PATTERN RECOGNITION!

S.Natchanon (MSc.)

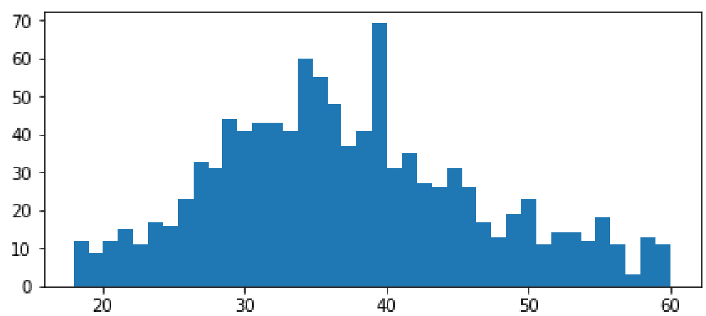
HOMEWORK2

6470177521

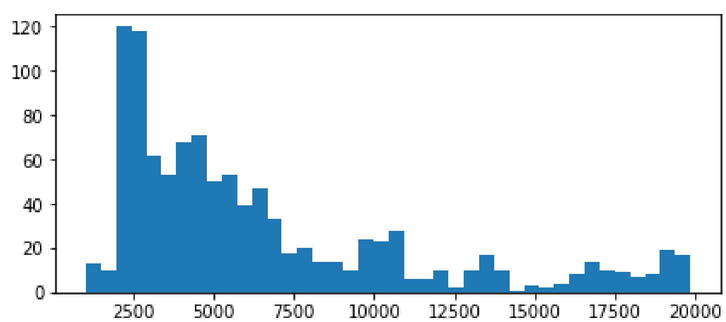
PART: Employee Attrition Prediction

T4: Observe the histogram for “Age, MonthlyIncome, DistanceFromHome” How many bins have zero counts? Do you think is a good discretization?

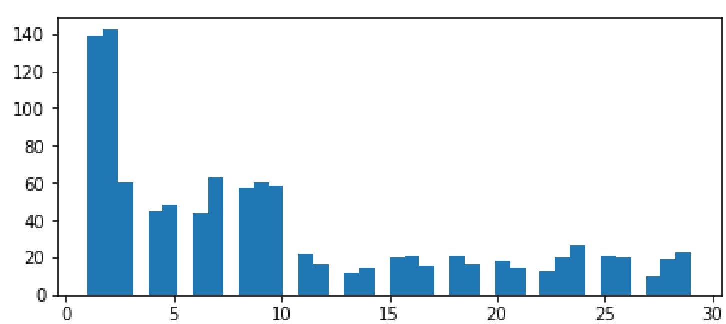
- **Ans:** 11 bins from DistanceFromHome have zero counts. This discretization is using by too many bins, that affect to some features which has small number of unique value (or too small size of data) shows a zero bin.



Age



MonthlyIncome



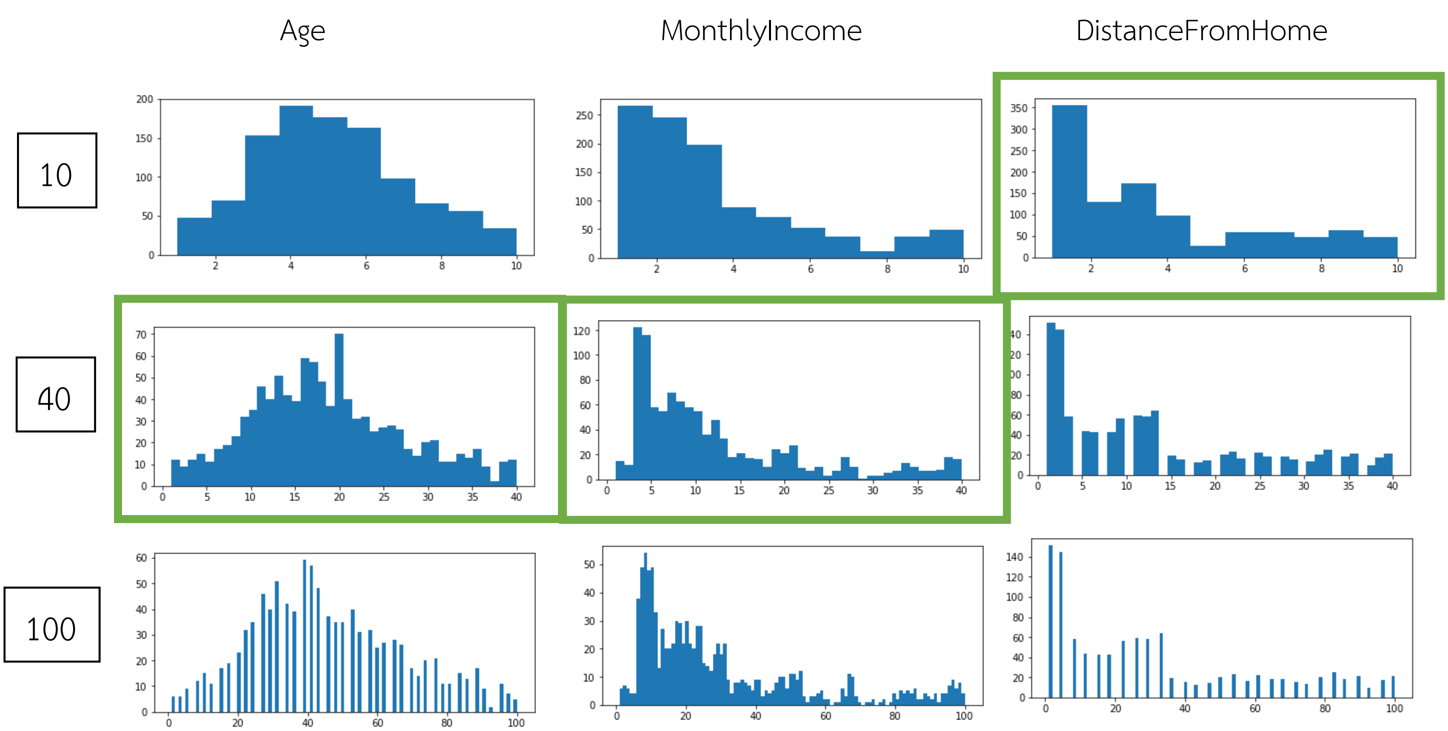
DistanceFromHome

T5: Can we use a Gaussian to estimate this histogram?

- **Ans**: Yes, we can (for some features). But this strategy may give us a less accurate than a straightforward method like counting histogram prob because we assume all data fit in gaussian curve even some are not i.e., Age -> look like gaussian, but other features is something else. However, GMM may be the best solution here because it mixed of many gaussian, and this mixed distribution can fit many distributions shape.

T6: histogram of 10, 40, and 100 bins. Which bin size is most sensible for each feature?

- **Ans:** More bin size gives more detailed information but in one condition that is it must not contain many zero bin (None zero bin is best).



T7: Which features should be discretized? And What are the criteria?

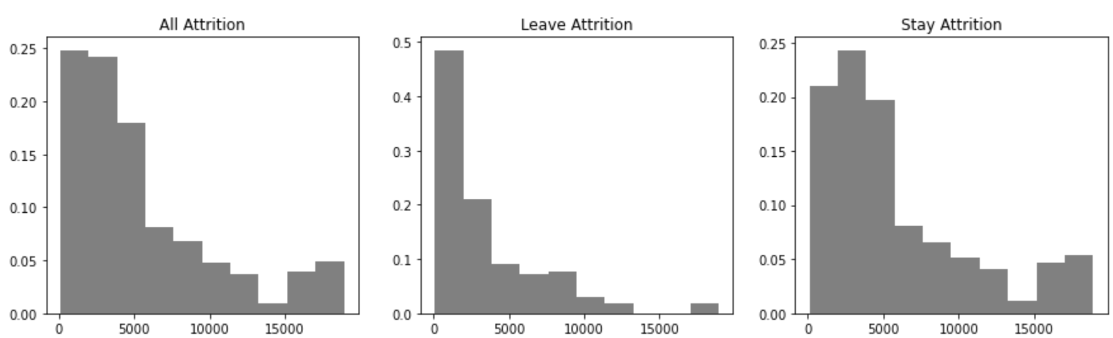
- **Ans**: Discretized features are shown in below picture. And my criteria are selecting feature that has continuous data or has many unique possible values (In this work, I choose the one that has unique value greater than 10)

```
Age  : unique -->  43
DailyRate  : unique -->  773
DistanceFromHome  : unique -->  29
HourlyRate  : unique -->  71
MonthlyIncome  : unique -->  1105
MonthlyRate  : unique -->  1143
PercentSalaryHike  : unique -->  15
TotalWorkingYears  : unique -->  39
YearsAtCompany  : unique -->  35
YearsInCurrentRole  : unique -->  19
YearsSinceLastPromotion  : unique -->  16
YearsWithCurrManager  : unique -->  18
```

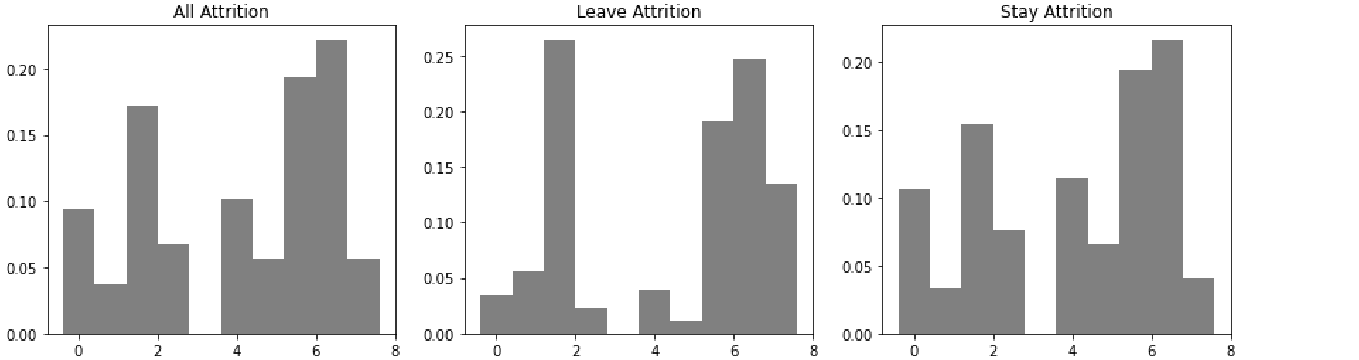
T8: What kind of distribution should we use to model histograms? What is MLE for the likelihood distribution? Plot the likelihood distributions for different “Attrition” values | bins= 10

- **Ans:** Multinomial Distribution (Prove is shown in very last page)

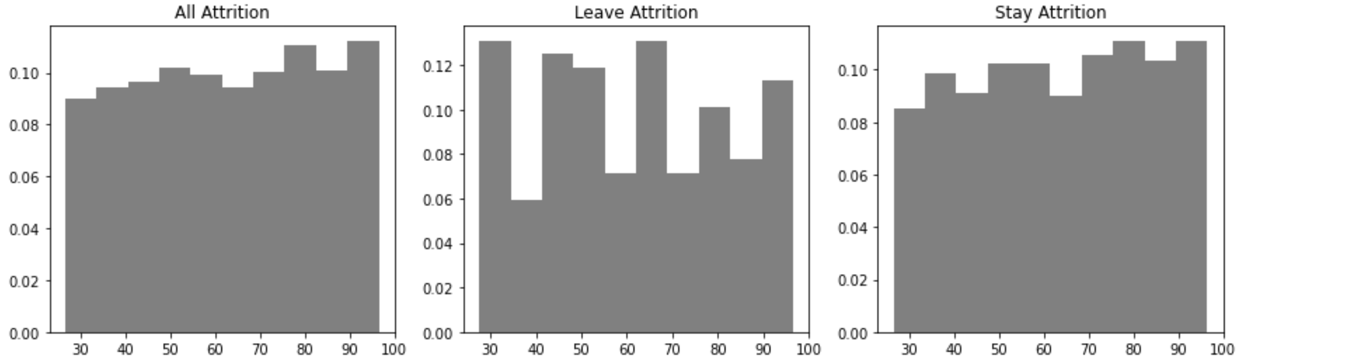
MonthlyIncome



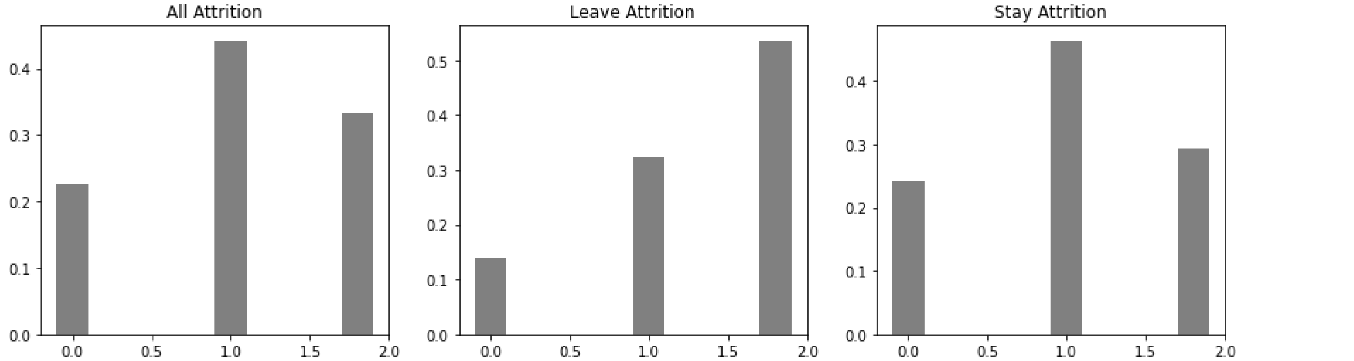
JobRole



HourlyRate

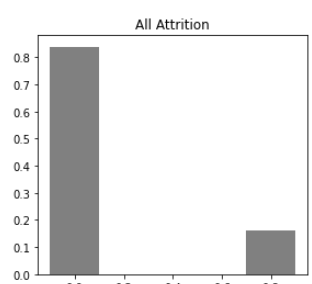


MaritalStatus



T9: What is the prior distribution of the two classes?

- **Ans:** “Leave” p = 0.161, “Stay” p = 0.839.



T10: Propose a method to fix zero encounter problem?

- **Ans:** Using flooring (add some tiny floating point to zero) strategy to make all product term not being zero.

T11: Implement your Naïve Bayes and report the Accuracy, Precision, Recall, and F1 score for this model.

- **Ans:** Implementation detail is attached as .ipynb and .py files.

```
Accuracy :  0.8513513513513513
Precision :  0.5384615384408284
Recall :  0.5833333333090278
F1 Score :  0.5599999994784
```

T12: Report the result from Gaussian pdf?

- **Ans:** Implementation detail is attached as .ipynb and .py files.

```
Accuracy :  0.8040540540540541
Precision :  0.40740740739231823
Recall :  0.4583333333142361
F1 Score :  0.4313725485044214
```

BASELINE COMPARISON

T13: Report the result from Random choice baseline?

- **Ans:** Implementation detail is attached as .ipynb and .py files.

```
Accuracy :  0.49324324324324326
Precision :  0.18518518518289895
Recall :  0.6249999999739584
F1 Score :  0.2857142853561905
```

T14: Report the result from Majority rule baseline?

- **Ans:** Implementation detail is attached as .ipynb and .py files.

```
Accuracy :  0.8378378378378378
Precision :  0.0
Recall :  0.0
F1 Score :  0.0
```

T15: Compare the two baselines with your Naïve Bayes classifier.

- **Ans:** The Naïve Bayes classifier outperform both baseline in both Accuracy and F1 score

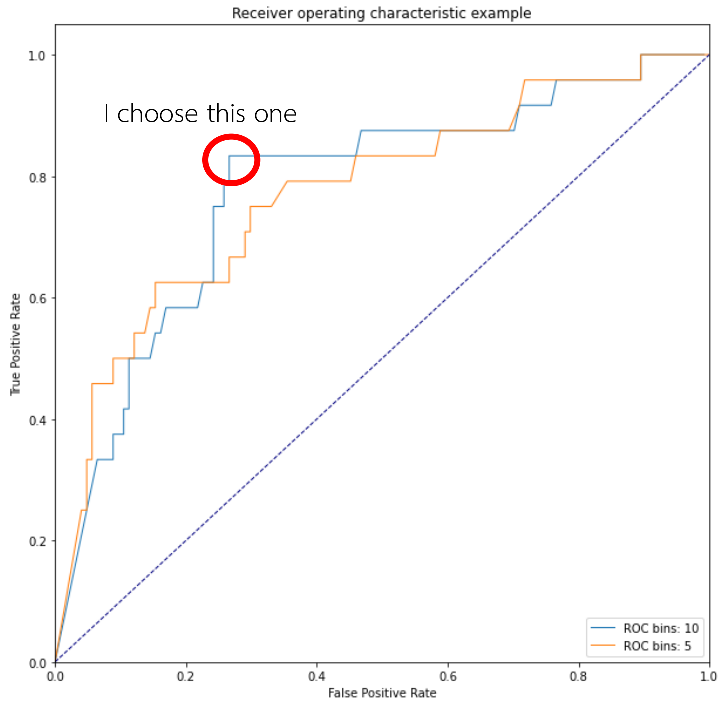
T16: Find the best acc, and F1 by threshold finding.

- **Ans:** Best threshold that gives the best accuracy (0.865) is 0.65

Best threshold that gives the best F1 score (0.57) is 0.099

T17: Plot RoC of your classifier.

- **Ans:**



T18: Change bins = 5. What happens to the RoC curve? Which discretization is better?

- **Ans:** At the same False Positive Rate (FPR) bins(5) has slightly better True Positive Rate (TPR) but doing worsen later. This one quite ambiguous to determine which one is better; it's depending on what application you are going to applied.

In this work (Leave or Stay) if I were CEO, I wouldn't want to let my employee leave. It's okay that I will get higher FPR but there is no downside if it false positive. So, I prefer more TPR which more FPR. I will use bins(10) model.

T19: Submit your code

- **Ans:** Sure!!

OT3: Shuffle the database for 10 times, calculate the mean and variance of accuracy.

- **Ans:**

```
Setting  1   : Accuracy = 0.764 , F1 = 0.407
Setting  2   : Accuracy = 0.804 , F1 = 0.453
Setting  3   : Accuracy = 0.845 , F1 = 0.566
Setting  4   : Accuracy = 0.736 , F1 = 0.381
Setting  5   : Accuracy = 0.845 , F1 = 0.549
Setting  6   : Accuracy = 0.709 , F1 = 0.218
Setting  7   : Accuracy = 0.764 , F1 = 0.462
Setting  8   : Accuracy = 0.770 , F1 = 0.433
Setting  9   : Accuracy = 0.831 , F1 = 0.590
Setting  10   : Accuracy = 0.804 , F1 = 0.508


Mean accuracy for 10 shuffles is 0.7872
Variance accuracy for 10 shuffles is 0.0437

Mean F1 score for 10 shuffles is 0.4567
Variance F1 for 10 shuffles is 0.0437
```

NOTE:

Code is provided in both .ipynb and .py

.ipynb contains all detail and implementation step for this work

.py is a code for running Naïve Bayes model including baseline and shuffle folds

Simply run ` **python NaïveBayes.py** `