

SEATTLE PACIFIC UNIVERSITY

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

Done By

Sheetal Murali

SPU ID no: 900409765

TABLE OF CONTENTS

SL NO	CONTENT	PAGE NO.
1.	Introduction	2
2.	Business Understanding	3
3.	Business Reason	4
4.	Data Understanding	4
5.	Data Preparation	5
6.	Data mining Algorithms	10
7	Analysis of the output	17
8	3 Ws	18
9	References	19

INTRODUCTION:

IBM is the global leader in business transformation through an open hybrid cloud platform and AI, serving clients in more than 170 countries around the world. Today 47 of the Fortune 50 Companies rely on the IBM Cloud to run their business, and IBM Watson enterprise AI is hard at work in more than 30,000 engagements. IBM has around **345,000** total number of employees in 2021 according to [macrotrends](#).

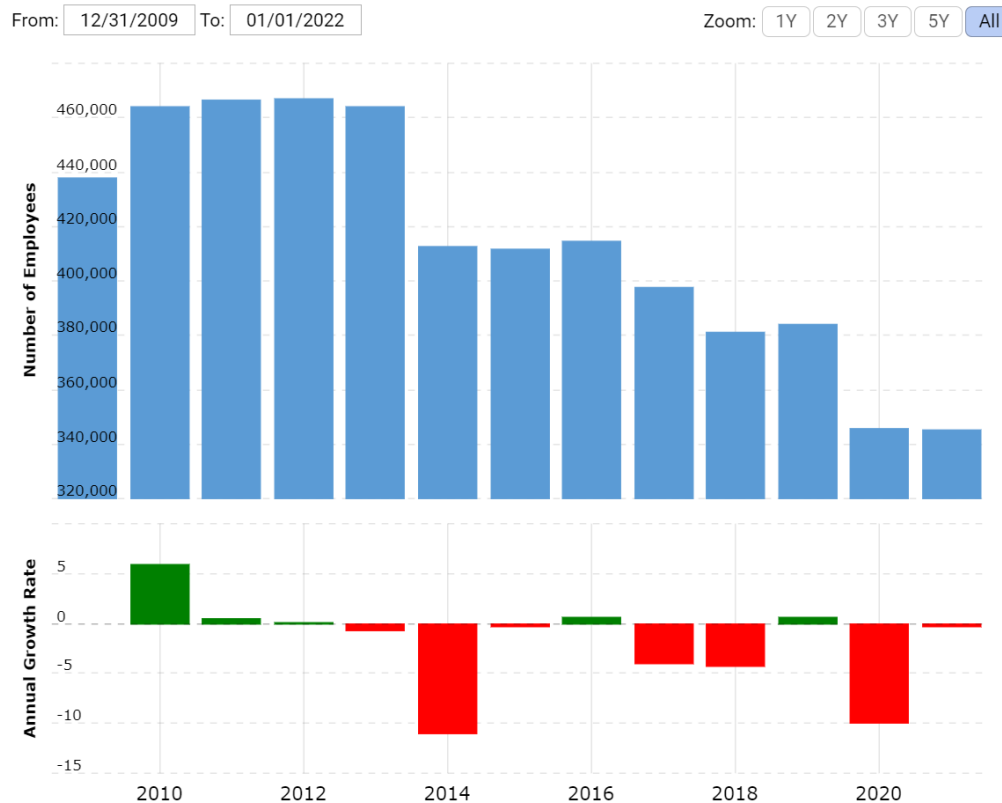
Employees are the backbone of any organization. Its performance is heavily based on the quality of the employees and retaining them. With employee attrition, organizations are faced with a few challenges:

1. Expensive in terms of both money and time to train new employees
2. Loss of experienced employees
3. Impact on productivity
4. Impact on profit

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency.

BUSINESS UNDERSTANDING

According to mactrotrends.com, We can see that IBM has a declining trend in the number of Employees over the years.



Interactive chart of IBM (IBM) annual worldwide employee count from 2010 to 2022 is shown as above,

IBM total number of employees in 2021 was 345,000, a 0.26% decline from 2020.

IBM total number of employees in 2020 was 345,900, a 9.87% decline from 2019.

IBM total number of employees in 2019 was 383,800, a 0.71% increase from 2018.

IBM total number of employees in 2018 was 381,100, a 4.2% decline from 2017.

Attrition in human resources refers to the gradual loss of employees over time. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture and motivation systems that help the organization retain top employees. Hence, it is important for the organization to uncover the factors that lead to employee attrition and retain the employees.

BUSINESS REASON:

IBM Employee Attrition Prediction

The goal of this project is to identify the causes of employee attrition in IBM through exploratory data analysis and analyze them using different classification models to determine whether an employee is likely to leave. This could significantly improve the HR department's capacity to act quickly to address the issue and stop attrition.

I have chosen WEKA as my tool to conduct this analysis.

WEKA is an open-source software that provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems.

DATA UNDERSTANDING:

The data collected is a high-level overview of what to expect in a data science pipeline and the tools that can be used along the way. It starts from framing the business question, to building and deploying a data model. The pipeline is demonstrated through the employee attrition problem.

First, I started with collecting raw data from the [IBM's official website](#) which had 1470 instances and 35 columns but I decided to explore more and find more data with relevance to the topic from [Kaggle](#). Here I could find a Capstone Project for Employee Attrition.

Data Structure: There are 37 features that describe each employee, their role in the company, his/her level of satisfaction, Education, salary, and share options available and their influence on the company.

Out of the above features, 2 features (Employee Number, Application ID) are redundant because they contain information that is not relevant to an employee's decision to leave the company. There are 2 other features (Over 18=yes, Standard Hours=80 hours) with 0 variance. i.e., the value is common to all hence they can be omitted.

This dataset is a supervised learning since the Attrition column is already given. This has 2 classes – Whether the employee is a current employee, or they have given voluntary resignation.

- Data: 23405 (rows) and 37 (Columns)
- Label: Attrition (Voluntary Resignation, Current Employee)
- Missing Values: 351
- Outliers: Yes
- Data type: Nominal, Numeric and String

DATA PREPARATION:

Once we have determined that we have the right data, then we move on to Data preparation. Sometimes we have too much data or too less data to generate the right model, hence we need the Data Preparation step. A lot of Data Mining projects spend almost 80% of the time on Data preparation. The Data preparation phase includes select the right amount of data for test and train, clean the data to reduce redundancy, deal with missing data, generate new attributes by examining the existing attributes.

For a machine learning algorithm to give acceptable accuracy, it is important to cleanse the data first. This is because the raw data collected from the field contains missing values, irrelevant columns, duplicate entries and so on

Loading the data:

When you click on the Explorer button in the Applications selector, On the top, you will see several tabs such as: • Preprocess • Classify • Cluster • Associate • Select Attributes • Visualize.

The data can be loaded from the following sources:

- Local file system
- Web
- Database

Data Preparation Performed:

1. **Rename:** This is an operator used to Name the attributes as we want. We change the name of attributes in the Parameters after using the Rename operator.
2. **Add ID:** Set an ID to the data set, which acts as an unique Identifier which we do not require hence the algorithm does not use it while calculating the data.
3. **Split Data/Cross Validation:** For my research I have chosen Supervised learning hence, I decided to try both Split data and Cross validation for multiple Algorithms.
 1. **Split data:** 70-30%
 2. **Cross Validation:** 5 Fold and 10 fold for all the algorithms

Test options

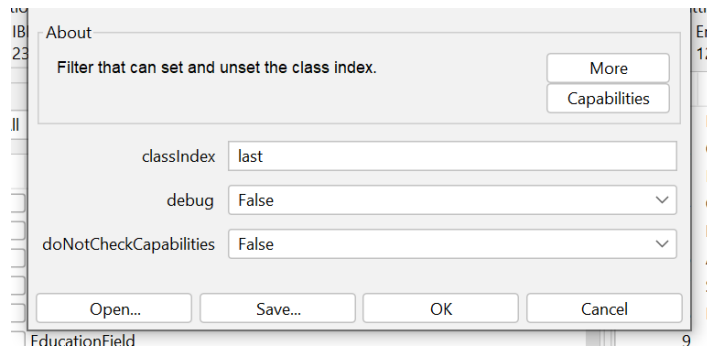
☐ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds
☒ Percentage split %

More options...

(Nom) Attrition ▼

Start Stop

4. **Set Role:** Here in WEKA, we call the set role as ClassAssigner, Where we set the Label for the data. Alternatively, WEKA uses default last Index as Class.



5. **Discretize:** In this project I have used Discretize by Binning for different attributes like Age, Monthly Income and 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.

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

6. **Select attributes:** Here we select the necessary Attributes which is used for the analysis. The algorithms only use the selected attributes which thus eliminates the irrelevant information from calculating. We might do data prep by getting rid of bad data first. The dataset contains 37 attributes of which few can be eliminated. For example, In this project we do not require Employee number, application ID, Over 18, standard hours (same for all), for the calculation, Hence these attributes can be eliminated.
7. **Normalize:** Normalization converts the value of numerical attributes to common scale. Normalizes all numeric values in the given dataset (apart from the class attribute if set). By default, the resulting values are in [0,1] for the data used to compute the normalization intervals.

8. **Filter outliers:** There are extreme value in the data set which will not fall into the bell curve and thus impact results. It is not recommended to Overfit the regression line because of an outlier as it cannot be used in general business problems. Hence this can be eliminated. I removed the following outliers.

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

9. **Nominal to Binary:** Converts all nominal attributes into binary numeric attributes.

Example: Overtime: yes=0, no=1 and

Gender: Female=0, Male=1

10. **Rename Nominal Values:** In Weka we do not have Nominal to numeric, instead we use Rename Nominal Attribute. We can easily convert those nominal values using the "Rename Nominal Value" filter by adding values.

About

Renames the values of nominal attributes.

More

Capabilities

debug False

doNotCheckCapabilities False

ignoreCase False

invertSelection False

selectedAttributes BusinessTravel

valueReplacements Travel_Rarely:1, Travel_Frequently:2, Non-Travel:3

Open... Save... OK Cancel

I converted the following attributes during Pre-Processed stage to generate accurate results:

- I. Age: '(-inf-21.5]':1, '(21.5-26.5]':2, '(26.5-27.5]':3, '(27.5-29.5]':4, '(29.5-33.5]':5, '(33.5-57.5]':6, '(57.5-58.5]':7, '(58.5-inf)':8
- II. Marital Status- Single:1, Divorced:2, Married:3
- III. BusinessTravel - Travel_Rarely:1, Travel_Frequently:2, Non-Travel:3
- IV. Department: Sales:1, Research & Development:2, Human Resources:3
- V. EducationField: Life Sciences:1, Technical Degree:2, Medical:3, Marketing:4, Other:5, Human Resources:6
- VI. JobRole: Sales Executive:1, Manager:2, Research Director: 3, Sales Representative:4, Laboratory Technician:5, Research Scientist: 6, Manufacturing Director:7, Healthcare Representative:8, Human Resources:9
- VII. Employee source- Referral:1, Company Website:2, Indeed:3, GlassDoor:4, LinkedIn:5, Adzuna:6, Seek:7, Recruit.net:8, Jora:9
- VIII. JobSatisfaction- Low:1, Medium:2, High:3, Very High:4
- IX. PerformanceRating: Low:1, Good:2, Excellent:3, Outstanding:4
- X. RelationshipSatisfaction: Low:1, Medium:2, High:3, Very High:4
- XI. WorkLifeBalance: Bad:1, Good:2, Better:3, Best:4
- XII. JobInvolvement: Low:1, Medium:2, High:3, Very High:4
- XIII. EnvironmentSatisfaction: Low:1, Medium:2, High:3, Very High:4
- XIV. Education: Below College:1, College:2, Bachelor:3, Master:4, Doctor:5

11. **Add Expression:** This works similar to generate attribute, where an instance filter that creates a new attribute by applying a mathematical expression to existing attributes. In this model, I used the following expression to create JobSatEval. This attribute is calculated by finding the average of similar attributes that affect job satisfaction which are:

- a. JobSatisfaction
- b. PerformanceRating
- c. RelationshipSatisfaction
- d. WorkLifeBalance,
- e. JobInvolvement,
- f. EnvironmentSatisfaction

debug: False

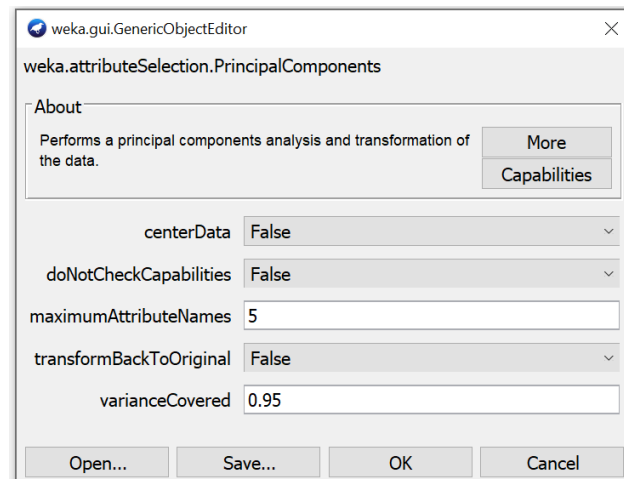
doNotCheckCapabilities: False

expression: JobSatEval

name: $(a7+a10+a13+a20+a21+a25)/6$

Buttons: Open..., Save..., OK, Cancel

12. **Principle Component Attribute:** After generating attributes, I still had 27 attributes which is a huge amount of data to process, hence, to further reduce attribute I used PCA (Feature Selection).



The Dimensionality reduction has ranked attributes based on their individual Evaluations. In this project the highest rank of an attribute had variance of 2.111. I used first 13 columns which has 50% threshold of the highest variance.

No.	Name
1	-0.409YearsAtCompany-0.391TotalWorkingYears-0.383JobLevel-0.375MonthlyIncome-0.357YearsInCurrentRole...
2	-0.527JobRole+0.461NumCompaniesWorked-0.424Department-0.247YearsWithCurrManager+0.23 MonthlyIncome...
3	0.653StockOptionLevel+0.648MaritalStatus+0.172Age-0.155Employee Source+0.147HourlyRate...
4	-0.385EducationField-0.321DistanceFromHome+0.318Department+0.305Age-0.262HourlyRate...
5	0.45 EducationField+0.377DistanceFromHome+0.298HourlyRate+0.286JobSatEval+0.269Department...
6	-0.587Education-0.457Age+0.305JobSatEval+0.231MonthlyIncome+0.225HourlyRate...
7	0.52 JobSatEval+0.457HourlyRate-0.438DistanceFromHome+0.262Education+0.247Age...
8	0.679OverTime=No+0.451TrainingTimesLastYear+0.345Gender=Male-0.211MonthlyRate+0.167DistanceFromHome...
9	0.65 Gender=Male-0.535TrainingTimesLastYear-0.383MonthlyRate+0.297BusinessTravel-0.143EducationField...
10	-0.612PercentSalaryHike-0.587BusinessTravel-0.472MonthlyRate+0.164HourlyRate-0.088JobSatEval...
11	0.875Employee Source+0.284BusinessTravel+0.173HourlyRate-0.166MonthlyRate-0.148Gender=Male...
12	-0.742PercentSalaryHike+0.5 BusinessTravel+0.365MonthlyRate-0.188Employee Source+0.106DistanceFromHome...
13	0.625MonthlyRate+0.435Gender=Male-0.421BusinessTravel+0.346Employee Source+0.223OverTime=No...
14	Attrition

I have uploaded the pre processed data in my GitHub Repository [here](#). This file can be downloaded and used to apply the classification algorithms.

DATA MINING ALGORITHMS:

For this analysis, I used the Classification method. The data collected is based on supervised learning. This dataset contains both input and output parameters, it is said to be labeled. In other words, the correct answer has already been assigned to the data. Here, I created a labeled data set which is the Attrition.

Classification is a commonly used technique for categorizing data points. It involves using algorithms that can be easily modified to improve the data quality. The primary goal of classification is to connect a variable of interest with Attrition.

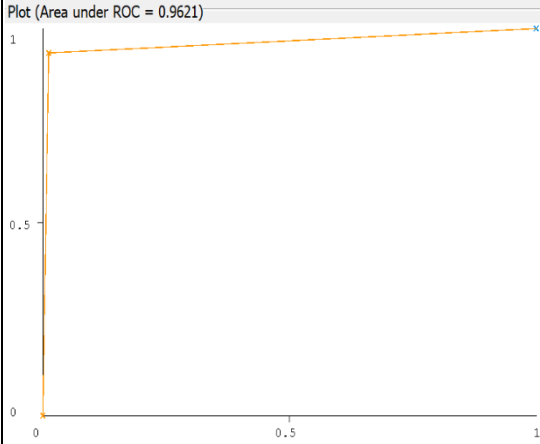
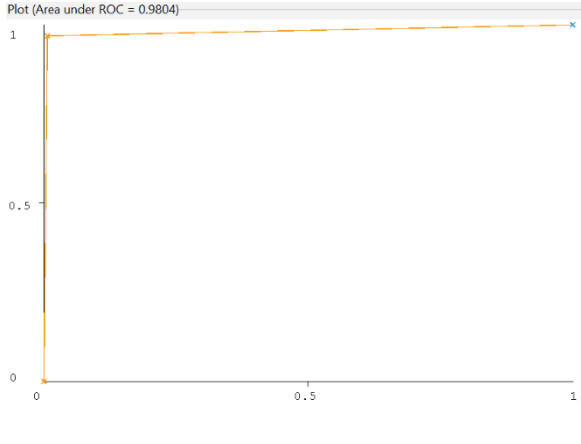
The algorithm establishes the link between the variables for prediction. The algorithm used for classification in data mining in WEKA is called the classifier, and observations made through the same are called the instances.

Firstly, I split data into Training set and Test set: 70-30 split.
Secondly, I used Cross Validation for better results: 5 folds and 10 folds.

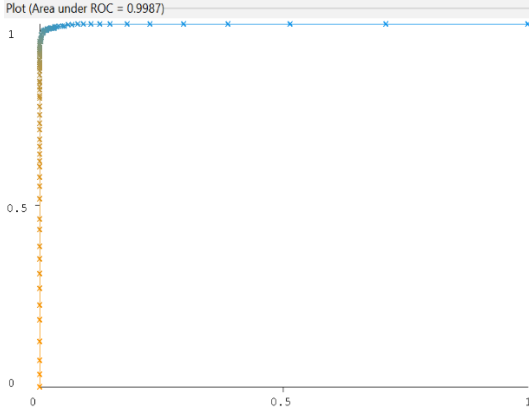
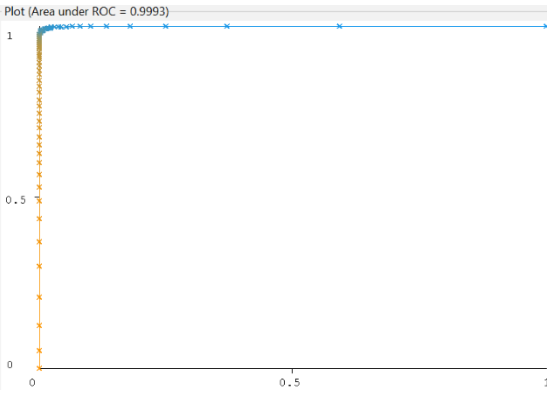
I have used the following algorithms to compare the performances:

- 1) Decision Tree (Random Tree in WEKA)
- 2) Random Forest
- 3) KNN
- 4) Support Vector Machine (SMO in WEKA)
- 5) Neural Network (Multilayer Perceptron in WEKA)
- 6) Ensemble method (Voting)

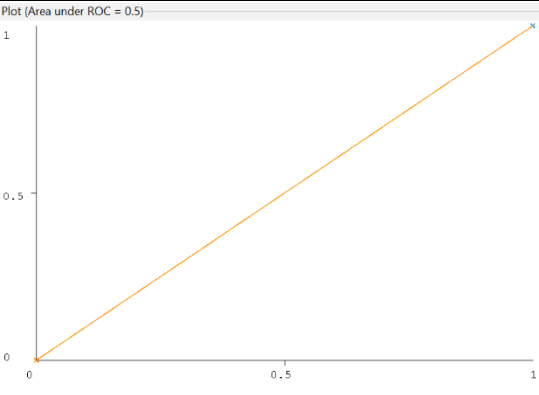
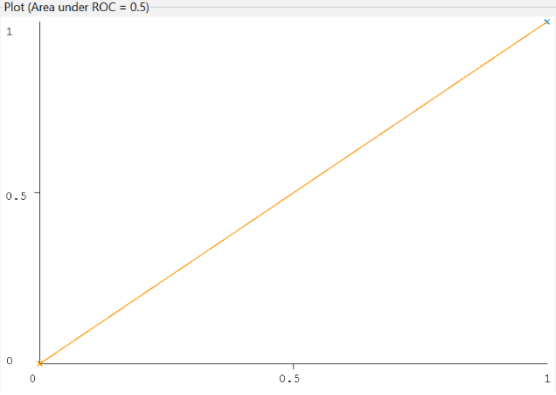
DECISION TREE

Performance	Split Data (70%-30%)	Cross-Validation (10 fold)
Correctly Classified Instances	97.8078 %	98.868 %
ROC Curve		
Confusion Matrix Classified	<pre> a b <-- classified as 1036 73 a = Voluntary Resignation 81 5835 b = Current employee </pre>	<pre> a b <-- classified as 3589 118 a = Voluntary Resignation 147 19563 b = Current employee </pre>

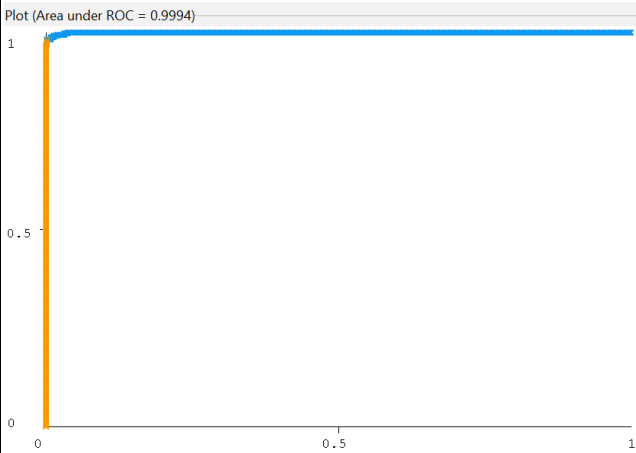
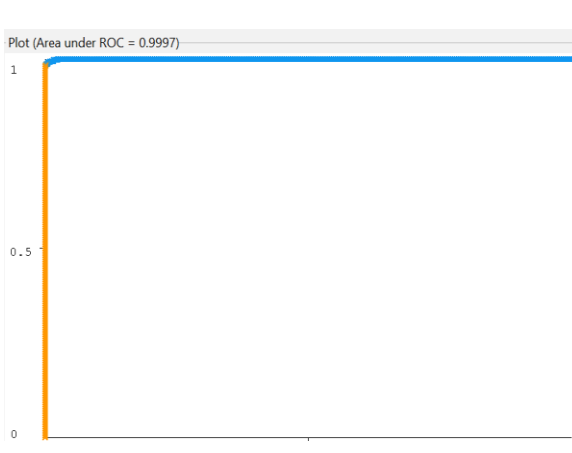
RANDOM FOREST – 1000 Trees

Performance	Split Data (70%-30%)	Cross-Validation (10 fold)
Correctly Classified Instances	98.7758 %	99.4192 %
ROC Curve		
Confusion Matrix	<pre> a b <-- classified as 1031 78 a = Voluntary Resignation 8 5908 b = Current employee </pre>	<pre> a b <-- classified as 3581 126 a = Voluntary Resignation 10 19700 b = Current employee </pre>

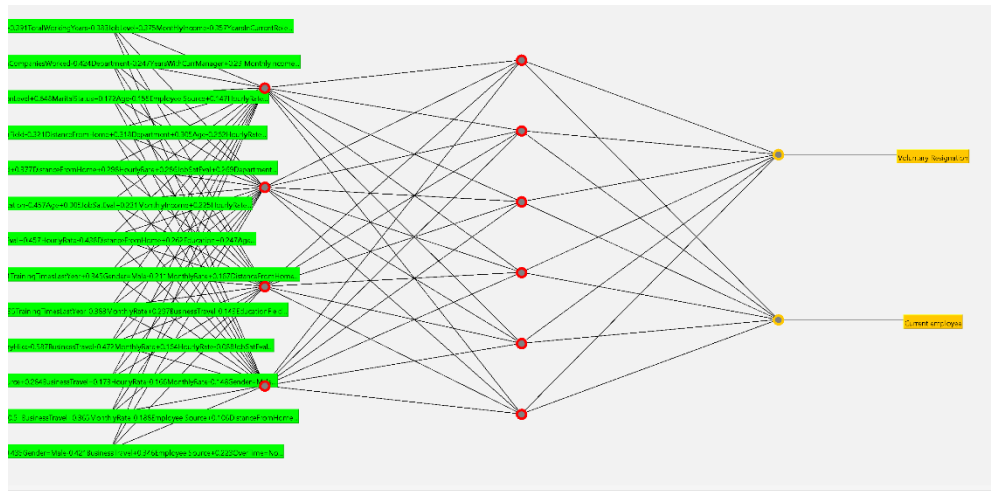
SMO : Sequential Minimal Optimization. SVM configuration in WEKA

Performance	Split Data (70%-30%)	Cross-Validation (10 fold)
Correctly Classified Instances	84.2135 %	84.1696 %
ROC Curve		
Confusion Matrix	<pre> a b <-- classified as 1031 78 a = Voluntary Resignation 8 5908 b = Current employee </pre>	<pre> a b <-- classified as 0 3707 a = Voluntary Resignation 0 19710 b = Current employee </pre>

K NEAREST NEIGHBOUR

Performance	Split Data (70%-30%)	Cross-Validation (10 fold)
Correctly Classified Instances	99.3594 %	99.4876 %
ROC Curve		
Confusion Matrix	<pre> a b <-- classified as 1086 23 a = Voluntary Resignation 22 5894 b = Current employee </pre>	<pre> a b <-- classified as 3647 60 a = Voluntary Resignation 60 19650 b = Current employee </pre>

Neural Network



Performance	Split Data (70%-30%)	Cross-Validation (5 fold)
Correctly Classified Instances	85.3808 %	84.9639 %
ROC Curve	Plot (Area under ROC = 0.6964) 	Plot (Area under ROC = 0.6967)
Confusion Matrix	<pre> a b <-- classified as 277 832 a = Voluntary Resignation 195 5721 b = Current employee </pre>	<pre> a b <-- classified as 860 2847 a = Voluntary Resignation 674 19036 b = Current employee </pre>

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

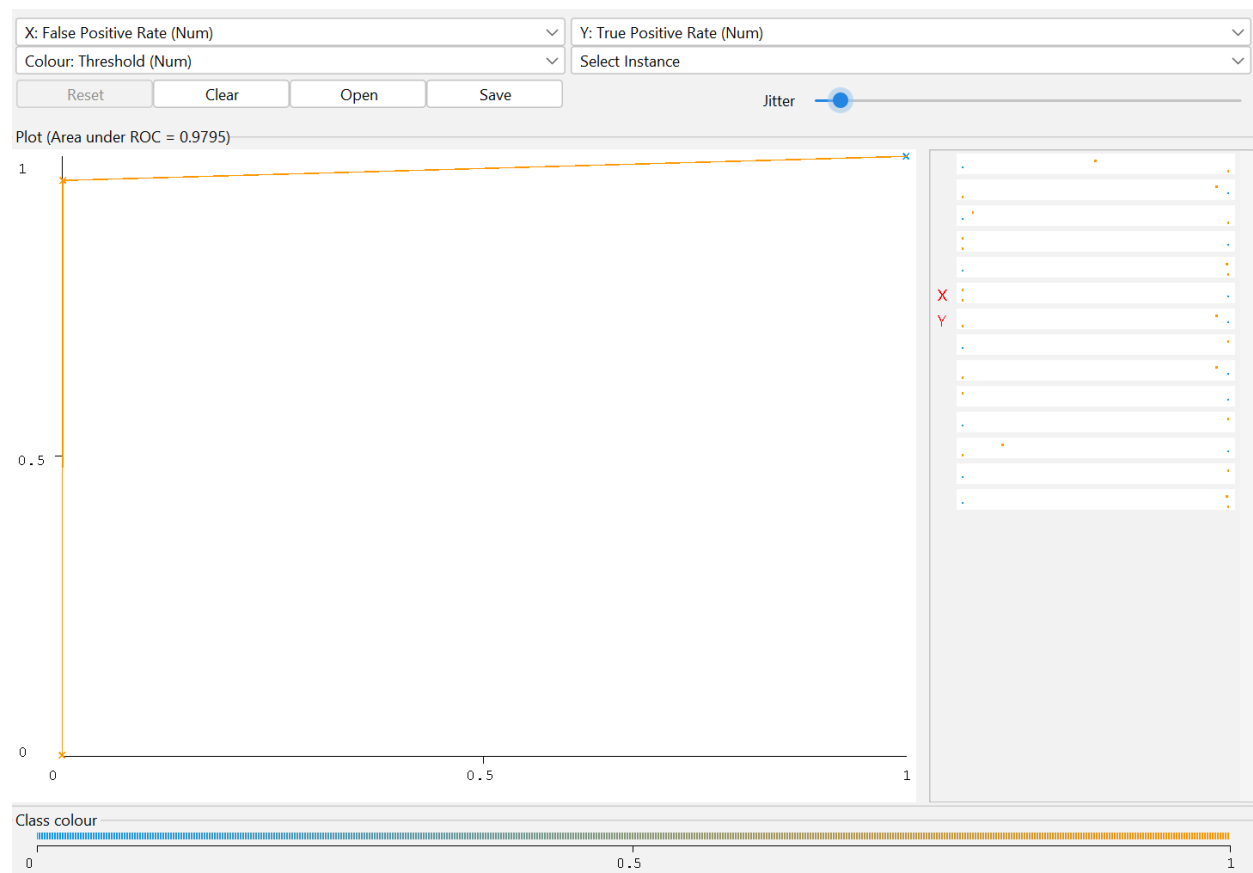
```
=== Confusion Matrix ===
```

```

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

```

ROC Curve:



ANALYSIS OF THE OUTPUT

- 1) From the above analysis, we found that cross validation has better accuracy than splitting 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 model can be used by the organization in determining whether a given employee is likely to resign or stay in the company.

3 W's

1) 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.

2) 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.

3) What would you use Differently Next Time?

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

REFERENCES

<https://www.ibm.com/us-en?ar=1>

[IBM: Number of Employees 2010-2022 | IBM | MacroTrends](#)

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

<https://github.com/sheetalmurali/DataMining/blob/main/preprocessData.csv.arff>

Dataset source:

The data is made available publicly 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

Further, I found the dataset used to build this model on Kaggle.com

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