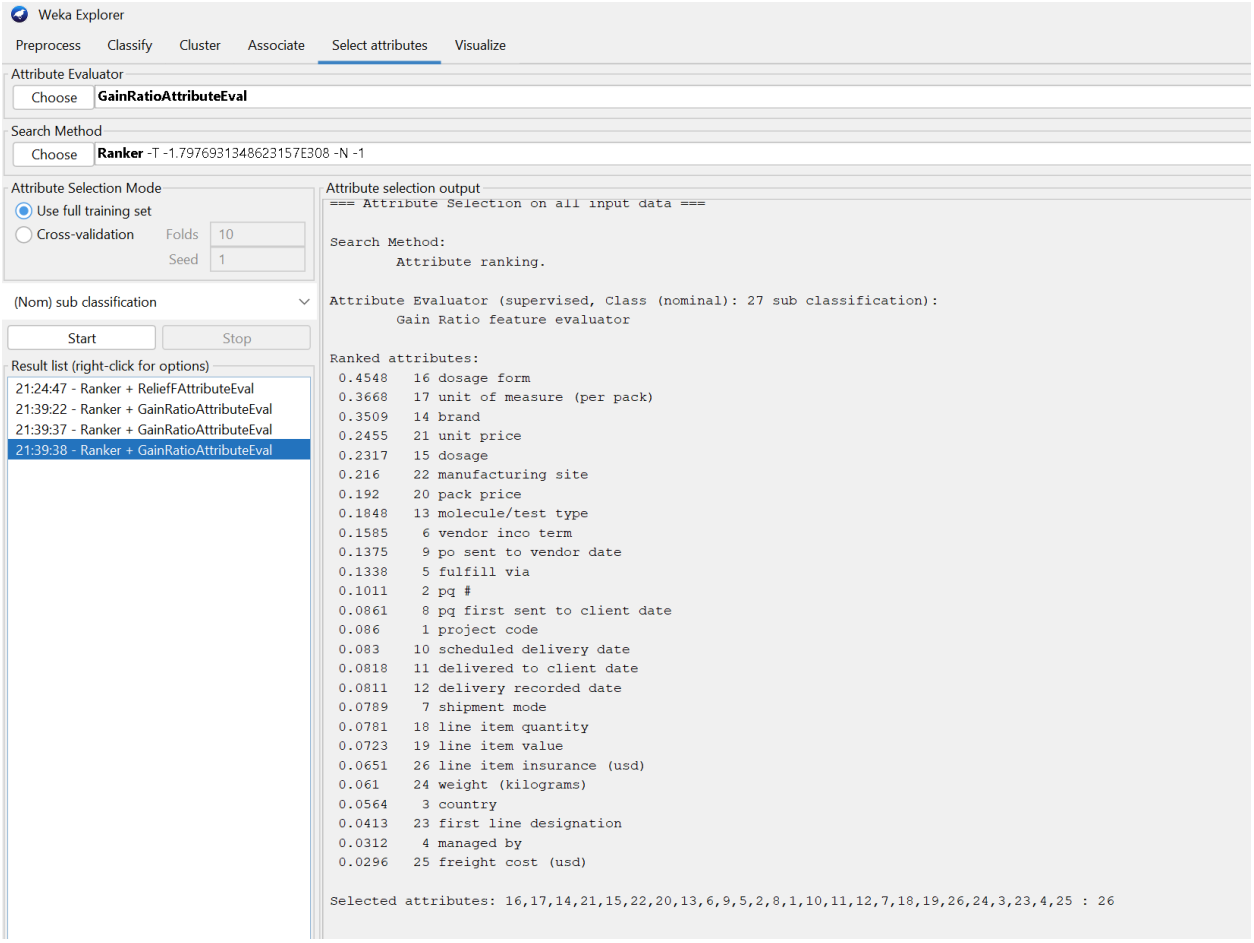
- ‘molecule/test type’
- ‘freight cost (usd)’
- ‘country’
- ‘project code’
- ‘pq first sent to client date’
- ‘pq #’
- ‘scheduled delivery date’
- ‘managed by’
- ‘delivered to client date’
- ‘delivery recorded date’

The dataset link for CorrelationAttributeEval is provided below:
https://drive.google.com/file/d/1uHdrEZbkVvp2ML1YXcq1V9uz4FIwNM9n/view?usp=drive_link

4.3 GainRatioAttributeEval

The GainRatioAttributeEval method assesses the significance of an attribute by calculating the gain ratio about the class. The gain ratio is determined using the formula: GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute), where H represents entropy. This method evaluates how well an attribute predicts the class, with a higher gain ratio

indicating a more informative attribute. We used the Ranker method so that we could rank the attributes based on their gain ratio.



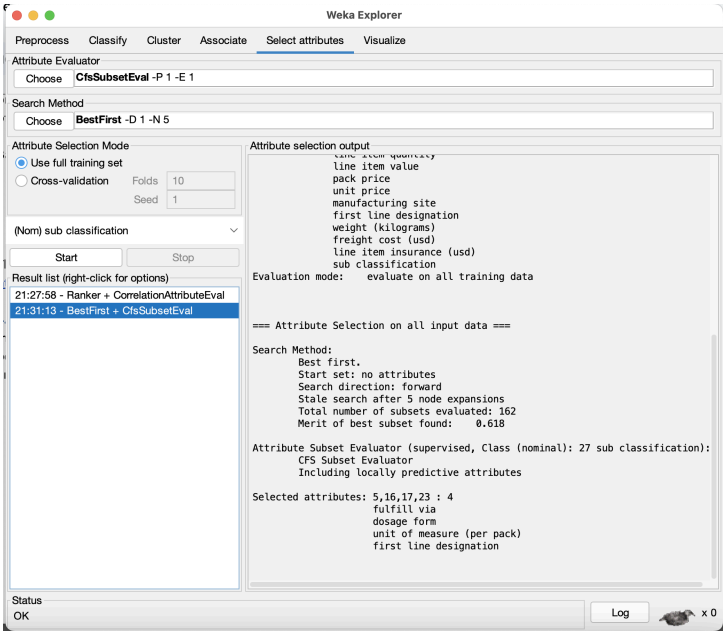
Using an arbitrary cutoff value of ≤ 0.1, the following attributes were removed:

- ‘pq first sent to client date’
- ‘project code’
- ‘scheduled delivery date’
- ‘delivered to client date’
- ‘delivery recorded date’
- ‘line item value’
- ‘shipment mode’
- ‘line item quantity’
- ‘line item insurance (usd)’
- ‘country’
- ‘weight (kilograms)’
- ‘managed by’
- ‘first line designation’
- ‘freight cost (usd)’

The dataset link for GainRatioAttributeEval is provided below:
https://drive.google.com/file/d/19ld0zUHt30W59wxF3YQaLgiGZs_1xO61/view?usp=drive_link

4.4 CfsSubsetEval

This evaluation method determines the value of a subset of attributes by considering both the individual predictive strength of each feature and the redundancy among them. Subsets that exhibit strong correlations with the class but minimal intercorrelation among features are prioritized. The BestFirst search method is used to identify and select the most effective subset of attributes.



Based on this analysis done in Weka, only four attributes were kept:

- ‘fulfill via’
- ‘dosage form’
- ‘unit of measure (per pack)’
- ‘first line designation’

The dataset link for CfsSubsetEval is provided below:
https://drive.google.com/file/d/1DiplUqFGNTSwEDeEGqUoRUrluI-NFQOY/view?usp=drive_link

4.5 Personal Attribute Selection

In our final method of attribute selection, we manually removed specific attributes based on personal judgment, with clear reasons behind each decision. Here are the attributes removed and the reasons as to why:

- ‘project code’ → This attribute is an ID and does not help with classification due to its relative uniqueness for each shipment
- ‘pq #’ → This is also an ID, and the actual ID code is very rarely mentioned throughout instances anyway.
- ‘managed by’ → PMO - US is the value for this column for all the instances except a handful
- ‘fulfill via’ → This attribute doesn’t have much value and doesn’t show much correlation with the class because its value is split half-and-half amongst all the instances.
- ‘first line designation’ → This attribute doesn’t have much value and doesn’t show much correlation with the class because its value is split half-and-half amongst all the instances.
- ‘pq first sent to client date’ → It is a date and is unique for every instance, and therefore
- ‘po sent to vendor date’ → It is a date and is unique for every instance
- ‘scheduled delivery date’ → It is a date and is unique for every instance
- ‘delivered to client date’ → It is a date and is unique for every instance
- ‘delivery recorded date’ → It is a date and is unique for every instance

The dataset link for Personal Attribute Selection is provided below:
https://drive.google.com/file/d/1oYYRldF_vPAVvZ28Rm_OX_vBV4GBGY3h/view?usp=drive_link

4.6 Train/Validation/Test Split

For our train/validation/test split, we decided on a 70/15/15 split with 70% of the stratified data going with the training, while validation and test each receive the remaining 15% of the data. The script below splits the data and sets it into the corresponding folders.

```
import pandas as pd
from sklearn.model_selection import train_test_split

folders = ['ReliefFAttributeEvalData', 'GainRatioAttributeEvalData',
           'CorrelationAttributeEvalData', 'CfsSubsetEvalData', 'PersonalAttributeData']
files = ['ReliefSampledDataset.csv', 'GainSampledDataset.csv',
         'CorrelationSampledDataset.csv', 'CfsSampledDataset.csv', 'PersonalSampledDataset.csv']

for idx, folder in enumerate(folders):
    df = pd.read_csv(f'{folder}/{files[idx]}')
    x = df.iloc[:, :-1]
    y = df.iloc[:, -1]
    X_train, X_temp, y_train, y_temp = train_test_split(x, y, test_size=0.30, stratify=y, random_state=42)
    X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.50, stratify=y_temp, random_state=42)
    train = pd.concat([X_train, y_train], axis=1)
    val = pd.concat([X_val, y_val], axis=1)
    test = pd.concat([X_test, y_test], axis=1)
    train.to_csv(f'{folder}/Train.csv', index=False)
    val.to_csv(f'{folder}/Val.csv', index=False)
    test.to_csv(f'{folder}/Test.csv', index=False)
```

Below are the final dataset folders which contain the original, train, validation, and test datasets for each attribute selection method.

ReliefFAttributeEval: *ReliefFAttributeEvalData*

CorrelationAttributeEval: *CorrelationAttributeEvalData*

GainRatioAttributeEval: *GainRatioAttributeEvalData*

CfsSubsetEval: *CfsSubsetEvalData*

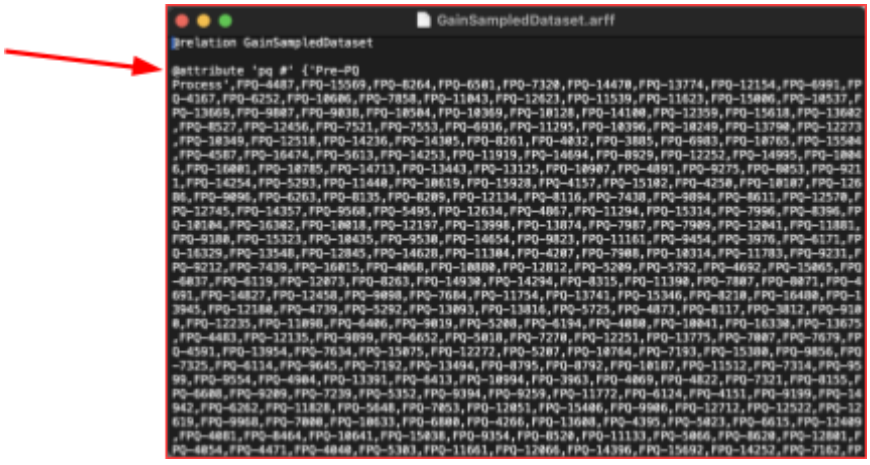
Personal Attribute Selection: *PersonalAttributeData*

4.7 Data Compatibility

We encountered issues when trying to run our model because the train and test datasets were incompatible due to differences in the number of attribute labels. To resolve this issue, we implemented the following steps:

First, we converted each CSV file into an ARFF file. Then, for every attribute selection method, we opened the FullSampledDataset ARFF file, copying from the top down to the beginning of the "@attribute" section to the "@data" signature and pasting it into both the train and test ARFF files for the attribute selection. We did this so that all the possible values could be grasped from both datasets without throwing any errors. Below show the top and bottom of the data regions to copy and paste (and replace) into the separate ARFF files.

Top:



Bottom:

Part 5: Results and Analysis

5.1 Models

bayes.NaiveBayes ⟶ This class implements a Naive Bayes classifier, which calculates the probability of a class based on feature independence assumptions. It determines numeric estimator precision values from the training data but isn't able to be updated with any new data unless the NaiveBayesUpdateable classifier is used. The default precision for numeric attributes in the updateable version is 0.1 when no training data is provided.

trees.J48 ⟶ This class implements the C4.5 decision tree algorithm, which can generate pruned or unpruned decision trees. The benefit of pruning is that it helps to avoid overfitting by removing less significant branches, and this makes the tree more generalizable and improves accuracy on unseen data.

rules.OneR ⟶ This class creates a classifier using the One Rule (1R) algorithm, which selects a single attribute that best predicts the target variable. It generates simple rules by discretizing numeric attributes and choosing the attribute with the lowest prediction error.

rules.RandomForest ⟶ This class builds a Random Forest, an ensemble of decision trees where each tree is trained on a random subset of the data and features. The final prediction is made by majority vote (classification) or averaging (regression), resulting in a model that reduces overfitting and improves accuracy, particularly for large datasets.

5.2 Results

ReliefFAttributeEval with Naive Bayes

```
=== Summary ===

Correctly Classified Instances         599               96.6129 %
Incorrectly Classified Instances        21                3.3871 %
Kappa statistic                          0.9361
Mean absolute error                      0.0139
Root mean squared error                  0.0927
Relative absolute error                  7.8083 %
Root relative squared error             31.1136 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.974     0.059     ACT
                 0.980    0.054    0.970      0.980    0.975      0.930    0.994     0.996     Adult
                 0.500    0.006    0.200      0.500    0.286      0.313    0.984     0.545     Malaria
                 0.907    0.010    0.955      0.907    0.930      0.915    0.992     0.978     Pediatric
Weighted Avg.    0.966    0.036    ?          0.966    ?          ?        0.994     0.990

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   0   0   1 |   c = ACT
   0   0   0 388   4   4 |   d = Adult
   0   0   0   1   1   0 |   e = Malaria
   0   0   0  11   0 107 |   f = Pediatric
```

ReliefFAttributeEval with J48

```
=== Summary ===

Correctly Classified Instances         618               99.6774 %
Incorrectly Classified Instances         2                0.3226 %
Kappa statistic                          0.9939
Mean absolute error                      0.0018
Root mean squared error                  0.0328
Relative absolute error                  1.0035 %
Root relative squared error             11.0072 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.009    0.995      1.000    0.997      0.993    0.998     0.997     Adult
                 0.500    0.000    1.000      0.500    0.667      0.707    0.897     0.508     Malaria
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Pediatric
Weighted Avg.    0.997    0.006    ?          0.997    ?          ?        0.997     0.995

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   1   1   0 |   e = Malaria
   0   0   0   0   0 118 |   f = Pediatric
```

ReliefFAttributeEval with OneR

```
=== Summary ===

Correctly Classified Instances         578               93.2258 %
Incorrectly Classified Instances        42                6.7742 %
Kappa statistic                          0.8649
Mean absolute error                      0.0226
Root mean squared error                  0.1503
Relative absolute error                 12.7064 %
Root relative squared error             50.4353 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.188    0.904      1.000    0.950      0.857    0.906     0.904     Adult
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.003     Malaria
                 0.669    0.000    1.000      0.669    0.802      0.788    0.835     0.732     Pediatric
Weighted Avg.    0.932    0.120    ?          0.932    ?          ?        0.906     0.883

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   2   0   0 |   e = Malaria
   0   0   0  39   0  79 |   f = Pediatric
```

ReliefFAttributeEval with Random Forest

```
=== Summary ===

Correctly Classified Instances         615               99.1935 %
Incorrectly Classified Instances         5                0.8065 %
Kappa statistic                          0.9847
Mean absolute error                      0.0138
Root mean squared error                  0.056
Relative absolute error                  7.7456 %
Root relative squared error             18.7901 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        1.000     1.000     ACT
                 1.000    0.022    0.988      1.000    0.994      0.983    1.000     1.000     Adult
                 0.500    0.000    1.000      0.500    0.667      0.707    0.998     0.700     Malaria
                 0.975    0.000    1.000      0.975    0.987      0.984    1.000     1.000     Pediatric
Weighted Avg.    0.992    0.014    ?          0.992    ?          ?        1.000     0.999

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   1   1   0 |   e = Malaria
   0   0   0   3   0 115 |   f = Pediatric
```

CorrelationAttributeEval with Naive Bayes

```
=== Summary ===

Correctly Classified Instances         527               85      %
Incorrectly Classified Instances        93               15      %
Kappa statistic                          0.7444
Mean absolute error                      0.0541
Root mean squared error                  0.2011
Relative absolute error                 30.4606 %
Root relative squared error             67.4909 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.018    0.450      1.000    0.621      0.665    0.999     0.960     HIV test -
                 0.904    0.017    0.904      0.904    0.904      0.887    0.992     0.906     HIV test
                 0.000    0.002    0.000      0.000    0.000      -0.002   0.477     0.003     ACT
                 0.836    0.040    0.974      0.836    0.899      0.768    0.962     0.971     Adult
                 0.500    0.058    0.027      0.500    0.051      0.106    0.832     0.255     Malaria
                 0.856    0.054    0.789      0.856    0.821      0.778    0.953     0.900     Pediatric
Weighted Avg.    0.850    0.039    0.916      0.850    0.877      0.783    0.964     0.944

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  85   0   1   8   0 |   b = HIV test
   0   0   0   0   0   1 |   c = ACT
  10   9   0 331  20  26 |   d = Adult
   0   0   1   0   1   0 |   e = Malaria
   1   0   0   8   8 101 |   f = Pediatric
```

CorrelationAttributeEval with J48

```
=== Summary ===

Correctly Classified Instances         615               99.1935 %
Incorrectly Classified Instances         5                0.8065 %
Kappa statistic                          0.9849
Mean absolute error                      0.003
Root mean squared error                  0.0462
Relative absolute error                  1.6623 %
Root relative squared error             15.515  %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test - Ancillary
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 0.992    0.009    0.995      0.992    0.994      0.983    0.996     0.995     Adult
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Malaria
                 0.992    0.006    0.975      0.992    0.983      0.979    0.996     0.993     Pediatric
Weighted Avg.    0.992    0.007    ?          0.992    ?          ?        0.996     0.994

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 393   0   3 |   d = Adult
   0   0   0   0   2   0 |   e = Malaria
   0   0   0   1   0 117 |   f = Pediatric
```

CorrelationAttributeEval with OneR

```
=== Summary ===

Correctly Classified Instances         578               93.2258 %
Incorrectly Classified Instances        42                6.7742 %
Kappa statistic                          0.8649
Mean absolute error                      0.0226
Root mean squared error                  0.1503
Relative absolute error                 12.7064 %
Root relative squared error             50.4353 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test - Ancillary
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.188    0.904      1.000    0.950      0.857    0.906     0.904     Adult
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.003     Malaria
                 0.669    0.000    1.000      0.669    0.802      0.788    0.835     0.732     Pediatric
Weighted Avg.    0.932    0.120    ?          0.932    ?          ?        0.906     0.883

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   2   0   0 |   e = Malaria
   0   0   0  39   0  79 |   f = Pediatric
```

CorrelationAttributeEval with Random Forest

```
=== Summary ===

Correctly Classified Instances         618               99.6774 %
Incorrectly Classified Instances         2                0.3226 %
Kappa statistic                          0.9939
Mean absolute error                      0.01
Root mean squared error                  0.0429
Relative absolute error                  5.6323 %
Root relative squared error             14.4085 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.998     0.500     ACT
                 1.000    0.009    0.995      1.000    0.997      0.993    1.000     1.000     Adult
                 0.500    0.000    1.000      0.500    0.667      0.707    0.999     0.833     Malaria
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Pediatric
Weighted Avg.    0.997    0.006    ?          0.997    ?          ?        1.000     0.999

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   1   1   0 |   e = Malaria
   0   0   0   0   0 118 |   f = Pediatric
```

GainRatioAttributeEval with Naive Bayes

```
=== Summary ===

Correctly Classified Instances         584               94.1935 %
Incorrectly Classified Instances        36                5.8065 %
Kappa statistic                          0.8904
Mean absolute error                      0.018
Root mean squared error                  0.1195
Relative absolute error                 10.1451 %
Root relative squared error             40.0916 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.002    0.900      1.000    0.947      0.948    1.000     1.000     HIV test -
                 1.000    0.008    0.959      1.000    0.979      0.976    0.995     0.922     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.900     0.016     ACT
                 0.965    0.094    0.948      0.965    0.956      0.877    0.987     0.994     Adult
                 1.000    0.011    0.222      1.000    0.364      0.469    0.997     0.450     Malaria
                 0.822    0.006    0.970      0.822    0.890      0.871    0.995     0.979     Pediatric
Weighted Avg.    0.942    0.062    ?          0.942    ?          ?        0.990     0.977

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   1   0   0   0   0   0 |   c = ACT
   0   4   0 382   7   3 |   d = Adult
   0   0   0   0   2   0 |   e = Malaria
   0   0   0  21   0  97 |   f = Pediatric
```

GainRatioAttributeEval with J48

```
=== Summary ===

Correctly Classified Instances         616               99.3548 %
Incorrectly Classified Instances         4                0.6452 %
Kappa statistic                          0.9878
Mean absolute error                      0.0031
Root mean squared error                  0.0463
Relative absolute error                  1.73   %
Root relative squared error             15.5516 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test - Ancillary
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.018    0.990      1.000    0.995      0.986    0.994     0.994     Adult
                 0.000    0.000    ?          0.000    ?          ?        0.616     0.005     Malaria
                 0.992    0.000    1.000      0.992    0.996      0.995    0.996     0.993     Pediatric
Weighted Avg.    0.994    0.011    ?          0.994    ?          ?        0.994     0.990

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   2   0   0 |   e = Malaria
   0   0   0   1   0 117 |   f = Pediatric
```

GainRatioAttributeEval with OneR

```
=== Summary ===

Correctly Classified Instances         578               93.2258 %
Incorrectly Classified Instances        42                6.7742 %
Kappa statistic                          0.8649
Mean absolute error                      0.0226
Root mean squared error                  0.1503
Relative absolute error                 12.7064 %
Root relative squared error             50.4353 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test - Ancillary
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.188    0.904      1.000    0.950      0.857    0.906     0.904     Adult
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.003     Malaria
                 0.669    0.000    1.000      0.669    0.802      0.788    0.835     0.732     Pediatric
Weighted Avg.    0.932    0.120    ?          0.932    ?          ?        0.906     0.883

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   2   0   0 |   e = Malaria
   0   0   0  39   0  79 |   f = Pediatric
```

GainRatioAttributeEval with Random Forest

```
=== Summary ===

Correctly Classified Instances         618               99.6774 %
Incorrectly Classified Instances         2                0.3226 %
Kappa statistic                          0.9939
Mean absolute error                      0.0128
Root mean squared error                  0.0486
Relative absolute error                  7.187  %
Root relative squared error             16.2953 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        1.000     1.000     ACT
                 1.000    0.009    0.995      1.000    0.997      0.993    1.000     1.000     Adult
                 0.500    0.000    1.000      0.500    0.667      0.707    1.000     1.000     Malaria
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Pediatric
Weighted Avg.    0.997    0.006    ?          0.997    ?          ?        1.000     1.000

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   1   1   0 |   e = Malaria
   0   0   0   0   0 118 |   f = Pediatric
```

CfsSubsetEval with Naive Bayes

```
=== Summary ===

Correctly Classified Instances         575               92.7419 %
Incorrectly Classified Instances        45                7.2581 %
Kappa statistic                          0.8568
Mean absolute error                      0.0301
Root mean squared error                  0.1425
Relative absolute error                 16.9109 %
Root relative squared error             47.8262 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.827     0.009     ACT
                 0.987    0.179    0.907      0.987    0.946      0.844    0.943     0.945     Adult
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Malaria
                 0.669    0.010    0.940      0.669    0.782      0.756    0.928     0.865     Pediatric
Weighted Avg.    0.927    0.116    ?          0.927    ?          ?        0.950     0.938

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 391   0   5 |   d = Adult
   0   0   0   0   2   0 |   e = Malaria
   0   0   0  39   0  79 |   f = Pediatric
```

CfsSubsetEval with J48

```
=== Summary ===

Correctly Classified Instances         603               97.2581 %
Incorrectly Classified Instances        17                2.7419 %
Kappa statistic                          0.9472
Mean absolute error                      0.0157
Root mean squared error                  0.0944
Relative absolute error                  8.8396 %
Root relative squared error             31.6832 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.076    0.959      1.000    0.979      0.941    0.954     0.950     Adult
                 0.000    0.000    ?          0.000    ?          ?        0.741     0.007     Malaria
                 0.881    0.000    1.000      0.881    0.937      0.926    0.948     0.919     Pediatric
Weighted Avg.    0.973    0.048    ?          0.973    ?          ?        0.959     0.948

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   2   0   0 |   e = Malaria
   0   0   0  14   0 104 |   f = Pediatric
```

CfsSubsetEval with OneR

```
=== Summary ===

Correctly Classified Instances         578               93.2258 %
Incorrectly Classified Instances        42                6.7742 %
Kappa statistic                          0.8649
Mean absolute error                      0.0226
Root mean squared error                  0.1503
Relative absolute error                 12.7064 %
Root relative squared error             50.4353 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.188    0.904      1.000    0.950      0.857    0.906     0.904     Adult
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.003     Malaria
                 0.669    0.000    1.000      0.669    0.802      0.788    0.835     0.732     Pediatric
Weighted Avg.    0.932    0.120    ?          0.932    ?          ?        0.906     0.883

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   2   0   0 |   e = Malaria
   0   0   0  39   0  79 |   f = Pediatric
```

CfsSubsetEval with Random Forest

```
=== Summary ===

Correctly Classified Instances         605               97.5806 %
Incorrectly Classified Instances        15                2.4194 %
Kappa statistic                          0.9536
Mean absolute error                      0.0143
Root mean squared error                  0.0883
Relative absolute error                  8.043  %
Root relative squared error             29.65   %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.497     0.002     ACT
                 1.000    0.067    0.964      1.000    0.981      0.948    0.985     0.987     Adult
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Malaria
                 0.881    0.000    1.000      0.881    0.937      0.926    0.977     0.947     Pediatric
Weighted Avg.    0.976    0.043    ?          0.976    ?          ?        0.985     0.980

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   0   2   0 |   e = Malaria
   0   0   0  14   0 104 |   f = Pediatric
```

Personal Selection (Non-Weka) with Naive Bayes

```
=== Summary ===

Correctly Classified Instances         508               81.9355 %
Incorrectly Classified Instances       112               18.0645 %
Kappa statistic                          0.7008
Mean absolute error                      0.0602
Root mean squared error                  0.2144
Relative absolute error                 33.8624 %
Root relative squared error             71.9445 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.889    0.026    0.333      0.889    0.485      0.535    0.997     0.926     HIV test -
                 0.936    0.021    0.889      0.936    0.912      0.896    0.993     0.925     HIV test
                 0.000    0.002    0.000      0.000    0.000      -0.002   0.486     0.003     ACT
                 0.775    0.040    0.972      0.775    0.862      0.706    0.964     0.972     Adult
                 0.500    0.049    0.032      0.500    0.061      0.117    0.833     0.505     Malaria
                 0.881    0.090    0.698      0.881    0.779      0.727    0.958     0.906     Pediatric
Weighted Avg.    0.819    0.046    0.893      0.819    0.845      0.734    0.966     0.949

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   8   1   0   0   0   0 |   a = HIV test - Ancillary
   1  88   0   1   4   0 |   b = HIV test
   0   0   0   0   0   1 |   c = ACT
  14  10   0 307  21  44 |   d = Adult
   0   0   1   0   1   0 |   e = Malaria
   1   0   0   8   5 104 |   f = Pediatric
```

Personal Selection (Non-Weka) with J48

```
=== Summary ===

Correctly Classified Instances         616               99.3548 %
Incorrectly Classified Instances         4                0.6452 %
Kappa statistic                          0.9878
Mean absolute error                      0.0031
Root mean squared error                  0.0463
Relative absolute error                  1.73   %
Root relative squared error             15.5516 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test - Ancillary
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.018    0.990      1.000    0.995      0.986    0.994     0.994     Adult
                 0.000    0.000    ?          0.000    ?          ?        0.616     0.005     Malaria
                 0.992    0.000    1.000      0.992    0.996      0.995    0.996     0.993     Pediatric
Weighted Avg.    0.994    0.011    ?          0.994    ?          ?        0.994     0.990

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   2   0   0 |   e = Malaria
   0   0   0   1   0 117 |   f = Pediatric
```

Personal Selection (Non-Weka) with OneR

```
=== Summary ===

Correctly Classified Instances         578               93.2258 %
Incorrectly Classified Instances        42                6.7742 %
Kappa statistic                          0.8649
Mean absolute error                      0.0226
Root mean squared error                  0.1503
Relative absolute error                 12.7064 %
Root relative squared error             50.4353 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test - Ancillary
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     ACT
                 1.000    0.188    0.904      1.000    0.950      0.857    0.906     0.904     Adult
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.003     Malaria
                 0.669    0.000    1.000      0.669    0.802      0.788    0.835     0.732     Pediatric
Weighted Avg.    0.932    0.120    ?          0.932    ?          ?        0.906     0.883

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   2   0   0 |   e = Malaria
   0   0   0  39   0  79 |   f = Pediatric
```

Personal Selection (Non-Weka) with Random Forest

```
=== Summary ===

Correctly Classified Instances         618               99.6774 %
Incorrectly Classified Instances         2                0.3226 %
Kappa statistic                          0.9939
Mean absolute error                      0.0053
Root mean squared error                  0.0372
Relative absolute error                  2.9657 %
Root relative squared error             12.4942 %
Total Number of Instances              620

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test -
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     HIV test
                 0.000    0.000    ?          0.000    ?          ?        1.000     1.000     ACT
                 1.000    0.009    0.995      1.000    0.997      0.993    1.000     1.000     Adult
                 0.500    0.000    1.000      0.500    0.667      0.707    0.999     0.833     Malaria
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Pediatric
Weighted Avg.    0.997    0.006    ?          0.997    ?          ?        1.000     0.999

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   9   0   0   0   0   0 |   a = HIV test - Ancillary
   0  94   0   0   0   0 |   b = HIV test
   0   0   0   1   0   0 |   c = ACT
   0   0   0 396   0   0 |   d = Adult
   0   0   0   1   1   0 |   e = Malaria
   0   0   0   0   0 118 |   f = Pediatric
```

5.3 Analysis

The best models, strictly evaluated for accuracy, that is to say the number of instances correctly or incorrectly classified, all performed similarly. Among these models, only two instances were misclassified: the singular "ACT" test point and one out of two "Malaria" test points. The best models were the following:

- Personal Selection with Random Forest
- Gain Ratio with Random Forest
- Correlation with Random Forest
- ReliefF with J48

All these models did an extremely good job of having very few misclassifications.

The incorrect identification of certain subclasses is quite common and repetitive due to the low incidence of the sub-classifications "ACT" and "Malaria". Considering a full dataset of 16 "ACT" entries and 30 "Malaria" entries, when a 40% sampling is conducted on the dataset, this number dropped down to 6 "ACT" and 12 "Malaria" instances. When the data was divided into training, validation, and testing subsets, only 4 "ACT" and 8 "Malaria" instances were allotted to the training set, while only 1 "ACT" and 2 "Malaria" entries went into the testing set. With less than 10 examples for training, this lack of data not only vastly affected the performance of the model in classifying both categories properly but also did terribly during the test.

In our model evaluation, we found that the focus on only accuracies and the confusion matrix gave similar results across all models. One might have been misled at the beginning to think that a Random Forest model would perform best due to its repetition in the models with high accuracy. Their actual error values, however, provided a different conclusion. Among four kinds of error metrics, including Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, and Root Relative Squared Error, the best selection method and model is **ReliefFAttributeEval combined with the J48 model**, showing the lowest error amongst all of the other models that performed equally well. It thus performs best to maintain minimum error.

ReliefFAttributeEval	CorrelationAttributeEval	GainRatioAttributeEval	CfsSubsetEval	Personal Selection
- project code - pq # - country - pq first sent to client date - molecule/test type - brand - dosage - dosage form - manufacturing site	- fulfill via - vendor inco term - shipment mode - po sent to vendor date - brand - dosage - dosage form - unit of measure (per pack) - line item quantity - line item value - pack price - unit price - manufacturing site - first line designation - weight (kilograms) - line item insurance (usd)	- pq # - fulfill via - vendor inco term - po sent to vendor date - molecule/test type - brand - dosage - dosage form - unit of measure (per pack) - pack price - unit price - manufacturing site	- fulfill via - dosage form - unit of measure (per pack) - first line designation	- country - vendor inco term - shipment mode - molecule/test type - brand - dosage - dosage form - unit of measure (per pack) - line item quantity - line item value - pack price - unit price - manufacturing site - weight (kilograms) - freight cost (usd) - line item insurance (usd)

It was noticed from the analysis of attribute selection methods that a few attributes were common in different methods and some were specific to particular approaches. This is shown in the table of listing of attributes which represents the various attributes found in each of the five selection methods; namely ReliefFAttributeEval, CorrelationAttributeEval, GainRatioAttributeEval, CfsSubsetEval, and Personal Selection. The majority of overlap in some features can be further depicted from it. However, the only attribute that appeared in all 5

selection methods was "dosage form". This indicated that dosage form is an important attribute, especially since it is a critical differentiator for pediatrics versus adults, as pediatrics formulations cannot be of tablet form.

The appearance of "dosage form" as a critical attribute across all attribute selection methods in the analysis is an indication of its importance in determining the proper classification for pharmaceutical products, especially in differentiating between pediatric and adult formulations. Drugs to be prescribed to children often have to be given in forms that are easy to take, such as oral solutions, chewable tablets, or suspensions, because small children often could not swallow tablets or capsules. The dosage form determined here in the dataset will give a better prediction for which target patient group the drug falls.

The data immediately shows that certain dosage forms, such as ‘tablets’, ‘capsules’, and ‘delayed-release capsules’, are quite clearly better suited for adults. For example, there are a high number of entries for regular tablets and "Tablet-FDC," and minor entries such as the "Tablet-FDC + blister" and "Tablet-blister." None of these are normally indicated for children since pediatric patients generally require alternative formulations. Tablets are in solid form and for exact dosing, hence generally not being indicated for younger patients; thus, it can be safely assumed that products in tablet form are for adults.

Contrary to this fact, the data shows that other forms like oral solution, chewable tablets, dispersible tablets, and oral suspension are more applicable to pediatric patients. This makes such dosage forms much more pediatrically appropriate, as ingestion is way easier either through liquid formulations or by making it chewable. This, therefore, gives a generally accurate distinction based on dosage form, hence enabling a more correct classification of the products and pinpointing those that are indeed pediatric formulations, even if this assumption doesn’t hold true in every instance.

This identification of a pattern underlines why "dosage form" is one of the top attributes in the model for all selection methods and why its presence is crucial to making correct classifications. This observation resulted in the ReliefFAttributeEval incorporated with the J48 decision tree algorithm as the overall best model that exhibited relatively the best performance in terms of the most balanced performance for these attributes. The table that outlines these attributes is placed after describing each of the attribute selection methods.

Part 6: Conclusion and Reproducibility

The ReliefFAttributeEval selection method, combined with the J48 algorithm, gave the best performance in our analysis. It outperformed other models when error metrics such as Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, and Root Relative Squared Error are taken into account. This model maintained a good balance of performance for correct classification and reduction of errors on account of the imbalance nature of this dataset, where certain sub-classifications like “ACT” and “Malaria” were highly underrepresented.

Generally speaking, we automate the identification of product subcategories in the supply chain dataset using Weka where the model would determine what sub classification the product is. In the process, we took some important lessons using Weka, which include how preprocessing steps of filling missing values, elimination of redundant columns, and normalizing numeric values are very necessary to ensure that the model would run without a hitch and results obtained would be reliable. Moreover, the application of different attribute selection methods helped us to understand much better what features were contributing most to the classification task.

Building the dataset in the future to include more instances of the underrepresented sub-classifications like “ACT” and “Malaria” would yield better results in the classification by the model of such categories. Also, further research could be done regarding more advanced machine learning techniques or ensemble methods that improve the performance of models dealing with highly imbalanced datasets and increase their applicability in real-world use.

Steps to Reproduce ReliefFAttributeEval with J48:

1. Take the initial Supply_Chain_Shipment_Pricing_Dataset.csv and run through all of the Pre-Processing steps as detailed below and above in Part 3 of the report
    a. Remove the special characters and apostrophes
    b. Fix missing values using the ‘ReplaceMissingValues’ filter on Weka under “unsupervised” and “attributes”
    c. Replace hidden missing values with the scripts shown in **“3.3 Hidden Value Correction”**
    d. Remove the ‘id’, ‘po / so #’, ‘asn/dn #’, ‘vendor’, ‘item description’, and ‘product group due to them being redundant or derived
    e. Normalize all of the numeric attributes on a scale from 1 to 1000 using the ‘Normalize’ filter on Weka under “unsupervised” and “attributes” and then clicking on the white space and changing the scale attribute to 1000
    f. Run the script as shown in **“3.6 Stratified Sampling”** to take a 40% stratified sample of the dataset

g. **Final pre-processed and sampled dataset can be found in Supply_Chain_Shipment_Pricing_Dataset_FinalSampled.csv**

2. Click on the "Select attributes" tab of Weka after loading in the Supply_Chain_Shipment_Pricing_Dataset_FinalSampled.csv
    a. Choose "ReliefFAttributeEval" under "Attribute Evaluator" and select "Yes" to allow for a Ranker Search Method
    b. After ensuring that the 'sub classification' is chosen for the class variable, hit "Start" and keep track of which attributes have higher than a 0.1 value and the class attribute
    c. Remove all other attributes in the Preprocess tab and save this dataset (Currently saved as ReliefSampledDataset.csv under the ReliefFAttributeEvalData folder)
    d. **Run the script in "4.6 Train/Validation/Test Split" to split the dataset into a 70/15/15 train/validation/test split (Saved as ReliefTrain.csv, ReliefVal.csv, and ReliefTest.csv under the ReliefFAttributeEvalData folder)**
    e. **Follow the process shown in "4.7 Data Compatibility" to ensure that the data is compatible when running the model (Saved as ReliefSampledDataset.arff, ReliefTrain.arff, and ReliefTest.arff under the ReliefFAttributeEvalData folder)**
3. Click on the "Classify" tab after loading in the ReliefTest.arff in the Preprocess tab
    a. Choose "trees" and "J48" under "Classifier"
    b. Click on "Supplied Test Set" and click the "Set" button and open the ReliefTest.arff file and set the class as "sub classification" and then click "Close"
    c. Make sure the class is selected to "(Nom) sub classification" on the "Classify" tab
    d. Click "Start" to run the J48 model
    e. **Right click on the model on the left bar of the screen and select "Save Model" to save the model (Saved as bestModelJ48Relief.model under the ReliefFAttributeEvalData folder)**

Part 7 - Team Members and Tasks Performed

Finding the Data & Building Proposal: Aarav Gupta
Preprocessing Initial Attempt: Rohith Yelisetty
Preprocessing & Project Update: Rohith Yelisetty
Non-Weka Attribute Selection Algorithm: Aarav Gupta
Attribute Selection Algorithms and Classifiers: Rohith Yelisetty
Results Output: Rohith Yelisetty
Results Analysis: Aarav Gupta
Building Final Report: Aarav Gupta

Part 8 - File Appendix and References

8.1 File Appendix

- Supply_Chain_Shipment_Pricing_Dataset.csv → Original dataset
- Supply_Chain_Shipment_Pricing_Dataset_FinalPreprocessing.csv → Full Dataset after all preprocessing is complete
- Supply_Chain_Shipment_Pricing_Dataset_FinalSampled.csv → 40% Sampled Dataset after all preprocessing was complete
- Scripts Folder → Contains all of the scripts necessary for preprocessing the data as detailed in Part 3 of the report
- Folder Structure (*ReliefFAttributeEvalData, CorrelationAttributeEvalData, GainRatioAttributeEvalData, CfsSubsetEvalData, PersonalAttributeData*)
    - (AttributeSelector)SampledDataset.csv (or .arff) → Dataset after removing attributes below cutoff value for each specific attribute selector
    - (AttributeSelector)Train.csv (or .arff) → 70% of (Attribute Selector) Sampled Dataset used for training the model
    - (AttributeSelector)Val.csv → 15% of (AttributeSelector)SampledDataset.csv used for validation
    - (AttributeSelector)Test.csv (or .arff) → 15% of (Attribute Selector) Sampled Dataset used for evaluating and testing the model
- ModelJ48Relief.model → Chosen model for the project

8.2 References

Dataset:

[1] Palacios, M. (2021). Supply Chain Shipment Pricing Dataset [Dataset]. In Data.gov. Doby. https://catalog.data.gov/dataset/supply-chain-shipment-pricing-data-07d29

Model Information Links:

[2] https://weka.sourceforge.io/doc.stable/weka/classifiers/bayes/NaiveBayes.html

[3] https://weka.sourceforge.io/doc.stable/weka/classifiers/trees/J48.html

[4] https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/OneR.html

[5] https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/RandomForest.html