SUBJECT: ISM 6359
DATA MINING

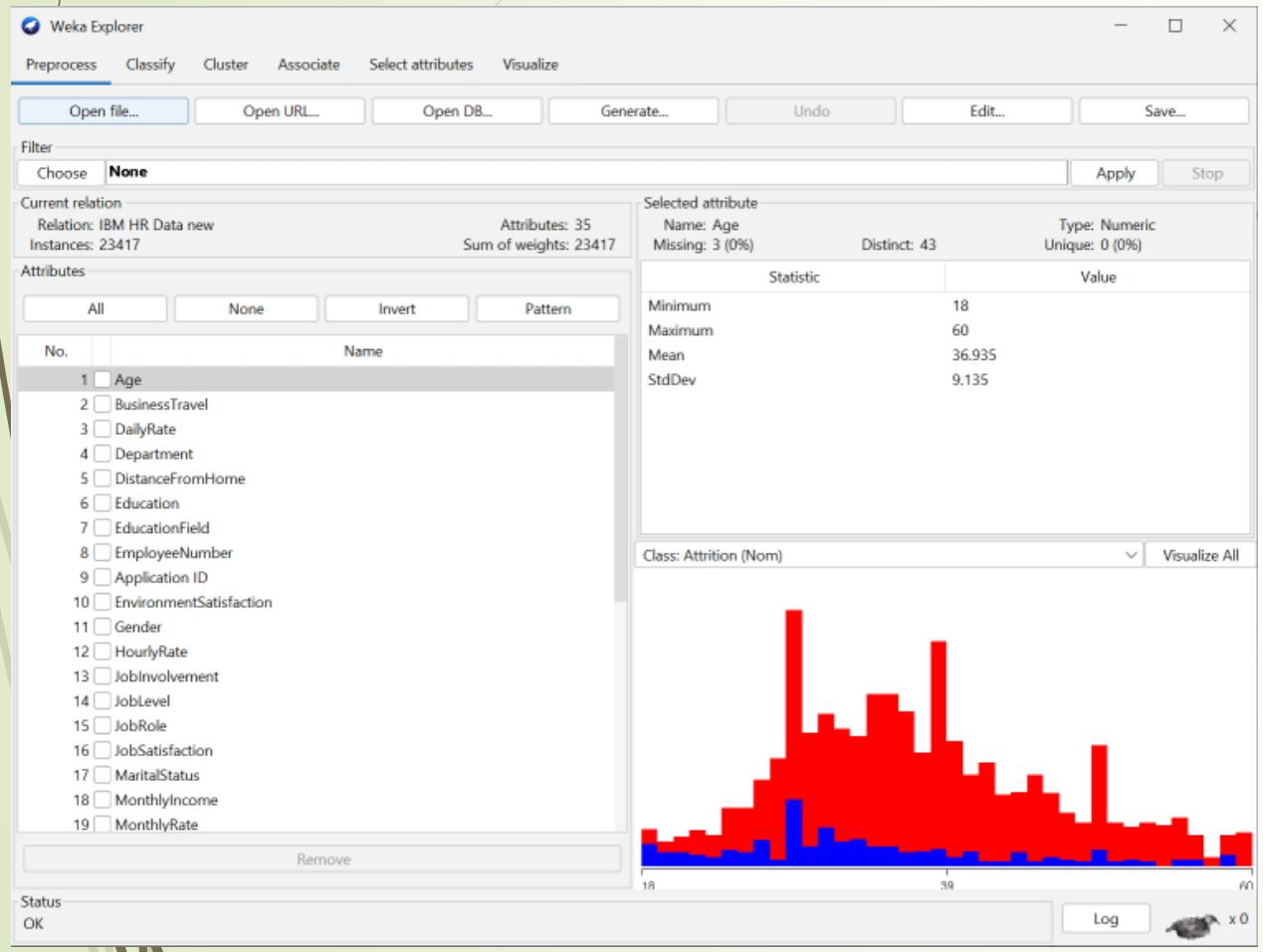
TOPIC: Employee Attrition
To build a machine learning model to predict the causes of Employee Attrition in IBM.

Tool: WEKA

BUSINESS REASON: Employee Attrition in IBM

- The goal of this project is to identify the causes of employee attrition with respect to IBM
- Human resource analytics (HR analytics) - applying analytic processes to the human resource department
- improving employee performance and getting a better return on investment
- HR department- act quickly to address the issue and stop attrition.

Tool: WEKA

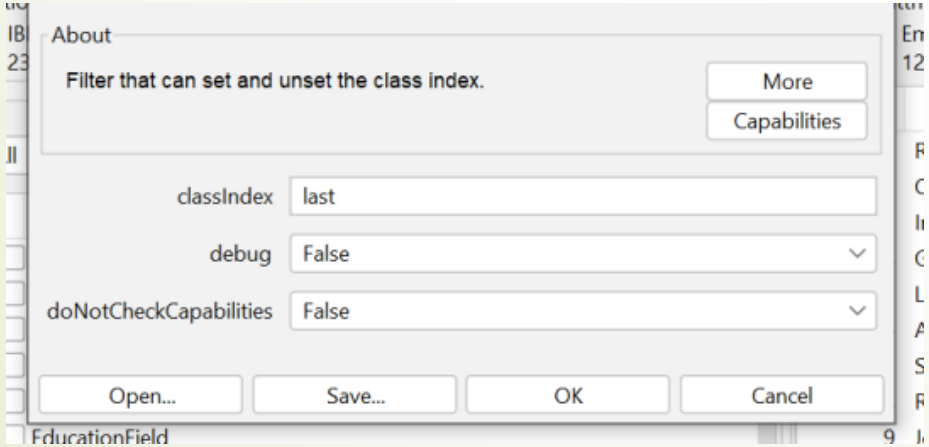


- WEKA is an open-source software provides tools for data preprocessing
- On the top, you will see several tabs as listed here:

- Preprocess
- Classify
- Cluster
- Associate
- Select Attributes
- Visualize

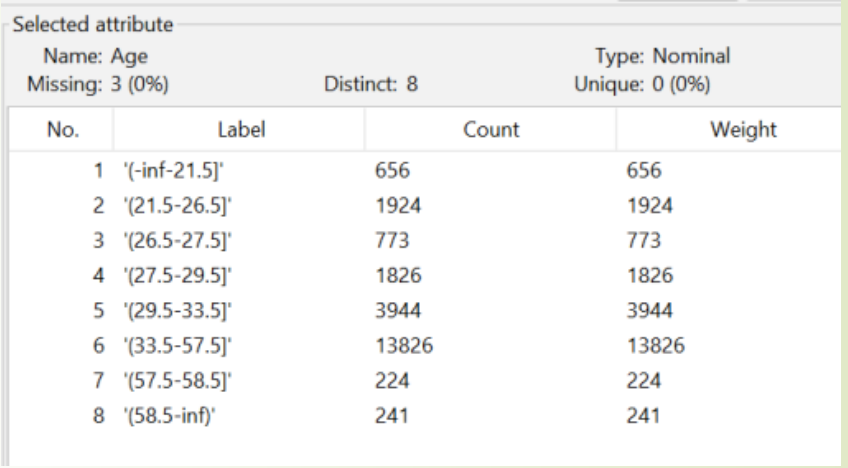
Data Preparation

- Loading the data
- Rename
- AddID
- Split data
- Cross Validation
- Set Role (ClassAssigner)



Data Preparation (continued)

- **Discretize -** Discretize by Binning for different attributes like Age, Monthly Income Frequency (**PKI Discretize**) for Education, Department. Although there is no option to discretize by User Specification. We can generate attribute by building an equation for the same.
- **RemoveMisclassified** in WEKA - Incorrectly classified instances, useful for removing outliers



Selected attribute

Name: Age Type: Nominal

Missing: 3 (0%) Distinct: 8 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-21.5]'	656	656
2	'(21.5-26.5]'	1924	1924
3	'(26.5-27.5]'	773	773
4	'(27.5-29.5]'	1826	1826
5	'(29.5-33.5]'	3944	3944
6	'(33.5-57.5]'	13826	13826
7	'(57.5-58.5]'	224	224
8	'(58.5-inf)'	241	241

Attribute Name	Values with Outliers	Values after removing outliers
Department	Sales, Research & Development, Human Resources, 1296	Sales, Research & Development, Human Resources
Gender	Male, Female, 1, 2	Male, Female
Marital Status	Single, Married, Divorced, 1, 2	Single, Married, Divorced
Overtime	Yes, No, Y	Yes, No
Employee source	Referral, Company Website, Indeed, GlassDoor, LinkedIn, Adzuna, Seek, Recruit.net, Jora, Test, 15, 1	Referral, Company Website, Indeed, GlassDoor, LinkedIn, Adzuna, Seek, Recruit.net, Jora

Data Preparation (continued)

- **Rename Nominal Values:** Here in Weka, we do not have Nominal to numeric. We require all numeric values for complex models like Neural Networks, SVM etc.

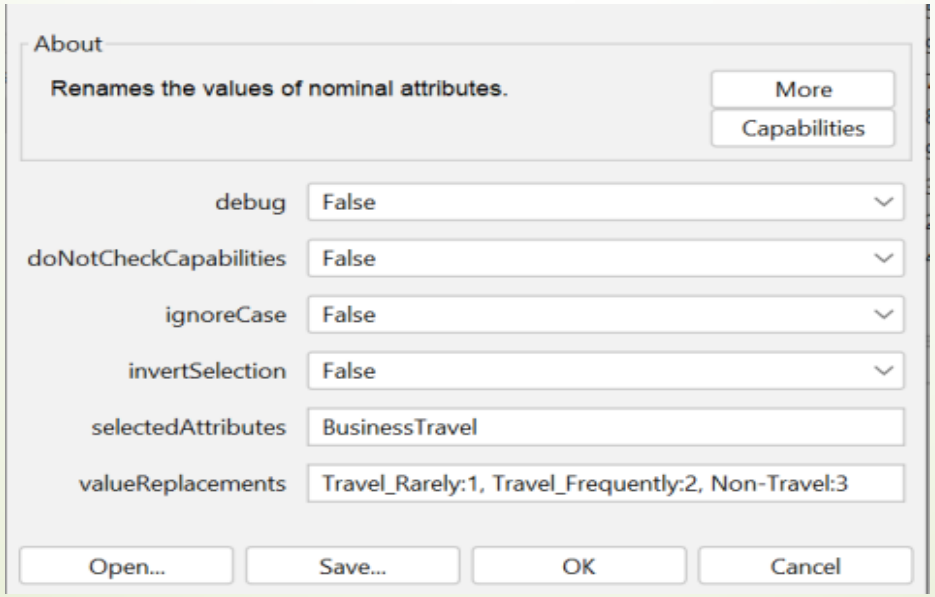
  Example: Age,Business travel, Department, JobRole

  JobSatisfaction, WorkLifeBalance,Employee source etc

- **NominaltoBinary-** convert Binary to numeric

  Gender, Overtime

- **Normalize** – huge magnitude of data to common numeric scale.



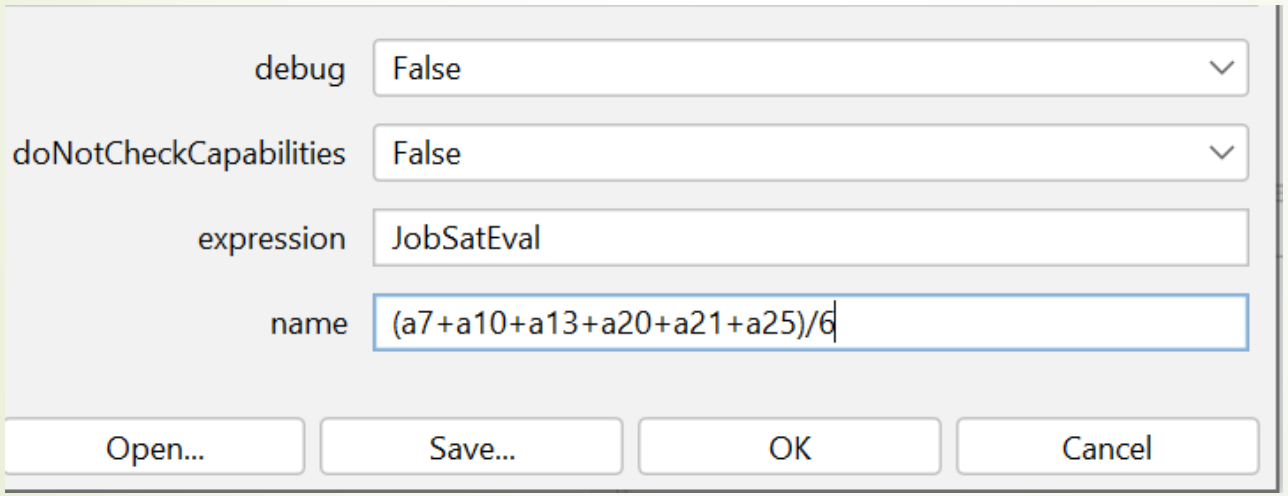
Data Preparation (continued)

- **Add Expression:** This works like generate attribute

  creates a new attribute by applying a mathematical expression

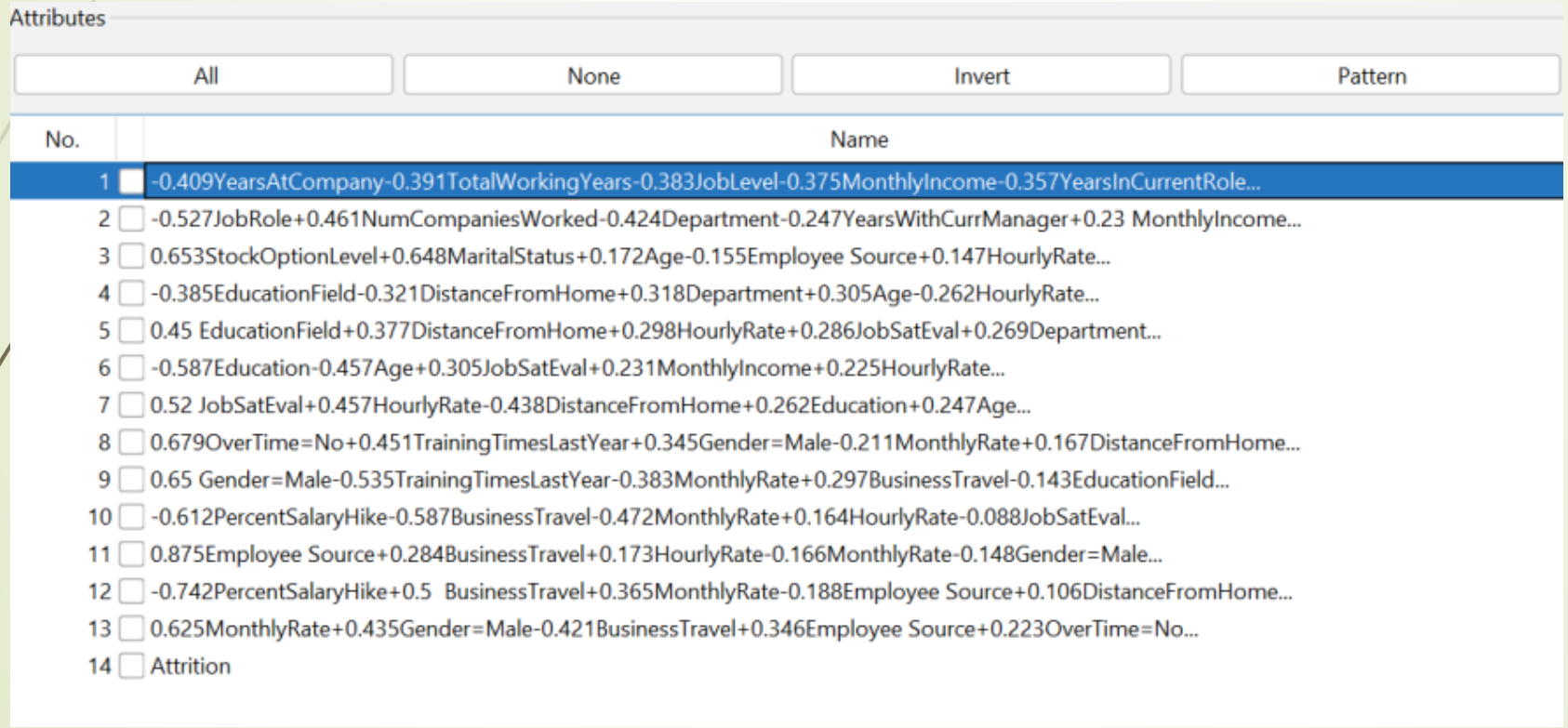
  to existing attributes

  JobSatisfaction, PerformanceRating, RelationshipSatisfaction, WorkLifeBalance, JobInvolvement, EnvironmentSatisfaction



Data Preparation (continued)

- **Select Attributes:** (Principal Component Attributes) In WEKA two ways:
  - Preprocessing stage
  - Select Attributes tab

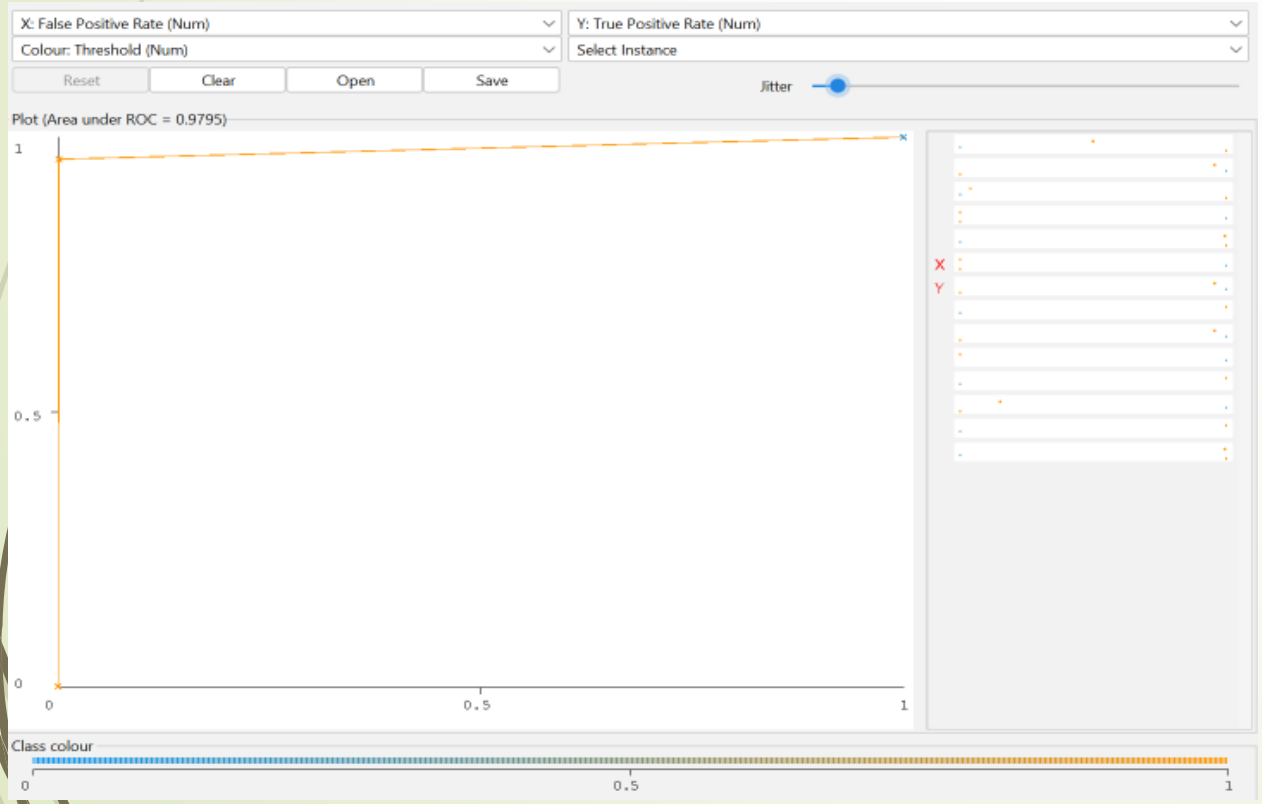


Accuracy of Correctly Classified Data in %

Model	Split Data 70%-30%	Cross-Validation 10 fold
Decision Tree	97.8078 %	98.868 %
Random Forest 1000 trees	98.7758 %	99.4192 %
Sequential Minimal Optimization (SVM)	84.2135 %	84.1696 %
K Nearest Neighbor K=20	99.3594 %	99.4876 %
Neural Network	85.3808 %	84.9639 %

Voting - Ensemble Method

Test Mode: Cross Validation – 10 folds
Correctly Classified Instances = 23258 which gave 99.321% accuracy
Incorrectly Classified Instances =159 with 0.679 %
Total Number of Instances = 23417




```
=== Confusion Matrix ===

     a     b   <-- classified as
  3557   150 |     a = Voluntary Resignation
     9 19701 |     b = Current employee
```

Analysis of the Output

1. From the above analysis, We found that cross validation has better accuracy than splitting the data.
   - Decision tree - 97.807%
   - Random Forest – 99.41%
   - SVM- 84.16%
   - KNN – 99.48%
   - Neural Network – 84.96%

2. Although the highest result is from Knn with K value 20, I have performed an Ensemble method to derive the best result for this model.
3. The Area under Curve is the highest for Knn Model which is 99.97%
4. From the Voting Algorithm I could determine that the Correctly Classified Instances have 99.321% accuracy and the Area under curve is 98%. This technique created multiple models and then combined them to produce improved results.
5. Every model makes a prediction (votes) for each test instance and the final output prediction is the one that receives more than half of the votes, this is the final prediction
6. After building the final model and determining the accuracy, this can be used by the organization to implement in determining whether a given employee is likely to resign or stay in the company.

3 W's

- **What went well?**

Topics taught in class could be easily related to the business problem I was solving.

Video lectures helped me simultaneously work on a new tool as we kept learning new features.

WEKA is an open source; I could find many learning materials from the University of Waikato which made it easier to understand the tool.

https://www.youtube.com/@WekaMOOC

- **What did not go well?**

The run time of the algorithms like SMO and Neural Network were taking too much heap space, at times the screen wasn't responding.

- **What would you use Differently Next Time?**

If given more time to explore, I would use new algorithms to know more about the tool.

Data Source

Initially I found this data on Kaggle.com

https://www.kaggle.com/datasets/rushikeshghate/capstone-projectibm-employee-attrition-prediction/code

The data is made available publicaly under the following license agreements:

https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem//?mhsrc=ibmsearch_a&mhq=attrition

https://github.com/IBM/employee-attrition-aif360/blob/master/data/emp_attrition.csv