# RATING PREDICTION
## SUPERMARKET SALES DATASET

GROUP 6

By Naufal Rasyid Sutansyah(2502006202), DataScience, naufal.sutansyah@binus.ac.id

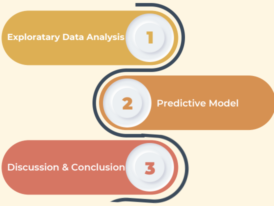
## INTRODUCTION

The growth of supermarkets in most populated cities are increasing and market competitions are also high. The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches for 3 months data. Predictive data analytics methods are easy to apply with this dataset.

## DATASET DESCRIPTION

- Invoice id: sales slip invoice
- Branch: 3 branches are available A, B and C
- City: Location
- Customer type: Customers member card/Normal
- Gender: Gender
- Product line: General item categorization groups
- Unit price: Price of product in $
- Quantity: Number of products purchased
- Date: Date of purchase from January 2019 to March 2019
- Time: Purchase time 10am to 9pm
- Payment: Payment (Cash, Credit card and Ewallet)
- COGS
- Gross margin percentage
- Gross income
- Rating: Customer rating on a scale of 1 to 10
- Tax: 5% tax fee for customer
- Total: Total price

## METHODS

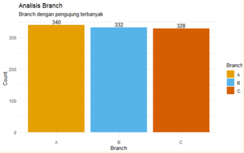


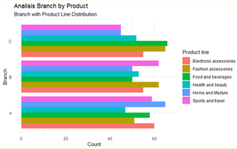
Exploratary Data Analysis 1

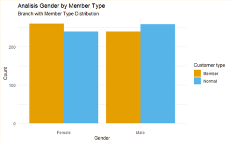
2 Predictive Model

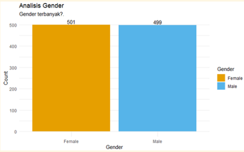
Discussion & Conclusion 3

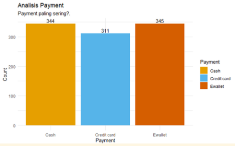
## EXPLORATARY DATA ANALYSIS

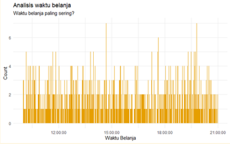








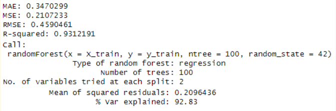




## PREDICTIVE MODELING →

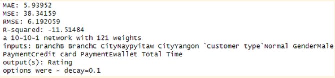
Using 3 Modeling to Predict customer rating (Rating ~ branch + city + customer + gender+ Total + Payment + Time

- **Random Forest**



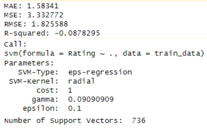
```
MAE: 0.3470299
MSE: 0.2107233
RMSE: 0.4590461
R-squared: 0.9312191
Call:
 randomForest(x = X_train, y = y_train, ntree = 100, random_state = 42)
              Type of random forest: regression
                    Number of trees: 100
No. of variables tried at each split: 2
          Mean of squared residuals: 0.2096436
                    % Var explained: 92.83
```

- **Neural Network**



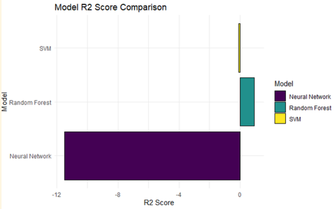
```
MAE: 5.93952
MSE: 38.34159
RMSE: 6.192059
R-squared: -11.51484
a 10-10-1 network with 121 weights
inputs: BranchB BranchC CityNaypyitaw CityYangon `Customer type`Normal GenderMale
PaymentCredit card PaymentEwallet Total Time
output(s): Rating
options were - decay=0.1
```

- **SVM**



```
MAE: 1.58341
MSE: 3.332772
RMSE: 1.825588
R-squared: -0.0878295
Call:
svm(formula = Rating ~ ., data = train_data)
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.09090909
    epsilon:  0.1
Number of Support Vectors:  736
```

- **Comparison For All Modeling**



## DISCUSSION

- Random Forest: This model has the highest R-squared score, indicating a good fit to the data. A higher R-squared score (closer to 1) suggests that the model explains a larger proportion of the variance in the target variable (customer rating).
- Neural Network: This model has a negative R-squared score, which means it performs poorly in explaining the variance in the target variable. A negative R-squared indicates that the model does not capture the patterns and relationships in the data well.
- SVM: The SVM model also has a negative R-squared score, indicating poor performance in explaining the variance in the target variable.

## CONCLUSION

Random Forest model proves to be the most effective in predicting customer ratings. It exhibits superior accuracy, as indicated by the lowest MAE, MSE, and RMSE values among the three models. Moreover, with an impressive R-squared value of 0.9312191, the random forest model demonstrates a strong fit to the data, explaining approximately 93.12% of the variance in the ratings. Therefore, the random forest model is highly recommended for accurate customer rating predictions.