Classification of Medical Supply Shipments

By: Aarav Gupta, Rohith Yelisetty

Statement and Project Goal

- Dataset provides information on shipments of antiretroviral drugs and HIV lab supplies to various countries

- Provides valuable insights into global spending on health commodities

- Proper categorization of key attributes can improve supply chain efficiency by reducing manual errors and optimizing logistics workflows

We are predicting the **sub-classification** of each shipment.

AG

Dataset Description

Possible sub-classifications and incidence rate:

- HIV Test (1,567)
- HIV Test - Ancillary (161)
- Pediatric (1,955)
- Adult (6,595)
- ACT (16)
- Malaria (30)

10,325 shipments with 32 attributes (Not including class)

Majority of shipments - Adult, Pediatric, HIV Test

Heavily skewed

Pre-Processing Steps

1. WEKA Preparation
2. Missing Value Correction
3. Hidden Value Correction
4. Removing Redundant and Derived Columns
5. Normalization by Scaling
6. Stratified Sampling

Weka Preparation

Côte d'Ivoire

item description
HIV, Reveal G3 Rapi

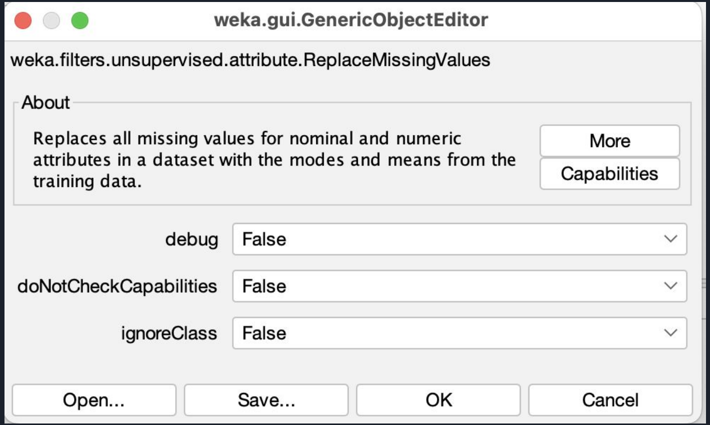
```python
import pandas as pd

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)
```

Missing Value Correction

‘shipment mode’ → 360 missing values

‘dosage’ → 1,736 missing values

‘line item insurance field (usd)’ → 287 missing values



AG

Hidden Value Correction

See ASN-93 (ID#:1281)

```python
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)
```

Replaces "Date Not Captured" with mode value of columns

po sent to vendor dat
Date Not Captured
Date Not Captured
Date Not Captured

Replaces "Weight Captured Separately" and "Freight Included in Commodity Cost" with the median value of columns

Freight Included in Commodity Cost

Removing Redundant and Derived Columns

Unique values for every instance:

- ‘Id’
- ‘po / so #’
- ‘asn/dn #’

Derived columns:

- ‘vendor’
- ‘item description’
- ‘product group’

AG

Normalization by Scaling and Stratified Sampling

Numeric Attributes in Dataset:

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

Scaled on a range from 1 - 1000 (Large due to outliers)

40% Stratified Sample

10,325 instances (shipments) got converted into 4,197 instances

New sub-classification counts:

- 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

Attribute Selection Methods

ReliefFAttributeEval → 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

CorrelationAttributeEval → Assesses the importance of an attribute by calculating the Pearson correlation between the attribute and the class

GainRatioAttributeEval → Assesses the significance of an attribute by calculating the gain ratio about the class

CfsSubsetEval → Determines the value of a subset of attributes by considering both the individual predictive strength of each feature and the redundancy among them

ReliefFAttributeEval

Arbitrary cutoff value of ≤ 0.1

```
Ranked attributes:
 0.79618   16 dosage form
 0.62659   13 molecule/test type
 0.62111   15 dosage
 0.504     14 brand
 0.46488   22 manufacturing site
 0.2416     1 project code
 0.22229    3 country
 0.14606    2 pq #
 0.14515    8 pq first sent to client date
 0.08989   10 scheduled delivery date
 0.08486   12 delivery recorded date
 0.08381   11 delivered to client date
 0.0654     9 po sent to vendor date
 0.06196    6 vendor inco term
 0.05561   17 unit of measure (per pack)
 0.05353    7 shipment mode
 0.02773   21 unit price
 0.01723   20 pack price
 0.00941   26 line item insurance (usd)
 0.00941    5 fulfill via
 0.00879   18 line item quantity
 0.00757   19 line item value
 0.00685   25 freight cost (usd)
 0.00195   24 weight (kilograms)
 0.00179    4 managed by
-0.0256    23 first line designation

Selected attributes: 16,13,15,14,22,1,3,2,8,10,12,11,9,6,17,7,21,20,26,5,18,19,25,24,4,23 : 26
```

CorrelationAttributeEval

Arbitrary cutoff value of ≤ 0.1

```
Ranked attributes:
 0.3658   16 dosage form
 0.3296   17 unit of measure (per pack)
 0.3294   14 brand
 0.2534    5 fulfill via
 0.2526   20 pack price
 0.2416    6 vendor inco term
 0.2194   18 line item quantity
 0.2006   21 unit price
 0.1991   15 dosage
 0.1755    7 shipment mode
 0.1702   26 line item insurance (usd)
 0.165    19 line item value
 0.1481    9 po sent to vendor date
 0.1403   24 weight (kilograms)
 0.1349   22 manufacturing site
 0.1133   23 first line designation
 0.0934   13 molecule/test type
 0.0643   25 freight cost (usd)
 0.0596    3 country
 0.0528    1 project code
 0.0265    8 pq first sent to client date
 0.0262    2 pq #
 0.0168   10 scheduled delivery date
 0.0164    4 managed by
 0.0161   11 delivered to client date
 0.016    12 delivery recorded date

Selected attributes: 16,17,14,5,20,6,18,21,15,7,26,19,9,24,22,23,13,25,3,1,8,2,10,4,11,12 : 26
```

AG

GainRatioAttributeEval

Arbitrary cutoff value of ≤ 0.1

```
Ranked attributes:
 0.4548   16 dosage form
 0.3668   17 unit of measure (per pack)
 0.3509   14 brand
 0.2455   21 unit price
 0.2317   15 dosage
 0.216    22 manufacturing site
 0.192    20 pack price
 0.1848   13 molecule/test type
 0.1585    6 vendor inco term
 0.1375    9 po sent to vendor date
 0.1338    5 fulfill via
 0.1011    2 pq #
 0.0861    8 pq first sent to client date
 0.086     1 project code
 0.083    10 scheduled delivery date
 0.0818   11 delivered to client date
 0.0811   12 delivery recorded date
 0.0789    7 shipment mode
 0.0781   18 line item quantity
 0.0723   19 line item value
 0.0651   26 line item insurance (usd)
 0.061    24 weight (kilograms)
 0.0564    3 country
 0.0413   23 first line designation
 0.0312    4 managed by
 0.0296   25 freight cost (usd)

Selected attributes: 16,17,14,21,15,22,20,13,6,9,5,2,8,1,10,11,12,7,18,19,26,24,3,23,4,25 : 26
```

CfsSubsetEval

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 162
        Merit of best subset found:    0.618

Attribute Subset Evaluator (supervised, Class (nominal): 27 sub classification):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 5,16,17,23 : 4
                     fulfill via
                     dosage form
                     unit of measure (per pack)
                     first line designation
```

Personal Attribute Selection

Removed Attributes:

- ‘project code’
- ‘pq #’
- ‘managed by’
- ‘fulfill via’
- ‘first line designation’
- ‘pq first sent to client date’
- ‘po sent to vendor date’
- ‘scheduled delivery date’
- ‘delivered to client date’
- ‘delivery recorded date’

Train/Validation/Test Split

70/15/15 split

Training → 2,890 instances

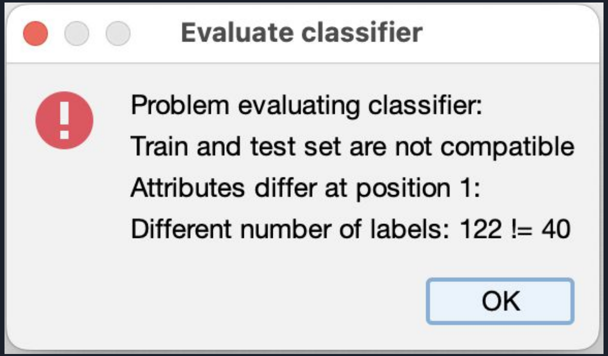
Validation → 619 instances

Testing → 620 instances

```
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)
```

Data Compatibility

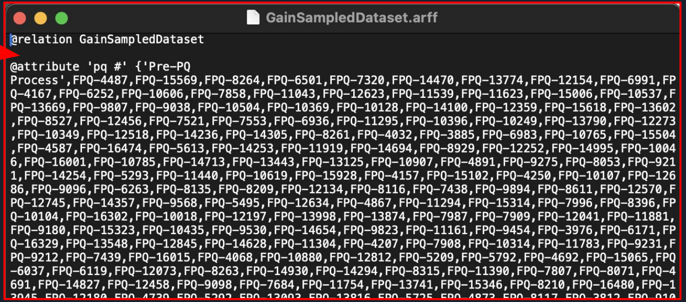
Train and test datasets had different attribute labels, leading to errors



Solution Implemented:

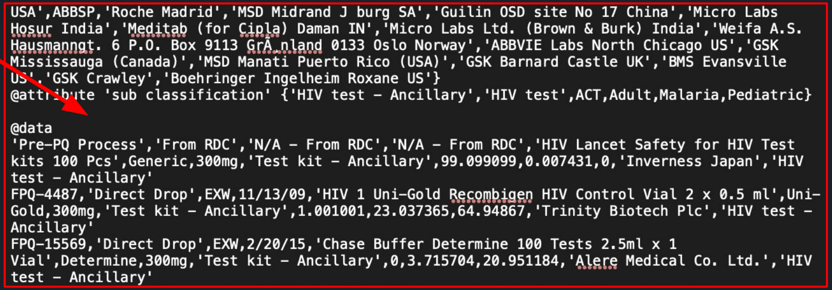
- Converted each CSV file into an ARFF file
- Opened the FullSampledDataset ARFF file for each attribute selection method
- Copied data from the “@attribute” section down to the “@data” signature
- Pasted this data into the train and test ARFF files for the attribute selection

Top:



```
GainSampledDataset.arff
@relation GainSampledDataset

@attribute 'pq #' {'Pre-PQ
Process',FPQ-4487,FPQ-15569,FPQ-8264,FPQ-6501,FPQ-7320,FPQ-14470,FPQ-13774,FPQ-12154,FPQ-6991,FP
Q-4167,FPQ-6252,FPQ-10606,FPQ-7858,FPQ-11043,FPQ-12623,FPQ-11539,FPQ-11623,FPQ-15006,FPQ-10537,F
PQ-13669,FPQ-9807,FPQ-9038,FPQ-10504,FPQ-10369,FPQ-10128,FPQ-14100,FPQ-12359,FPQ-15618,FPQ-13602
,FPQ-8527,FPQ-12456,FPQ-7521,FPQ-7553,FPQ-6936,FPQ-11295,FPQ-10396,FPQ-10249,FPQ-13790,FPQ-12273
,FPQ-10349,FPQ-12518,FPQ-14236,FPQ-14305,FPQ-8261,FPQ-4032,FPQ-3885,FPQ-6983,FPQ-10765,FPQ-15504
,FPQ-4587,FPQ-16474,FPQ-5613,FPQ-14253,FPQ-11919,FPQ-14694,FPQ-8929,FPQ-12252,FPQ-14995,FPQ-1004
6,FPQ-16001,FPQ-10785,FPQ-14713,FPQ-13443,FPQ-13125,FPQ-10907,FPQ-4891,FPQ-9275,FPQ-8053,FPQ-921
1,FPQ-14254,FPQ-5293,FPQ-11440,FPQ-10619,FPQ-15928,FPQ-4157,FPQ-15102,FPQ-4250,FPQ-10107,FPQ-126
86,FPQ-9096,FPQ-6263,FPQ-8135,FPQ-8209,FPQ-12134,FPQ-8116,FPQ-7438,FPQ-9894,FPQ-8611,FPQ-12570,F
PQ-12745,FPQ-14357,FPQ-9568,FPQ-5495,FPQ-12634,FPQ-4867,FPQ-11294,FPQ-15314,FPQ-7996,FPQ-8396,FP
Q-10104,FPQ-16302,FPQ-10018,FPQ-12197,FPQ-13998,FPQ-13874,FPQ-7987,FPQ-7909,FPQ-12041,FPQ-11881,
FPQ-9180,FPQ-15323,FPQ-10435,FPQ-9530,FPQ-14654,FPQ-9823,FPQ-11161,FPQ-9454,FPQ-3976,FPQ-6171,FP
Q-16329,FPQ-13548,FPQ-12845,FPQ-14628,FPQ-11304,FPQ-4207,FPQ-7908,FPQ-10314,FPQ-11783,FPQ-9231,F
PQ-9212,FPQ-7439,FPQ-16015,FPQ-4068,FPQ-10880,FPQ-12812,FPQ-5209,FPQ-5792,FPQ-4692,FPQ-15065,FPQ
-6037,FPQ-6119,FPQ-12073,FPQ-8263,FPQ-14930,FPQ-14294,FPQ-8315,FPQ-11390,FPQ-7807,FPQ-8071,FPQ-4
691,FPQ-14827,FPQ-12458,FPQ-9098,FPQ-7684,FPQ-11754,FPQ-13741,FPQ-15346,FPQ-8210,FPQ-16480,FPQ-1
```

Bottom:



```
USA',ABBSP,'Roche Madrid','MSD Midrand J burg SA','Guilin OSD site No 17 China','Micro Labs
Hosur India','Meditab (for Cipla) Daman IN','Micro Labs Ltd. (Brown & Burk) India','Weifa A.S.
Hausmanngt. 6 P.O. Box 9113 GrÂ nland 0133 Oslo Norway','ABBVIE Labs North Chicago US','GSK
Mississauga (Canada)','MSD Manati Puerto Rico (USA)','GSK Barnard Castle UK','BMS Evansville
US','GSK Crawley','Boehringer Ingelheim Roxane US'}
@attribute 'sub classification' {'HIV test - Ancillary','HIV test',ACT,Adult,Malaria,Pediatric}

@data
'Pre-PQ Process','From RDC','N/A - From RDC','N/A - From RDC','HIV Lancet Safety for HIV Test
kits 100 Pcs',Generic,300mg,'Test kit - Ancillary',99.099099,0.007431,0,'Inverness Japan','HIV
test - Ancillary'
FPQ-4487,'Direct Drop',EXW,11/13/09,'HIV 1 Uni-Gold Recombigen HIV Control Vial 2 x 0.5 ml',Uni-
Gold,300mg,'Test kit - Ancillary',1.001001,23.037365,64.94867,'Trinity Biotech Plc','HIV test -
Ancillary'
FPQ-15569,'Direct Drop',EXW,2/20/15,'Chase Buffer Determine 100 Tests 2.5ml x 1
Vial',Determine,300mg,'Test kit - Ancillary',0,3.715704,20.951184,'Alere Medical Co. Ltd.','HIV
test - Ancillary'
```

Models

bayes.NaiveBayes → Calculates the probability of a class based on feature independence assumptions and determines numeric estimator precision values from the training data

trees.J48 → Implements the C4.5 decision tree algorithm, which can generate pruned or unpruned decision trees

rules.OneR → Creates a classifier using the One Rule (1R) algorithm, which selects a single attribute that best predicts the target variable

rules.RandomForest → Builds a Random Forest, an ensemble of decision trees where each tree is trained on a random subset of the data and features

Results - Accuracy Percentage

		Attribute Selection Methods				
		ReliefF	Correlation	GainRatio	CfsSubset	Personal Selection
Models	NaiveBayes	96.61	85	94.19	92.74	81.94
	J48	99.68	99.19	99.35	97.26	99.35
	OneR	93.23	93.23	93.23	93.23	93.23
	RandomForest	99.19	99.68	99.68	97.58	99.68

Results - Best Model Error Comparison

	Attribute Selection Method & Model			
	ReliefF & J48	Correlation & Random Forest	Gain Ratio & Random Forest	Personal Selection & Random Forest
Mean Absolute Error	0.0018	0.01	0.0128	0.0053
Root Mean Squared Error	0.0328	0.0429	0.0486	0.0372
Relative Absolute Error (%)	1.0035	5.6323	7.187	2.9657
Root Relative Squared Error (%)	11.0072	14.4085	16.2953	12.4942

Analysis

- All of best models misclassified same two instances
    - Singular ‘ACT’ point
    - One out of two ‘Malaria’ points misclassified

- Attribute found in all 5 selection methods → ‘dosage form’
    - Good indicator of which type of shipment it is
    - Certain types of dosage are indicative of pediatric or adult shipments

Conclusion

Best Performance: ReliefFAttributeEval selection method, combined with the J48 algorithm

Outperformed other models regarding:

- Root Mean Squared Error
- Relative Absolute Error
- Root Relative Squared Error
- Mean Absolute Error

Good balance of performance considering underrepresented sub-classifications

How to Reproduce Our Best Model

1. Preprocessing
    a. Weka Preparation
    b. Missing and Hidden Value Correction
    c. Removing Redundant and Derived Columns
    d. Normalization by Scaling
    e. Stratified Sampling
    f. Train/Val/Test Split
2. Attribute Selection
    a. Conduct ReliefFAttributeEval
3. Classification
    a. J48

Next Steps and Future Studies

Building dataset with more “ACT” and “Malaria”

Further research regarding more advanced ML techniques for highly imbalanced datasets

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