มหาวิทยาลัยสงขลานครินทร์

MACHINE LEARNING PROJECT

ONLINE TRANSACTION DATA ANALYSIS AND FORECASTING FOR PRINCE OF SONGKHLA UNIVERSITY COOPERATIVE CREDIT AND SAVING, LIMITED

Canva

HTTPS://WWW.CANVA.COM/DESIGN/DAGT1J1-C84/BGZ2IX7LCSMSBHWMT4AITG/VIEW?UTM_CONTENT=DAGT1J1-C84&UTM_CAMPAIGN=DESIGNSHARE&UTM_MEDIUM=LINK&UTM_SOURCE=EDITOR

Wathunyu Phetpaya
Code 6710120039, Computer Engineering

Outline

มหาวิทยาลัย สงขลานครินทร์

HTTPS://GITHUB.COM/WATHUNYU123/ML

Problem	3
Data Transformation	4
Data Distribution	5 - 7
Data Preparation	8
Data Training	9 - 11
Result	12

Problem : Background

Emphasizing the need for capacity planning to accommodate growth

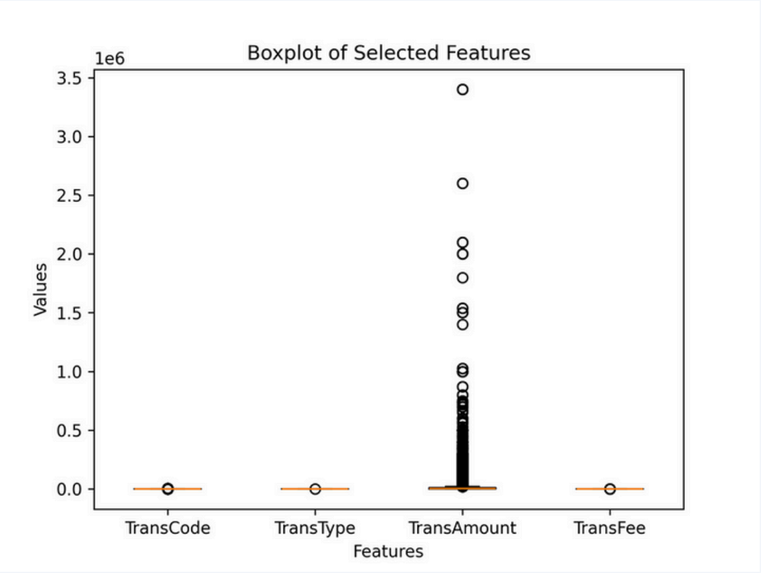
- The savings cooperative is experiencing exponential growth in its mobile app usage as members increasingly adopt digital banking services. The organization needs to develop a robust strategy to accommodate this surge in demand and ensure seamless user experiences.

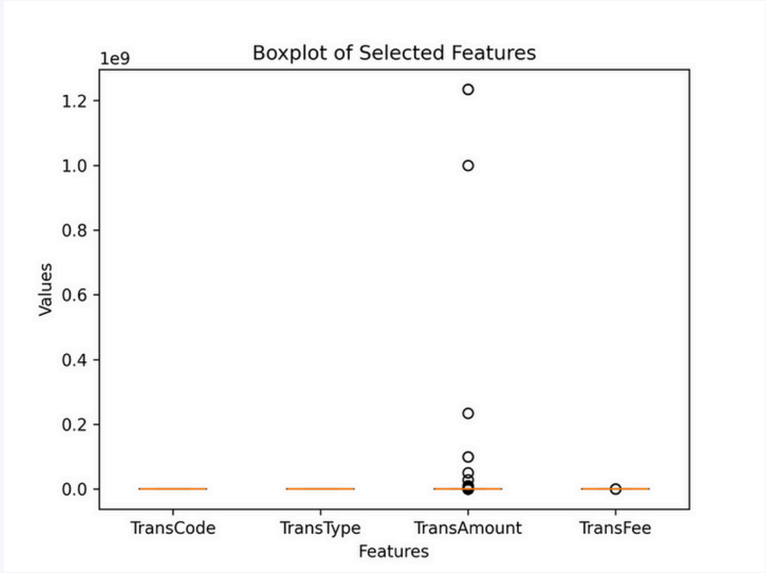
Data Transformation

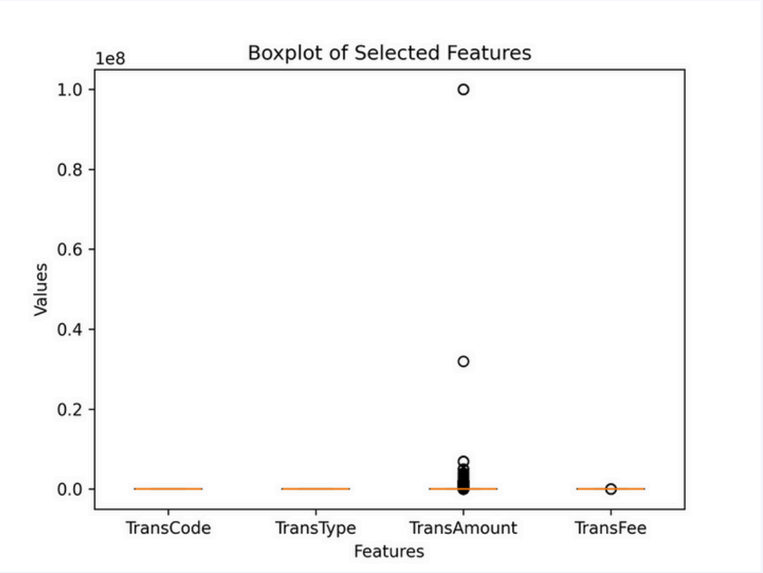
```python
data2567, data2566, data2565, data2564 = import_data(path, files, '2567'), import_data(path, files, '2566'), import_data(path, files, '2565'), import_data(path, files, '2564')
datasets = [
    ("data2567", data2567),
    ("data2566", data2566),
    ("data2565", data2565),
    ("data2564", data2564),
]
show_detail(
    datasets,
    shape=True,
    column=True,
    info=True,
    describe=True,
    is_null=True,
    dtype=True,
)

# Columns to be plotted
columns = ['TransCode', 'TransType', 'TransAmount', 'TransFee']
labels = ['TransCode', 'TransType', 'TransAmount', 'TransFee']
# Iterate through each dataset
for label, df in datasets:
    # Save the boxplot
    plot_boxplot(df, columns, labels, filename=f'{image_path}boxplot_{label}.png')

plot_histograms(
    datasets=datasets,
    column='TransAmount',
    filename=f'{image_path}hist_{datasets[0][0]}_{datasets[1][0]}_{datasets[2][0]}_{datasets[3][0]}{png}'
)

# Iterate through each dataset
for label, df in datasets:
    # Convert Tran_Date to datetime and extract the month and year
    df['Tran_Date'] = pd.to_datetime(df['Tran_Date'], format='%Y%m%d')
    df['YearMonth'] = df['Tran_Date'].dt.to_period('M')

    # Get the unique TransCode counts by month
    transcode_counts = df.groupby(['YearMonth', 'TransCode']).size().unstack(fill_value=0)

    # Save the DataFrame as an image
    save_dataframe_as_image(df=transcode_counts, filename=f'{image_path}transcode_counts_{label}{png}')
```

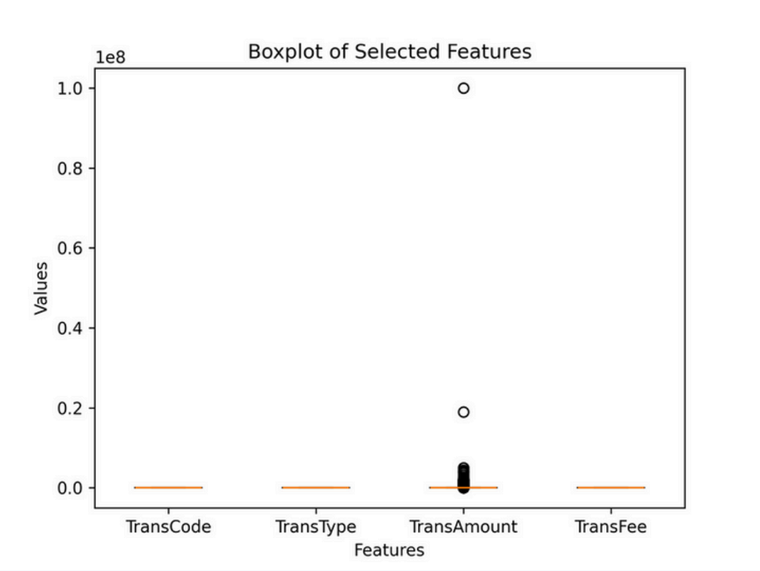
- Convert .xlsx to Dataframe
- Convert string type of datetime to datetime type
- Group age to 6 range
  - 0-18
  - 19-30
  - 31-40
  - 41-50
  - 51-60
  - 61+

Data Distribution

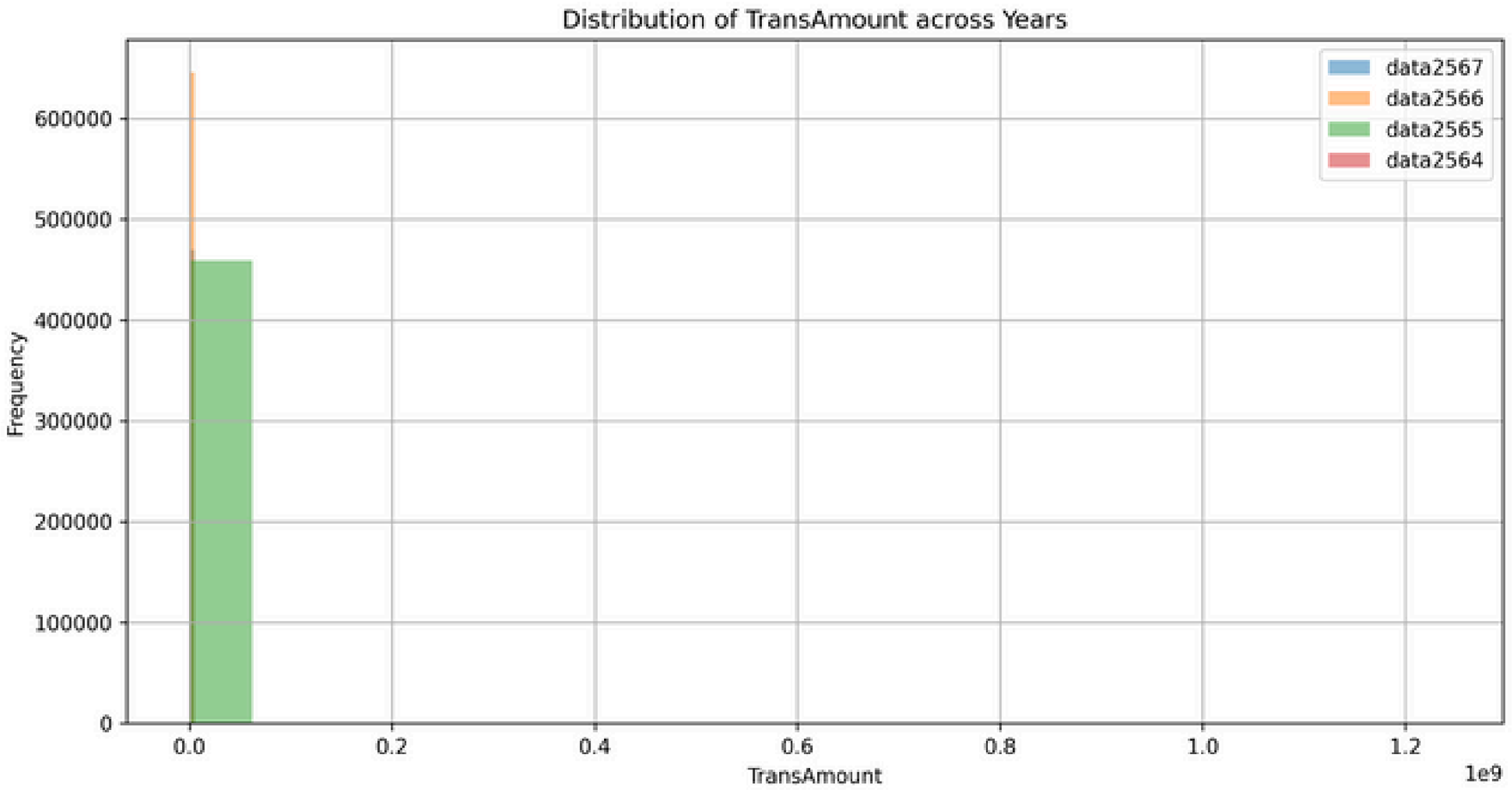




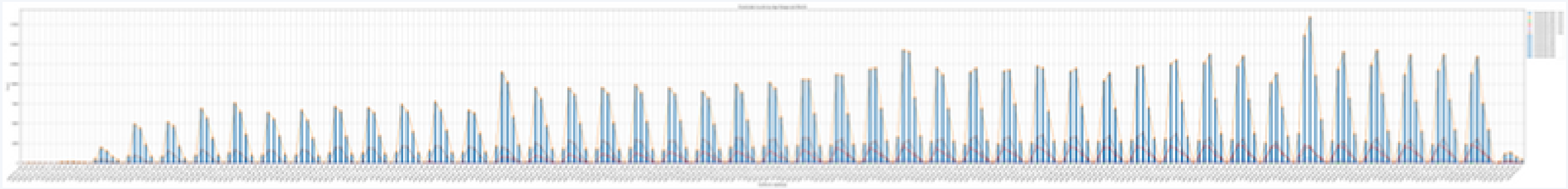




Data Distribution



Data Distribution

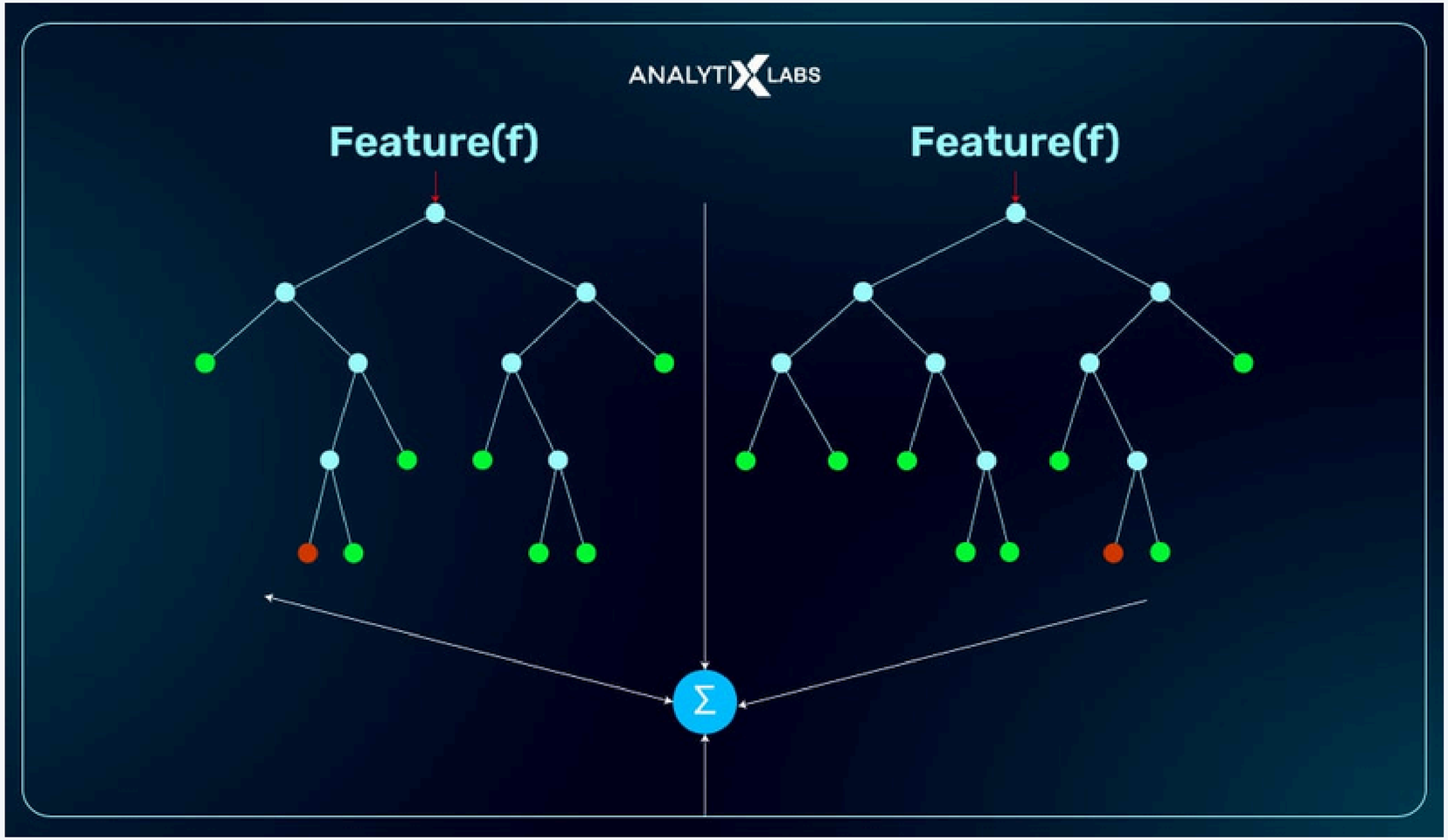


Data Preparation

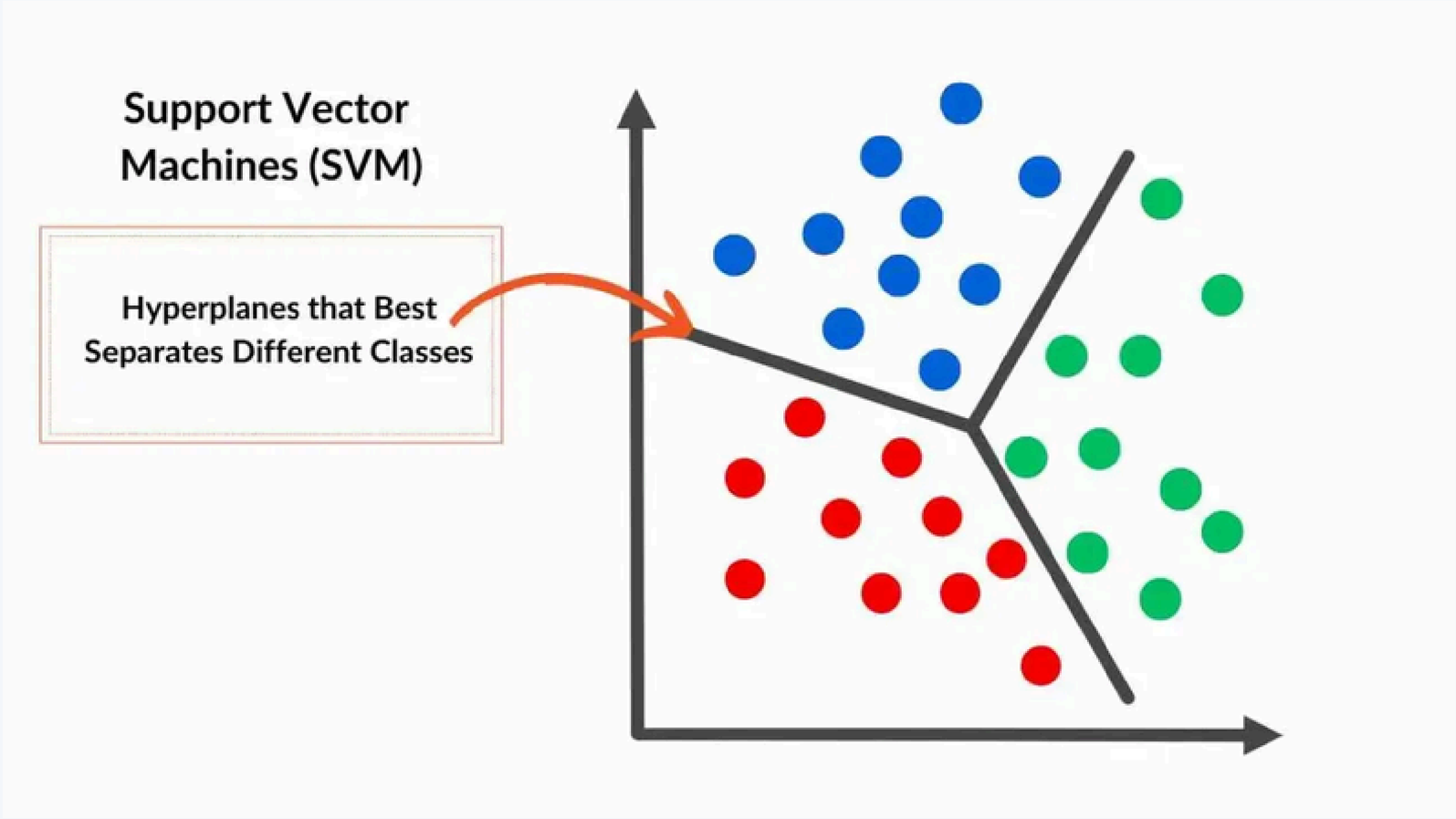
```
    model_rf, model_svm, model_nn= None, None, None
    X_train, X_test, y_train, y_test = None, None, None, None

    # Prepare the data for the model
    data = transcode_counts_by_age.reset_index()
    data['YearMonth'] = data['YearMonth'].astype(str)

    # Convert categorical variables to dummy/indicator variables
    X = pd.get_dummies(data[['AgeRange']], drop_first=True)  # Only AgeRange for now
    X = pd.concat([X, data[transcode_counts_by_age.columns]], axis=1) # Add TransCode columns

    # Create the target variable (e.g., predicting counts for a specific TransCode)
    target_columns = transcode_counts_by_age.columns  # All TransCodes as targets
    y = X[target_columns]

    # Ensure column names are strings
    X.columns = X.columns.astype(str)  # Convert all column names to strings
    y.columns = y.columns.astype(str)  # Convert all column names to strings

    # Split the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # Get the dummy column names from training data
    age_range_dummy_cols = X_train.columns[X_train.columns.str.startswith('AgeRange_')]
```

- Split dataset (2564 - 2567)
  - 80 % for traing
  - 20 % for testing

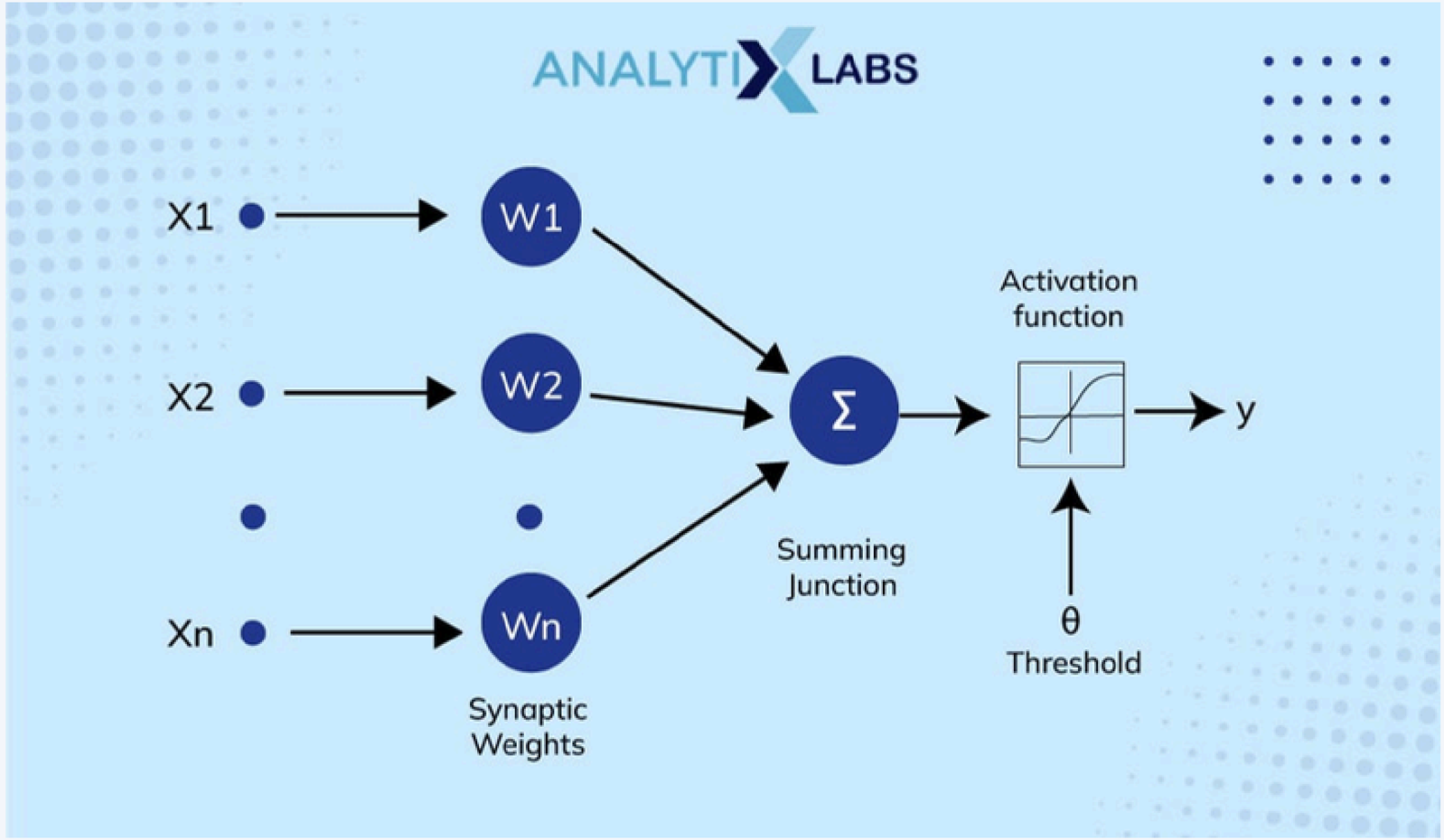
Method 1: Random Forest Regression



Method 2: Support Vector Machine



Method 3: Neural Network



Error Comparison Table

Metric	Random Forest	SVM	Neural Network
MSE	12780.29179814815	3135324.6340004043	127261186621465.58
MAE	46.469074074074065	778.4731053463337	3748915.6730751633

Neural Network's best choice for the data result where the age range of 0-18 remains 0, which is most consistent with reality


Q
A

มหาวิทยาลัย
สงขลานครินทร์

Thank You

For Your Attention