TUGAS INDUSTRIAL DATA ANALYTICS
MACHINE LEARNING PREDICTIVE ANALYSIS WITH POWER BI

Nama : Mochammad Aditya Putra Suhendar
NIM : 120450058
Program Studi : Sains Data

1. Bisnis dan Tujuan Bisnis

Pada tugas kali ini, saya menggunakan data *customer churn* pada sebuah bank dimana dataset berisi data-data nasabah meliputi:

```
RowNumber
CustomerId
Surname
CreditScore
Geography
Gender
Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
Exited
```

Dataset diatas akan diolah untuk memprediksi apakah nasabah-nasabah dalam dataset diatas akan *churn* atau tidak dengan menggunakan *machine learning* dalam *predictive analysis*.

2. Penggunaan Machine Learning

Machine learning yang digunakan pada tugas kali ini yaitu:

1. Decision Trees
2. Random Forest
3. XGBoost

Ketiga machine learning diatas akan membantu memberikan hasil prediksi seberapa banyak customer yang akan *churn.*

3. Dataset

CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Dataset diatas adalah output dari df.head() yang menampilkan beberapa row dan kolom teratas, dalam dilihat bahwa dataset diatas mengandung informasi detail nasabah dari bank tersebut.

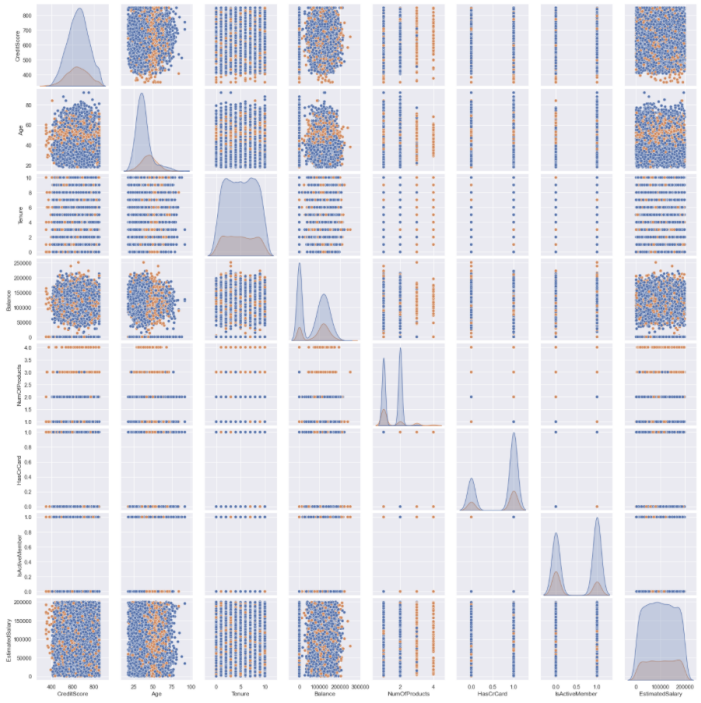
4. Data Preprocessing

Pada *data preprocessing* dilakukan beberapa langkah diantara:

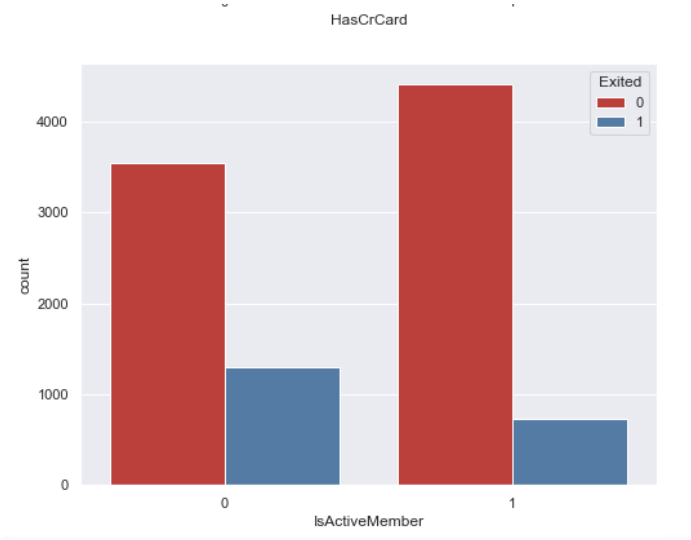
1. Menampilkan apakah ada nilai kosong dalam kolom

```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

2. Visualisasi data



Visualisasi diatas menampilkan/menggambarkan nilai-nilai pada setiap kolom.



Visualisasi ini merupakan salah satu gambaran visualisasi dari beberapa visualisasi yang menggambarkan nasabah yang sudah keluar atau belum, dalam artian "0" untuk "Active Member" dan "1" untuk yang "Exited".

3. Pembuatan dataset baru dari beberapa kolom

CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveM
619	France	Female	42	2	0.00	1	1	
608	Spain	Female	41	1	83807.86	1	0	
502	France	Female	42	8	159660.80	3	1	
699	France	Female	39	1	0.00	2	0	
850	Spain	Female	43	2	125510.82	1	1	

CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	Exited	Geography_France
619	42	2	0.00	1	101348.88	1	True
608	41	1	83807.86	1	112542.58	0	False
502	42	8	159660.80	3	113931.57	1	True
699	39	1	0.00	2	93826.63	0	True
850	43	2	125510.82	1	79084.10	0	False

Dataset ini nantinya akan disatukan kembali dan dijadikan skala , tidak semua kolom akan terambil, akan ada penggabungan data lebih lanjut.

5. Data Splitting

Langkah selanjutnya adalah dengan membagi data antara *train* dan *test* dengan pembagian 9:1.

6. Model Machine Learning

1. Decision Tree

```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', max_depth=2, random_state=1)
```

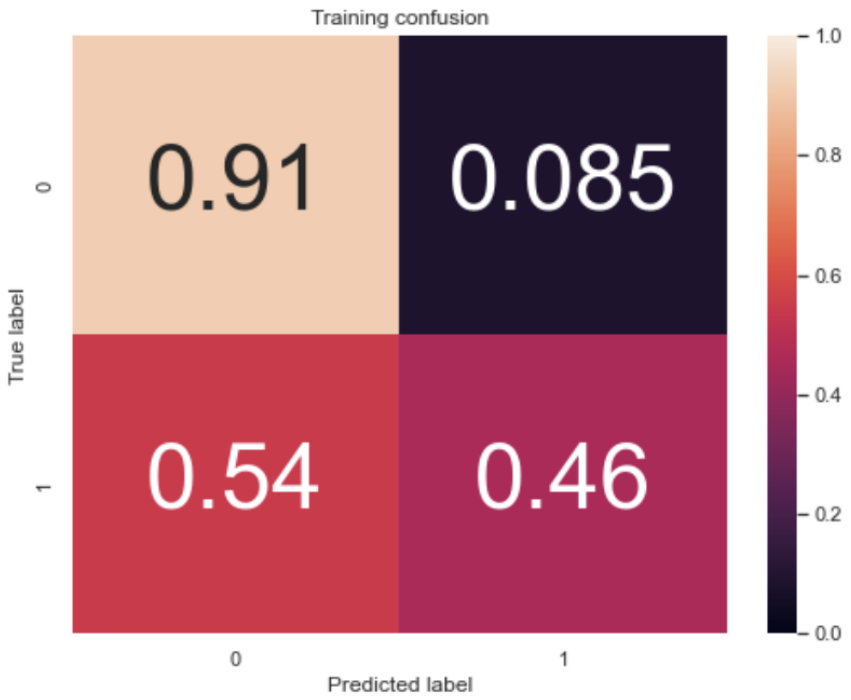
Di atas adalah model Decision Tree yang dibuat.

	Variable	Feature Importance Score
0	Age	0.622346
1	NumOfProducts	0.195552
2	IsActiveMember_1	0.182102
3	CreditScore	0.000000
4	Tenure	0.000000
5	Balance	0.000000
6	EstimatedSalary	0.000000
7	Geography_France	0.000000
8	Geography_Germany	0.000000
9	Geography_Spain	0.000000
10	Gender_Female	0.000000
11	Gender_Male	0.000000
12	HasCrCard_0	0.000000
13	HasCrCard_1	0.000000
14	IsActiveMember_0	0.000000

Melakukan perhitungan dan pengaturan Feature Importance (FI) untuk suatu model, kemungkinan model Decision Tree (DT). Langkah-langkah yang dilakukan melibatkan iterasi melalui setiap kolom fitur (kecuali kolom target 'Exited'), menghitung Feature Importance menggunakan model DT, dan mencetak nilai FI untuk setiap fitur. Setiap nilai FI kemudian disusun dalam DataFrame baru ('fi') yang berisi informasi tentang variabel dan skor Feature Importance. DataFrame ini kemudian ditambahkan ke dalam DataFrame 'final_fi'. Jika 'final_fi' sudah ada, DataFrame baru akan disatukan dengan yang sudah ada, dan jika tidak, DataFrame baru tersebut akan menjadi 'final_fi'. Akhirnya, hasil FI diurutkan secara menurun berdasarkan skor Feature Importance, dan DataFrame 'final_fi' dihasilkan untuk memberikan gambaran yang jelas tentang signifikansi masing-masing fitur dalam model.

```
Training Accuracy is:  0.8213333333333334
Testing Accuracy is:  0.822
```

Hasil akurasi dari data latih dan data uji.



Confusion matrix adalah tabel yang digunakan untuk mengukur kinerja model klasifikasi. Tabel ini menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas. Confusion matrix pada gambar menunjukkan hasil dari tes kebingungan pelatihan model klasifikasi.

Tabel ini memiliki empat sel, yang masing-masing mewakili satu kemungkinan hasil klasifikasi. Sel-sel tersebut adalah:

- **True Positive (TP)**: Prediksi yang benar untuk kelas positif.
- **False Positive (FP)**: Prediksi yang salah untuk kelas positif.
- **True Negative (TN)**: Prediksi yang benar untuk kelas negatif.
- **False Negative (FN)**: Prediksi yang salah untuk kelas negatif.

Berdasarkan informasi yang tertera pada gambar, dapat disimpulkan bahwa model klasifikasi ini memiliki kinerja yang baik untuk kelas positif. Jumlah TP untuk kelas positif adalah 0.91, yang berarti bahwa model berhasil memprediksi dengan benar bahwa 0.91 data positif adalah benar-benar positif. Namun, model ini memiliki kinerja yang kurang baik untuk kelas negatif. Jumlah FN untuk kelas negatif adalah 0.085, yang berarti bahwa model gagal memprediksi bahwa 0.085 data negatif adalah benar-benar negatif

Secara keseluruhan, model klasifikasi ini memiliki kinerja yang cukup baik. Namun, model ini masih dapat ditingkatkan kinerjanya untuk kelas negatif.

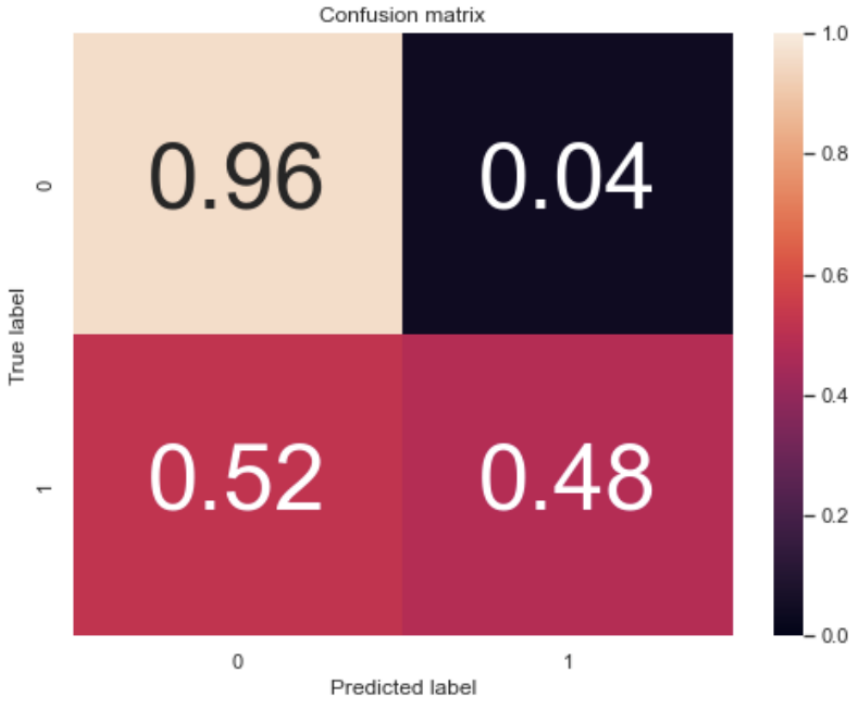
Kesimpulan dari model *Decision Tree* dapat dilihat pada gambar dibawah

```
Tingkat True Positive / Recall per kelas:  [0.914772 0.455489]
Presisi per kelas:  [0.868034 0.577163]
Tingkat False Alarm per kelas:  [0.544511 0.085228]
Tingkat Miss per kelas:  [0.085228 0.544511]
Kesalahan klasifikasi setiap kelas:  [0.178667 0.178667]
Akurasi setiap kelas:  [0.821333 0.821333]

Rata-rata Recall:  0.685130369325603
Rata-rata Presisi:  0.7225985220264771
Rata-rata False Alarm:  0.31486963067439694
Rata-rata tingkat Miss:  0.31486963067439694
Rata-rata Kesalahan klasifikasi:  0.17866666666666667
Rata-rata Akurasi:  0.8213333333333334
```

2. Random Forest

```
Akurasi pada Data Latih:  1.0
Akurasi pada Data Uji:  0.861
```



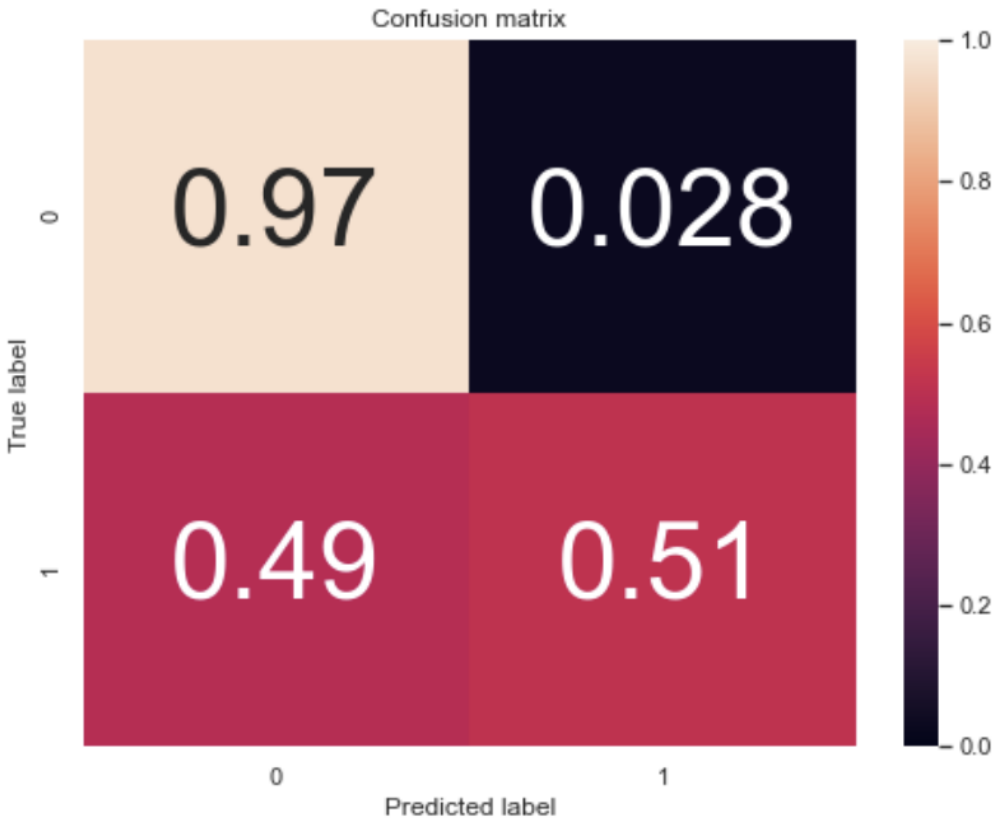
Confusion matrix pada gambar menunjukkan hasil dari pengujian model klasifikasi pada data uji. Tabel ini memiliki dua baris dan dua kolom, yang masing-masing mewakili kelas positif dan negatif.

Jumlah TP untuk kelas positif adalah 0.96, yang berarti bahwa model berhasil memprediksi dengan benar bahwa 0.96 data positif adalah benar-benar positif. Jumlah FN untuk kelas positif adalah 0.04, yang berarti bahwa model gagal memprediksi bahwa 0.04 data positif adalah benar-benar negatif.

Jumlah TN untuk kelas negatif adalah 0.48, yang berarti bahwa model berhasil memprediksi dengan benar bahwa 0.48 data negatif adalah benar-benar negatif. Jumlah FP untuk kelas negatif adalah 0.52, yang berarti bahwa model memprediksi salah bahwa 0.52 data negatif adalah benar-benar positif.

Berdasarkan informasi yang tertera pada gambar, dapat disimpulkan bahwa model klasifikasi ini memiliki kinerja yang baik untuk kelas positif. Namun, model ini memiliki kinerja yang kurang baik untuk kelas negatif.

3. XGBoost



Dapat disimpulkan bahwa model XGBoost memiliki kinerja yang cukup baik, dengan akurasi sebesar 0,861. Akurasi adalah rasio antara jumlah prediksi yang benar dengan total jumlah

prediksi. Dalam hal ini, model XGBoost berhasil memprediksi dengan benar 86,1% dari total data uji.

Namun, model XGBoost ini masih dapat ditingkatkan kinerjanya untuk kelas negatif. Hal ini ditunjukkan oleh nilai recall yang rendah untuk kelas negatif, yaitu 0,8. Recall adalah rasio antara jumlah prediksi yang benar untuk kelas positif dengan total jumlah data positif. Dalam hal ini, model XGBoost hanya berhasil memprediksi dengan benar 80% dari total data positif.

Nilai recall yang rendah untuk kelas negatif menunjukkan bahwa model XGBoost sering kali gagal memprediksi bahwa data negatif adalah benar-benar negatif. Hal ini dapat menyebabkan masalah serius, terutama jika model XGBoost digunakan untuk aplikasi yang membutuhkan akurasi yang tinggi untuk kelas negatif.

Berdasarkan informasi yang tertera pada gambar, dapat disimpulkan bahwa:

- Model XGBoost berhasil memprediksi dengan benar 91 data positif dan 48 data negatif.
- Model XGBoost memprediksi salah 52 data negatif.

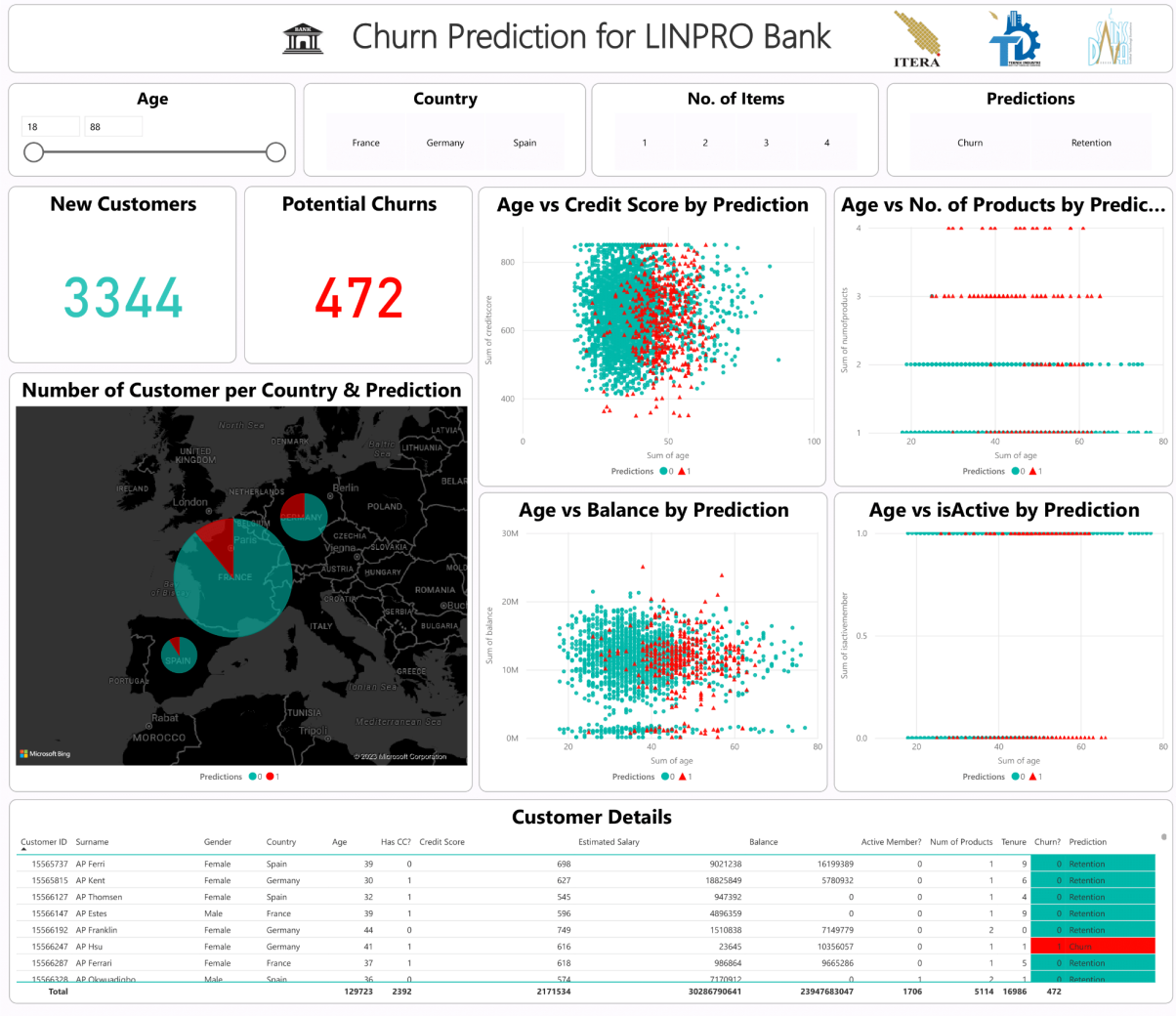
7. Hasil Prediksi

creditscore	geography	gender	age	tenure	balance	numofproducts	hascrcard	isactivemember	estimatedsalary	Predictions - Churn or Not	Predictions - Probability to Churn	Predictions - Churn or Not Desc
619	France	Female	42	2	0	1	1	1	101348.88	0	0.330071	Retention
608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0	0.193652	Retention
502	France	Female	42	8	159660.8	3	1	0	113931.57	1	0.944657	Churn
699	France	Female	39	1	0	2	0	0	93826.63	0	0.065576	Retention
850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0	0.097383	Retention

Hasil prediksi yang dilakukan model dimasukan pada dataset awal dengan menambah beberapa kolom, kolom-kolom ini yang nantinya akan membantu membuat hasil *predictive analysis* pada dashboard Power BI.

8. Dashboard

Data hasil prediksi disimpan dalam bentuk .csv kemudian di-*import* kedalam software Power BI yang pada tahap selanjutnya akan dibuat *dashboard* dan dilakukan analisis tujuan bisnis dari bank tersebut untuk melihat berapa nasabah yang akan menetap dan akan *churn*.



Dashboard memiliki peran penting untuk tujuan bisnis yang dilakukan oleh LINPRO Bank dengan beberapa visualisasi yang interaktif:

1. Age Slider

   Bagan interaktif yang dapat diatur sesuai dengan keperluan pengguna agar dapat meganalisis sesuai rentang umur nasabah yang diperlukan.

2. Country Selector

   Bagan interaktif yang dapat diatur sesuai dengan keperluan pengguna untuk melihat sebaran nasabah dari tiga negara untuk berapa jumlah nasabah baru dan potensi untuk *churn.*

3. No. of Items Selector

   Bagan interaktif yang dapat diatur sesuai dengan keperluan pengguna untuk melihat sebaran nasabah untuk berapa produk yang nasabah miliki, ini dapat menjadi solusi atau target untuk marketing sales melihat potensi penawaran produk baru LINPRO Bank bagi nasabah untuk meningkatkan profit dan elektabilitas LINPRO Bank.

4. Prediction Selector

   Bagan interaktif yang dapat diatur sesuai dengan keperluan pengguna untuk melihat sebaran nasabah antara prediksi yang akan *churn* dan akan menetap, bagan ini dibuat untuk memfokuskan pengguna/analyst kepada nasabah yang akan *churn* dengan membuat penawaran baru dan bonus untuk nasabah agar menetap.

5. New Customers & Potential Churn Cards

   Bagan ini dibuat sebagai informasi yang mudah terlihat untuk mendeskripsikan total jumlah nasabah pada LINPRO Bank dan jumlah kemungkinan potensi nasabah yang akan churn. Bagan ini penting untuk mengetahui sejauh mana kinerja customer relation pada LINPRO Bank.

6. Age vs Credit Score by Prediction

   Bagan dibuat dengan tujuan untuk melihat persebaran nasabah berdasarkan umur dengan tingkat credit score berdasarkan prediksi nasabah yang akan churn dan yang menetap dengan warna merah segitiga untuk yang akan churn dan warna hijau untuk yang menetap. Dapat dilihat pada bagan tersebut bahwa rata-rata usia di rentang 40 - 50 tahun memiliki credit score yang cukup tinggi diantara 600 hingga 800 dengan potensi nasabah churn di umur 50 tahun dengan credit score tersebar rapat di antara 600 - 800. Hasil ini dapat di analisis bahwa perlu ada tindakan lebih lanjut untuk usia di rentang 40 - 50 tahun untuk menaikkan credit score sehingga nasabah yang memiliki potensial churn berkemungkinan akan menetap.

7. Age vs No. Products by Predictions

   Dapat dilihat pada bagan tersebut bahwa semakin banyak produk yang dimiliki oleh nasabah adalah nasabah yang memiliki potensi churn. Dapat dilihat pada indeks jumlah produk 3 dan 4 diisi oleh prediksi nasabah potensial churn yang artinya perlu adanya tindakan lebih lanjut dari bagian marketing, sales, dan customer relations untuk memberikan penawaran yang lebih baik, bonus yang ekstra, dan penanganan lebih lanjut agar nasabah yang memiliki banyak produk terkesan dan juga mengurungkan niat agar untuk menetap.

8. Age vs Balance by Predictions

   Dapat dilihat pada bagan tersebut terdapat persebaran nasabah yang memiliki potensial churn di umur 40 - 50 tahun dengan balance di antara 10 juta. Dapat disimpulkan bahwa perlu dilakukan evaluasi lebih lanjut mengenai biaya tambahan seperti admin bank, biaya transaksi, dan juga bonus transaksi (jika ada).

9. Age vs isActive by prediction

   Pada bagan tersebut 0 merupakan nasabah yang tidak aktif dan 1 merupakan nasabah yang aktif dengan rentang umur 40 - 60 tahun nasabah yang aktif memiliki potensial churn yang tinggi. Terdapat beberapa faktor yang memengaruhi salah satunya adalah umur dimana nasabah dengan umur di atas 50 tahun biasanya sudah memasuki usia pensiun dengan artian nasabah minim bertransaksi pada usia tersebut, sehingga alih-alih nasabah menyimpan uangnya dan akan tergerus oleh biaya administrasi nasabah diluar usia produktif lebih memilih untuk menonaktifkan rekeningnya.

10. Number of Customer per Country & Prediction

    Bagan diatas menunjukkan peta persebaran prediksi berdasarkan 3 negara dengan gambaran pie chart sesuai dengan jumlah, semakin besar pie chart pada negara tersebut semakin mewakili jumlah nasabah pada negara tersebut. Peta ini sangat bermanfaat untuk mengetahui negara mana yang harus difokuskan untuk meningkatkan jumlah nasabah maupun membuat nasabah yang ada agar menetap.

11. Customer Details

    Bagan ini berbentuk tabel yang dapat di scroll dengan berisi keterangan semua nasabah pada bank tersebut dengan output di sisi paling kanan menunjukkan prediksi apakah nasabah tersebut menetap atau pun churn. Dapat dilakukan filtering pada tabel tersebut untuk memfokuskan kepada nasabah churn kemudian dilakukan pendekatan lebih lanjut kepada para nasabah tersebut untuk dilakukan penawaran ekstra agar nasabah tersebut menetap sebagai nasabah LINPRO Bank.