Nomor Peserta : ODD10097
Nama Peserta 1 : Sarah Shabrina
Nama Peserta 2 : Farah Qotrunnada

LEMBAR JAWAB FINAL DOKTER DATA 2020

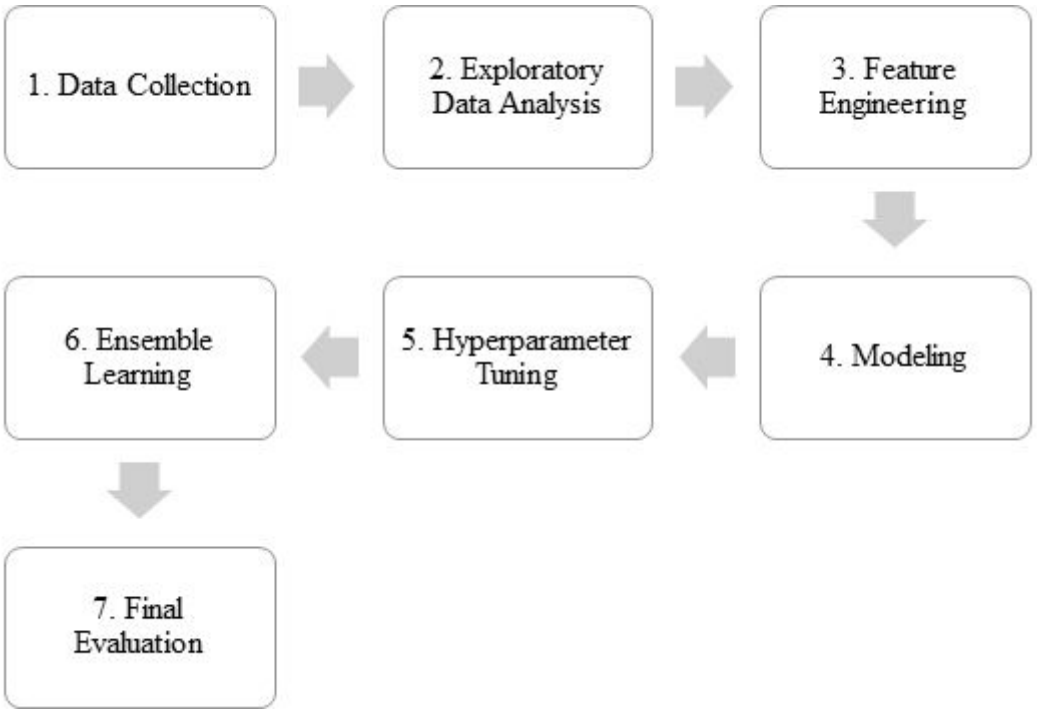
1. Problem Statement

Masalah : Analisis kondisi penggunaan akses informasi dan komunikasi dari pelanggan

Tujuan :

1. Mengekstrak informasi dan insight dari data
2. Memprediksi *flag customer churn or active* berdasarkan data dengan model klasifikasi terbaik
3. Mengetahui variabel utama yang mempengaruhi *flag customer churn or active*

2. Workflow Penyelesaian Masalah



1. Data Collection

a. Attribute Information

Mengekstrak informasi awal dari data yang dimiliki tanpa perlu analisis mendalam

2. Exploratory Data Analysis

Melakukan eksplorasi data untuk memahami dan mendapatkan gambaran dari data yang dimiliki untuk pemodelan berikutnya.

a. Statistika Deskriptif

Statistik Deskriptif adalah statistik yang berfungsi untuk mendeskripsikan atau membari gambaran terhadap obyek yang diteliti melalui data sampel atau populasi

sebagaimana adanya tanpa melakukan analisis dan membuat kesimpulan yang berlaku secara umum.

b. Mendeteksi Type Data

Secara umum, tipe data statistik adalah sebagai berikut :

1. Nominal. Nilai atribut bertipe nominal tersusun atas simbol-simbol yang berbeda, yaitu suatu himpunan terbatas. Pada tipe nominal, tidak ada urutan ataupun jarak antar atribut. Tipe ini sering juga disebut kategorikal atau enumerasi. Secara umum, tipe output pada supervised learning adalah data nominal.
2. Ordinal. Nilai ordinal memiliki urutan, sebagai contoh 4 > 2 > 1. Tetapi jarak antar suatu tipe dan nilai lainnya tidak harus selalu sama, seperti 4-2 =/= 2-1. Atribut ordinal terkadang disebut sebagai numerik atau kontinu

c. Mendeteksi Incomplete Values

Seperti data yang kosong (missing value)

d. Mendeteksi Noisy Data

Seperti error dan outlier (pencilan)

e. Mendeteksi Keseimbangan Data

Keseimbangan data berarti proporsi antara tiap feature dengan label terkait

f. Mendeteksi Korelasi antar Feature

Korelasi dideteksi untuk memeriksa hubungan kelinearan antar Feature dan Feature dengan Label Kelas.

3. Feature Engineering

-Remove Outlier

Outlier yang tidak memiliki makna dan berpotensi memperburuk kondisi data dapat dihapus

-Imputing Missing Value

Secara umum ada 2 cara untuk menangani missing value yaitu dengan menghapusnya atau menggantinya dengan nilai yang lain

-Feature Selection

Untuk memilih variable input yang dapat mengefesiensi input data, mengurangi noise dan irrelevant variable dan tetap mendukung hasil prediksi yang bagus.

a. Correlation

Menggunakan koefisien korelasi Pearson’s untuk mengukur hubungan, asosiasi antar dua peubah kontinu. Bernilai antara 0 dengan +-1, dengan 0 berarti taka da korelasi antara 2 peubah.

b. Feature Important

Memilih feature yang relevann dan memberikan pengaruh bagus bagi hasil akurasi prediksi

-Feature Scaling

Mengkonversi nilai feature ke dalam skala tertentu.

a. Normalisasi

Adalah proses mengubah feature numerik ke jangkauan standar dari suatu nilai. Jangkauan nilai dapat berupa [-1,1] atau [0,1]

b. Standardisasi

Data diubah sehingga rataan menjadi nol dan standar deviasi konstan

C. Transformasi Logaritmik

Menghitung logaritma dari setiap nilai dalam data untuk mengkompress jangkauan yang lebar ke jangkauan yang lebih sempit.

4. Modelling

-Data Splitting

Pada umumnya data yang telah diproses sebelumnya dibagi menjadi data training, validation/validation, dan testing data. Mesin dilatih menggunakan training data, saat proses training, performance measure diukur berdasarkan kemampuan mengenali/ mengeneralisasi validation data. Perlu diketahui, performance measure diukur menggunakan validation data untuk menghindari overfitting dan underfitting.

Overfitting adalah keadaan ketika model memiliki kinerja baik hanya untuk training data/seen examples tetapi tidak memiliki kinerja baik untuk unseen examples. Underftting adalah keadaan ketika model memiliki kinerja buruk baik untuk training data dan unseen examples.

Cross validation adalah teknik untuk menganalisis apakah suatu model memiliki generalisasi yang baik (mampu memiliki kinerja yang baik pada unseen examples). Saat proses training, kita latih model dengan training data serta dievaluasi menggunakan validation data. Teknik cross validation bekerja dengan prinsip yang sama, yaitu membagi sampel asli menjadi beberapa subsampel dengan partisi sebanyak K (K-fold)

-Pemodelan

Pemodelan dilakukan berdasarkan tipe data dan ada/tidaknya target variable/label. Klasifikasi digunakan ketika terdapat data kategorikal dan terdapat target variable. Jenis-jenis klasifikasi :

-Multilayer Perception

Multilayer Perception merupakan bagian dari Artificial Neurral Network. MLP merupakan neural network yang terdiri dari input layer, hidden layer dan output layer.

-Support Vector Machine

Support Vector Machine (SVM) merupakan sebuah algoritma klasifikasi untuk data linear dan non-linear. SVM menggunakan *mapping* non-linear untuk mentransformasikan *training* data awal ke dimensi yang lebih tinggi

-Random Forest Classifier

Random forest (RF) adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi *random forest* dilakukan melalui penggabungan pohon (*tree*) dengan melakukan *training* pada sampel data yang dimiliki. Penggunaan pohon (tree) yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan *random forest* diambil berdasarkan hasil *voting* dari *tree* yang terbentuk. Pemenang dari *tree* yang terbentuk ditentukan dengan *vote* terbanyak. Pembangunan pohon (*tree*) pada *random forest* sampai dengan mencapai ukuran maksimum dari pohon data. Akan tetapi,pembangunan pohon *random forest* tidak dilakukan pemangkasan (*pruning*) yang merupakan sebuah metode untuk mengurangi kompleksitas ruang. Pembangunan dilakukan dengan penerapan metode *random feature selection* untuk meminimalisir kesalahan. Pembentukan pohon (*tree*) dengan sample data menggunakan variable yang diambil secara acak dan menjalankan klasifikasi pada semua *tree* yang terbentuk.

-KNN

K-nearest neighbors atau KNN adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (*train data sets*), yang diambil dari k tetangga terdekatnya (*nearest neighbors*). K-nearest neighbors melakukan klasifikasi dengan proyeksi data pembelajaran pada ruang berdimensi banyak. Ruang ini dibagi menjadi bagian-bagian yang merepresentasikan kriteria data pembelajaran. Setiap data pembelajaran direpresentasikan menjadi titik-titik ***c*** pada ruang dimensi banyak.

-Decision Tree

Decision tree merupakan suatu metode klasifikasi yang menggunakan struktur pohon, dimana setiap node merepresentasikan atribut dan cabangnya merepresentasikan nilai dari atribut, sedangkan daunnya digunakan untuk merepresentasikan kelas. Node teratas dari decision tree ini disebut dengan root.

-Gaussian Naïve Bayes

Gaussian Naïve Bayes adalah variasi dari Naïve Bayes yang mengikuti Distribusi Normal Gaussian dengan support berupa data yang kontinu. Naïve Bayes merupakan kelompok klasifikasi yang berdasarkan teorema Bayes.

-Logistic Regression

Logistic Regression adalah sebuah algoritma klasifikasi untuk mencari hubungan antara fitur (input) diskrit/kontinu dengan probabilitas hasil output diskrit tertentu

-Gradien Boosting
Gradient boosting adalah algoritma machine learning yang menggunakan ensemble dari decision tree untuk memprediksi nilai

5. Hyperparameter Tuning

Hyperparameter tuning adalah pemilihan optimal hyperparameter untuk suatu learning algorithm. Hyperparameter adalah parameter yang nilainya telah ada sebelum proses learning dimulai.
Ada 2 cara untuk mencari optimal hyperparameter:
a. Grid Search : mencari secara 'exhaustively' untuk hyperparameter yang optimal.
b. Random Search : mengambil subset spesifik hyperparameter secara random daripada mencari secara manual.

6. Ensemble Learning

Ensemble learning adalah paradigma dari machine learning di mana model yang banyak (atau biasa disebut "weak learners") dilatih untuk menyelesaikan masalah yang sama dan dikombinasikan untuk mendapatkan hasil yang lebih baik. Hypotesis dasarnya adalah ketika model yang lemah dikombinasikan maka akan terbentuk model dengan akurasi yang lebih tinggin dan/atau robust model.

Ada 3 jenis ensemble learning yaitu :
- bagging, memperhatikan weak learners yang sejenis, mempelajari secara independent satu sama lain secara parallel dan mengkombinasikan dengan suatu proses deterministic rata-rata.
- boosting, memperhatikan weak learners yang sejenis, mempelajari secara bertahap (base model bergantung pada model sebelumnya) dan mengkombinasikan mereka dengan strategy deterministik
- stacking, memperhatikan weak learners yang berbeda jenis, dipelajari secara parallel dan kemudian dikombinasikan.

7. Final Evaluation

Setelah selesai dilatih, maka model hasil pembelajaran dievaluasi dengan testing data. Training, validation, dan testing data tersusun oleh data yang independen satu sama lain (tidak beririsan) untuk memastikan model yang dihasilkan memiliki generalisasi cukup baik.

Bentuk Evaluasi :

1. Accuracy,

Akurasi adalah salah satu cara untuk mengevaluasi model klasifikasi.
Akurasi bernilai perbandingan jumlah prediksi yang benar dengan total prediksi.
Untuk klasifikasi biner, akurasi dapat dihitung sebagai berikut :

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$

Di mana TP = True Positives, TN = True Negatives, FP = False Positives, dan FN = False Negatives.

2. F-measure

F-score membantu untuk mengukur recall dan precision di waktu yang sama.

3. AUC, ROC

AUC merupakan area di bawah kurva (Area under the Curve of) ROC (Receiver Operating Characteristic), suatu kurva yang menggambarkan probabilitas dengan variabel sensitivitas dan kekhususan (specificity) dengan nilai batas antara 0 hingga 1. Area di bawah kurva memberikan gambaran tentang keseluruhan pengukuran atas kesesuaian dari model yang digunakan.

3. Hasil Analisis

1. Data Collection

Sumber Data :
Soal Final Dokter Data 2020

Attribute Information :

Variabel	Keterangan	Pendefinisian
Y	Flag customer churn or active (1 churner/non active; 0 active)	Kategori Pelanggan (1 : Churner 2. : Aktif)
X1	Total revenue for all internet products in last 2 month	Jumlah pendapatan dari produk internet dalam 2 bulan terakhir
X2	Total revenue for all pots products in last 2 month (all pots = local, SLJJ, mobile, SLI)	Jumlah pendapatan untuk semua produk pots (plain old telephone service) dalam 2 bulan terakhir
X3	Total payment for all pots products in last 2 month (all pots = local, SLJJ, mobile, SLI)	Jumlah pembayaran untuk produk pots dari pelanggan dalam 2 bulan terakhir
X4	Total upload for internet in last 2 month	Jumlah unggahan di internet dalam 2 bulan terakhir

X5	Total download for internet in last 2 month	Jumlah unduhan di internet dalam 2 bulan terakhir
X6	Total call for all pots products in last 2 month	Jumlah telfon dalam semua produk pots dalam 2 bulan terakhir

2. EDA (Exploratory Data Analysis)

Dilakukan EDA untuk memahami data, memperoleh dan mengekstrak informasi dari data.

Statistika Deskriptif

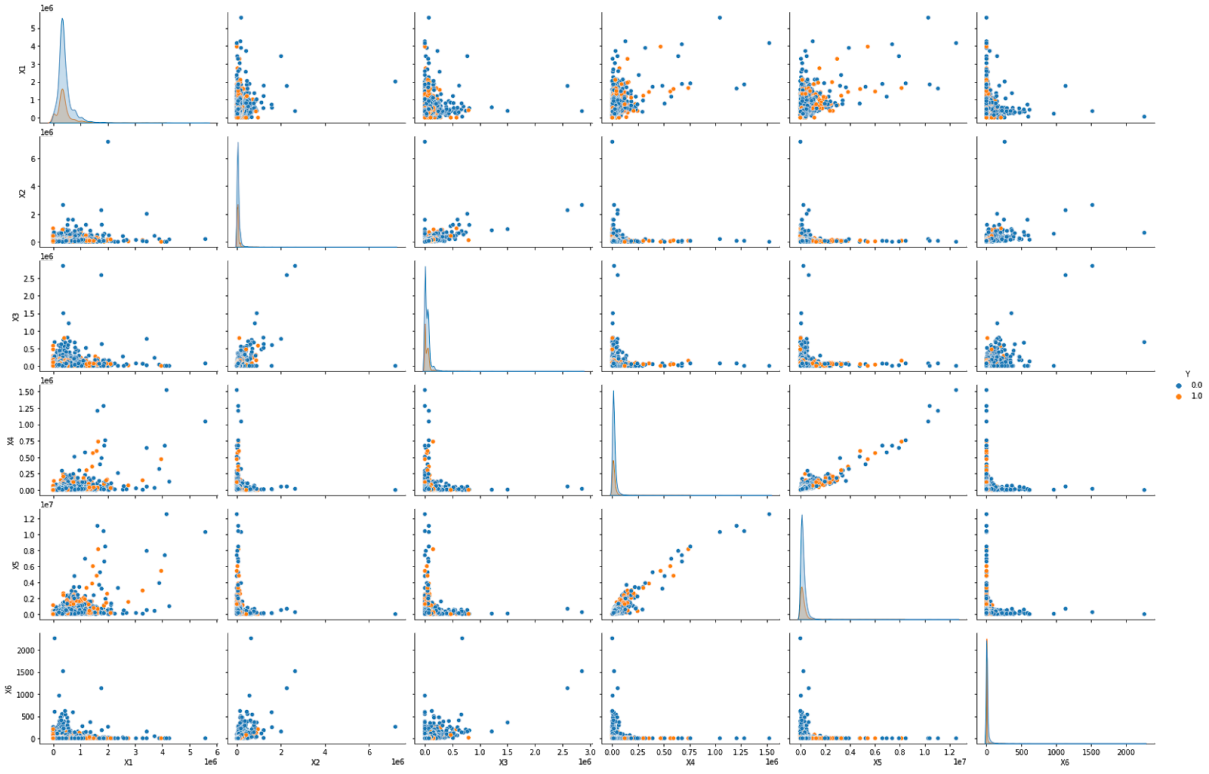
	X1	X2	X3	X4	X5	X6	Y
count	9.000000e+03	9.000000e+03	9.000000e+03	9.000000e+03	9.000000e+03	9000.000000	9000.000000
mean	4.129862e+05	5.639445e+04	3.733415e+04	1.798978e+04	2.226881e+05	8.366889	0.252667
std	2.971791e+05	1.156696e+05	7.058762e+04	3.997671e+04	4.104448e+05	45.399525	0.434566
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000
25%	2.640000e+05	1.650000e+04	0.000000e+00	5.800028e+03	6.700649e+04	0.000000	0.000000
50%	3.465000e+05	4.400000e+04	2.750000e+04	1.154101e+04	1.446450e+05	0.000000	0.000000
75%	4.730000e+05	6.600000e+04	5.500000e+04	2.014478e+04	2.728671e+05	1.000000	1.000000
max	5.577000e+06	7.195431e+06	2.854794e+06	1.522577e+06	1.252349e+07	2264.000000	1.000000

```
 #   Column  Non-Null Count
---  ------  --------------
 0   X1      9000 non-null
 1   X2      9000 non-null
 2   X3      9000 non-null
 3   X4      9000 non-null
 4   X5      9000 non-null
 5   X6      9000 non-null
 6   Y       9000 non-null
```

Insight dari data tersebut diantaranya adalah :

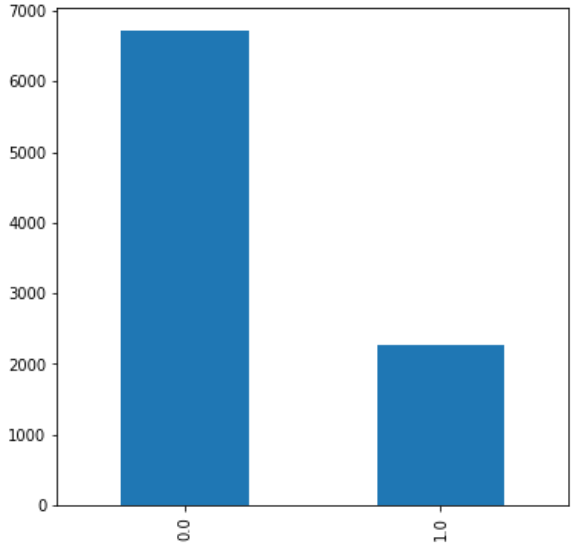
a. Dapat dilihat dari statistika deskriptif tersebut tidak ada missing value (data yang kosong) karena jumlah (count) dari semua feature atau variabel adalah 9000 data.
b. Data feature bertipe numerik sedangkan label/target variable berupa data bertipe kategorikal
c. Range/Jangkauan antar variabel tidak terlalu berbeda jauh yaitu memiliki nilai minimum 0 dan maksimum adalah angka dengan 6-8 digit

Visualisasi Data

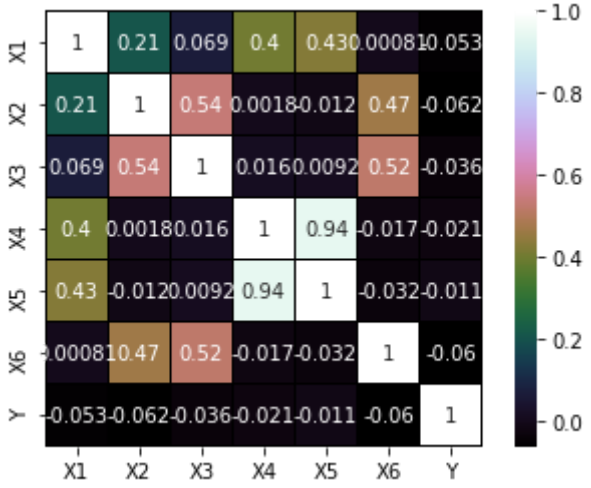


d. Variabel X4 (Jumlah unggahan di internet dalam 2 bulan terakhir) dan X5 (Jumlah unduhan di internet dalam 2 bulan terakhir) berkorelasi positif

e. Data ada yang tidak konsisten. Data Y (kategori pelanggan) aktif namun data featurenya (X1 sampai X6) semua bernilai nol artinya tidak ada aktivitas dalam telekomunikasi tersebut.



f. Berdasarkan diagram batang di atas, jumlah kategori pelanggan churn atau tidak aktif lebih banyak daripada pelanggan yang aktif. Data label tidak seimbang.

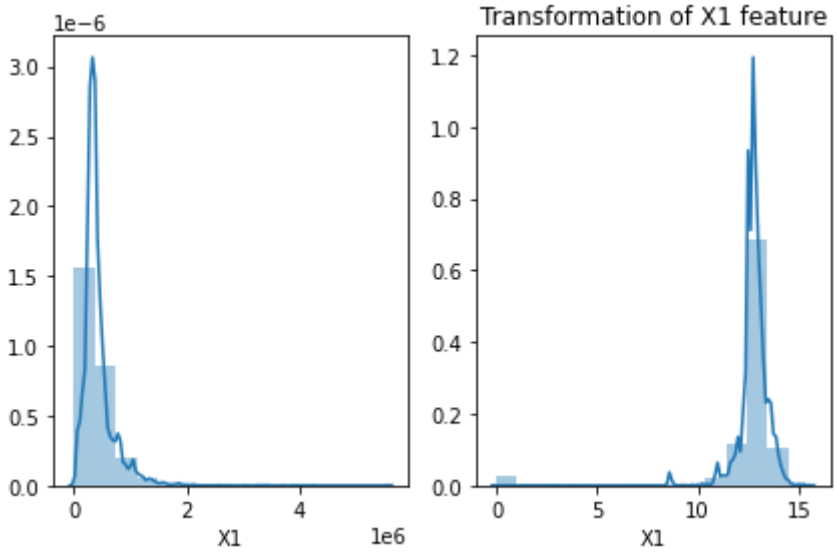


g. Berdasarkan diagram heat map di atas yang menggambarkan korelasi antar feature/variabel, dapat dilihat bahwa analisis korelasi sebelumnya yang mengatakan bahwa X4 (Jumlah unggahan di internet dalam 2 bulan terakhir) dan X5 (Jumlah unduhan di internet dalam 2 bulan terakhir) berkorelasi positif benar. Nilai korelasi antara variabel X4 dan X5 yaitu 0.94. Artinya jumlah unggahan di internet dan jumlah unduhan di internet tersebut mempunyai hubungan kelinearan yang tinggi. Sedangkan variabel-variabel lain tidak memiliki korelasi antar variabel yang cukup tinggi. Artinya hubungan kelinearan dari variabel-variabel tersebut cukup tinggi.

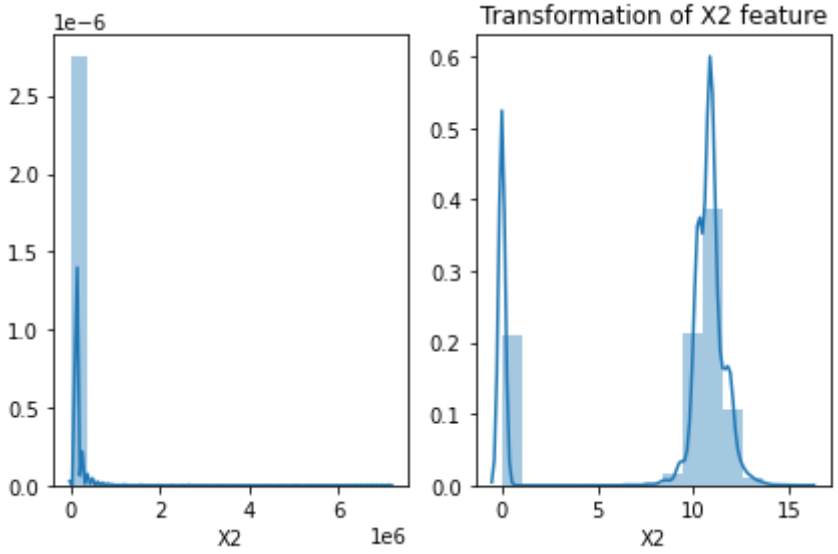
3. Feature Engineering

a. Outlier pada data tidak dihapus
b. Tidak dilakukan imputing *missing value* karena tidak ada data yang kosong
c. Feature Selection
1. Drop Inconsistent Data, yang telah diketahui pada EDA
2. Berdasarkan heatmap analysis, tidak ada variabel yang signifikan berkorelasi dengan variabel target/label. Maka, seluruh feature akan digunakan untuk pemodelan selanjutnya.

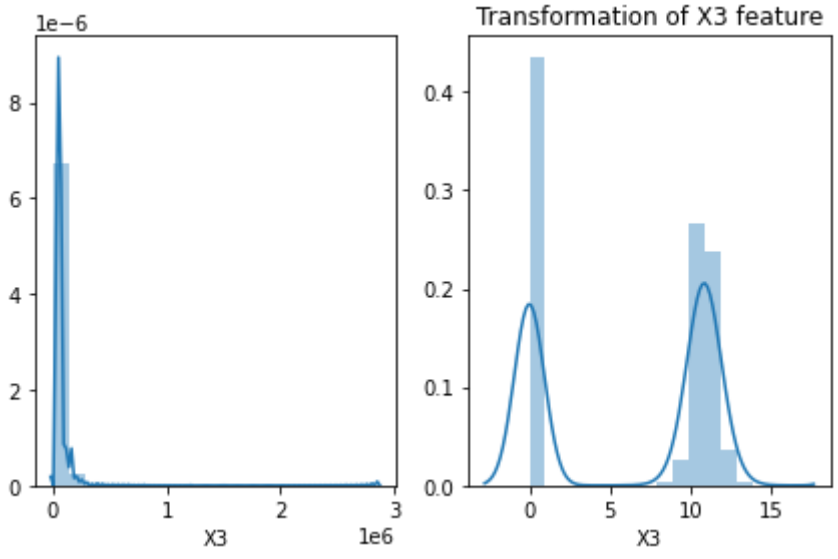
d. Feature Scaling
1. Transformasi Logaritmik


1e−6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
X1
1e6

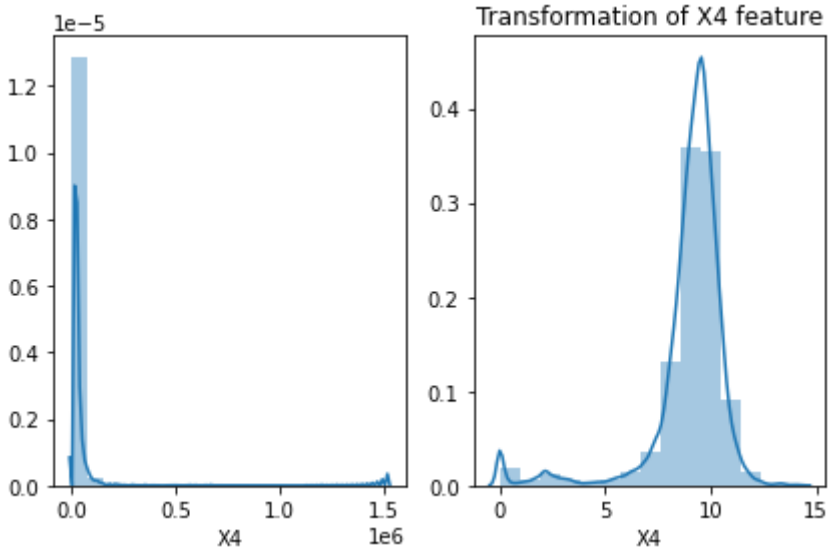
Transformation of X1 feature
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
X1


1e−6
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
X2
1e6

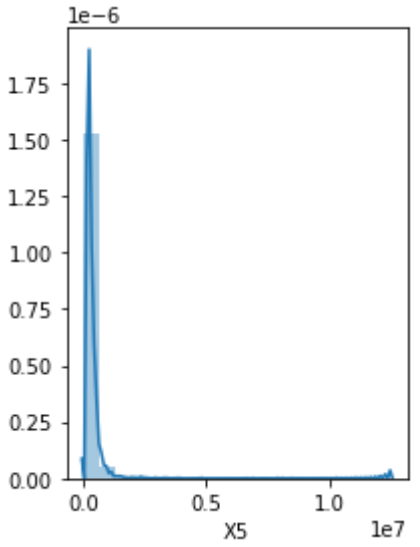
Transformation of X2 feature
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
X2

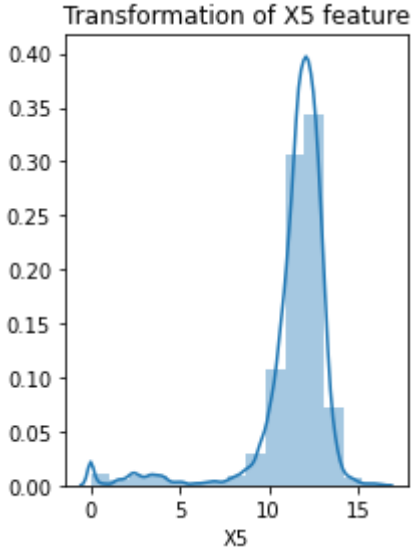

1e−6
8
6
4
2
0
0
1
2
3
X3
1e6

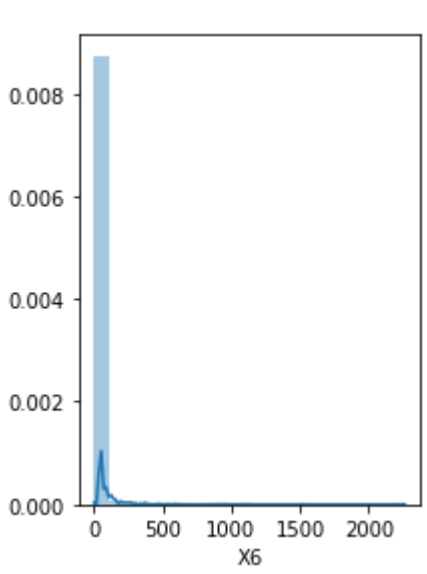
Transformation of X3 feature
0.4
0.3
0.2
0.1
0.0
0
5
10
15
X3

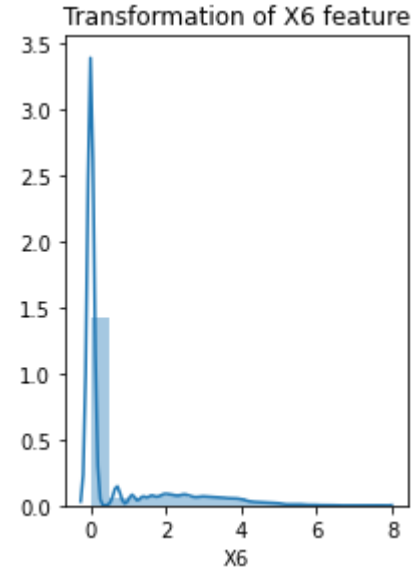

1e−5
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
X4
1e6

Transformation of X4 feature
0.4
0.3
0.2
0.1
0.0
0
5
10
15
X4


1e−6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
X5
1e7


Transformation of X5 feature
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
X5


0.008
0.006
0.004
0.002
0.000
0
500
1000
1500
2000
X6


Transformation of X6 feature
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
X6

Masing-masing distribusi dari variabel tersebut *skew/miring*. Kemudian dilakukan peninjauan proses transformasi logaritmik dari data. Namun, variabel X1, X2, X3, X6 tetap berdistribusi miring. Sedangkan variabel X4 dan X5 berubah menuju ke distribusi normal. Akhirnya dilakukan transformasi logaritmik untuk variabel X4 dan X5.

2. Standardisasi
   Dilakukan proses standardisasi agar semua variabel dalam data tersebut berdistribusi normal dengan rataan 0 dan variansi konstan.
3. Normalisasi
   Tidak dilakukan proses normalisasi

4.Modeling

a. Data Splitting

Data latih yang ada dipartisi data training : data uji = 70% : 30%. Data training tersebut digunakan untuk melakukan pembelajaran data. Sedangkan data uji digunakan untuk mengevaluasi pemodelan secara final
Kemudian data training tersebut diolah lagi menggunakan prinsip k-fold cross validation dengan untuk menghindari model yang overfitting.
Hasil dari cross validation model adalah :

```
CV Score of mplc : 0.7544928053140885
CV Score of SVM : 0.7557645389317437
CV Score of RFC: 0.7315951498318479
CV Score of KNN : 0.7040848513823385
CV Score of Decision Tree : 0.7513122697314751
CV Score of Gaussian Nb : 0.39179804541309526
CV Score of LogReg : 0.7536973868435332
CV Score of Grad_Bos : 0.7529022213284462
```

Dapat dilihat bahwa CV dari model berada di rentang 0.7 kecuali Gaussian Nb. Angka ini cukup mengatakan bahwa selain gaussian nb, model cukup tidak mengalami overfitting.

b. Pemodelan

Digunakan beberapa model klasifikasi dengan machine learning diataranya : Multilayer Perception, Support Vector Machine, Random Forest Classifier, KNN, Decision Tree, Gaussian Naïve Bayes, Logistic Regression, Gradien Boosting

Berdasarkan nilai akurasi tertinggi diperoleh 2 model terbaik yaitu DT dan MPLC. Berdasarkan nilai AUC tertinggi dan mendekati 1 diperoleh 2 model terbaik yaitu Gradient Boosting dan Logistic Regression.

Confusion Matrix dari setiap model adalah :

```
mplc

array([[1995,   34],
       [ 621,   46]], dtype=int64)

SVM

array([[1995,   34],
       [ 619,   48]], dtype=int64)

RFC

array([[1878,  151],
       [ 556,  111]], dtype=int64)

KNN

array([[1783,  246],
       [ 541,  126]], dtype=int64)

Decision Tree

array([[1997,   32],
       [ 611,   56]], dtype=int64)

Gaussian Nb

array([[ 440, 1589],
       [ 101,  566]], dtype=int64)

Logistic Regression

array([[1996,   33],
       [ 623,   44]], dtype=int64)

Gradient Boosting

array([[1989,   40],
       [ 616,   51]], dtype=int64)
```

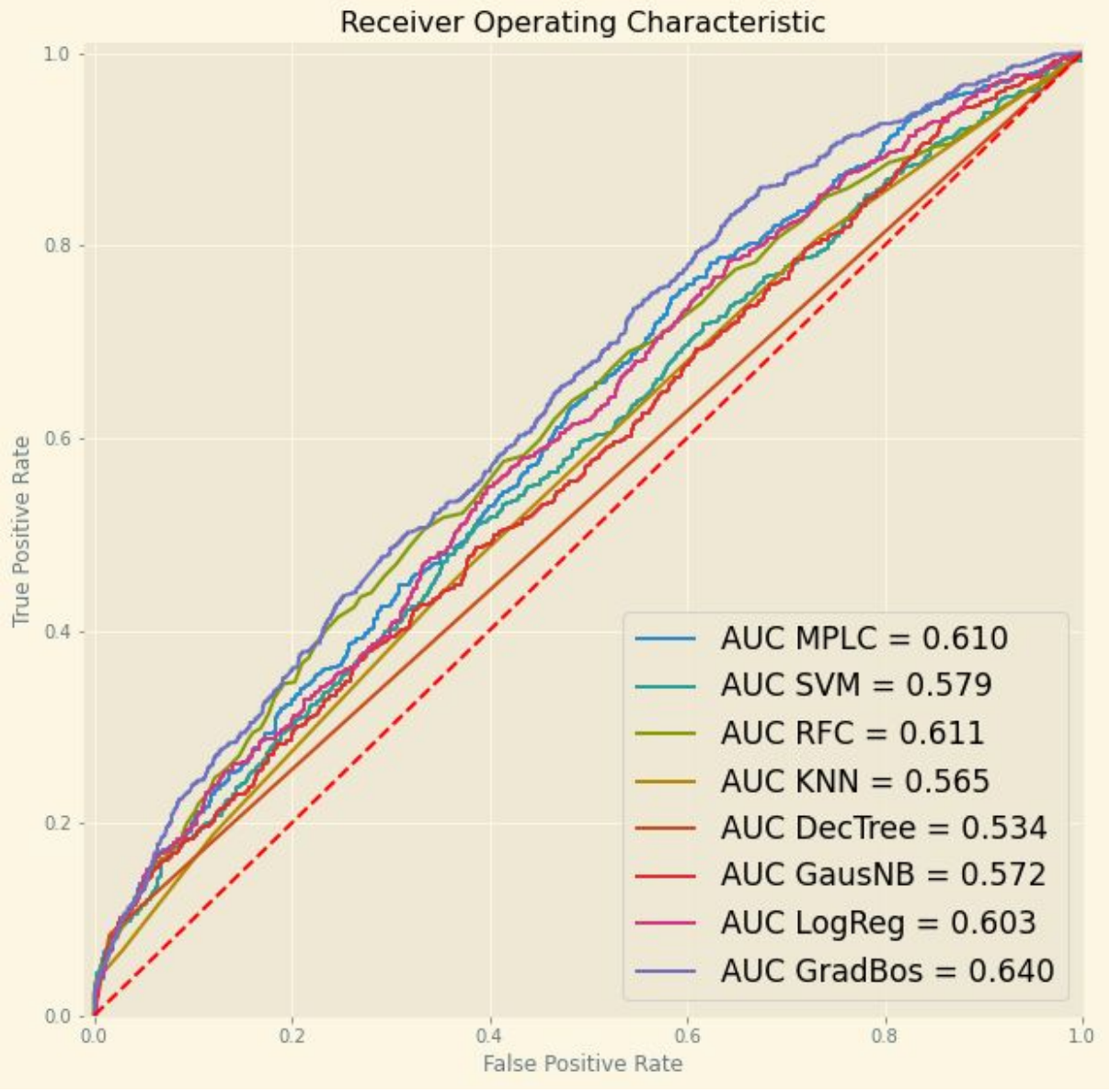
```
Accuracy score
mplc : 0.7570474777448071
SVM : 0.7577893175074184
RFC: 0.737759643916914
KNN : 0.7080860534124629
Decision Tree : 0.7614985163204748
Gaussian Nb : 0.3731454005934718
LogReg  : 0.7566765578635015
Grad_Bos : 0.7566765578635015


Precision score
mplc : 0.575
SVM : 0.5853658536585366
RFC: 0.42366412213740456
KNN : 0.3387096774193548
Decision Tree : 0.6363636363636364
Gaussian Nb : 0.26264501160092807
LogReg  : 0.5714285714285714
Grad_Bos : 0.5604395604395604


recall_score
mplc : 0.06896551724137931
SVM : 0.07196401799100449
RFC: 0.1664167916041979
KNN : 0.1889055472263868
Decision Tree : 0.08395802098950525
Gaussian Nb : 0.848575712143928
LogReg  : 0.06596701649175413
Grad_Bos : 0.07646176911544228


cohen_kappa_score
mplc : 0.07409427052965811
SVM : 0.07823799580714552
RFC: 0.11554302493276347
KNN : 0.07945868273449219
Decision Tree : 0.09622081989981679
Gaussian Nb : 0.037415846985068946
LogReg  : 0.07068970409425446
Grad_Bos : 0.07990847862888573
```

```
F-Score
mplc : 0.12315930388219544
SVM : 0.1281708945260347
RFC: 0.23896663078579117
KNN : 0.24254090471607312
Decision Tree : 0.14834437086092717
Gaussian Nb : 0.40113394755492554
LogReg  : 0.11827956989247314
Grad_Bos : 0.13456464379947228
```



5.Hyperparameter Tuning

Dari model terbaik, dilakukan hyperparameter tuning untuk meningkatkan akurasi. Diperoleh hasil sebagai berikut

a. KNN

Hyperparameter Tuning untuk parameter grid berikut
param_grid = {'n_neighbors':np.arange(5, 50), 'weights':['distance', 'uniform']}
Diperoleh hyperparameter terbaik adalah

```
{'n_neighbors': 48, 'weights': 'uniform'}
Accuracy Score:
0.7570474777448071
```

Dapat dilihat bahwa nilai akurasi naik dari sebelum dilakukan tuning

b. Namun untuk Decision Tree, untuk hyperparameter tuning malah menurunkan akurasi menjadi 0.7158753709198813 (sebelumnya 0.76)

{'random_state': 43, 'min_samples_split': 4, 'min_samples_leaf': 11, 'max_features': 'log2'}

c. XGBoost

```
{'subsample': 0.9, 'random_state': 43, 'n_estimators': 650, 'min_child_weight': 1, 'max_depth': 4, 'learning_rate': 0.01, 'c
olsample_bytree': 0.9}
Accuracy Score:
0.7577893175074184
```

d. Logistic Regression

```
{'penalty': 'l2', 'max_iter': 5000, 'C': 21.544346900318832}
Accuracy Score:
0.7566765578635015
```

6.Ensemble Learning

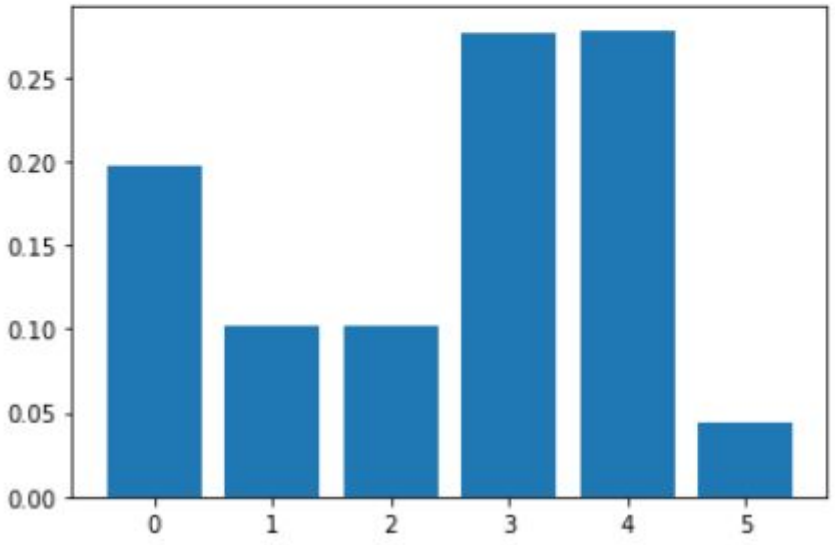
Masuk ke dalam pemodelan

7. Final Evaluation

Model terbaik yang dipilih yaitu decision tree dengan nilai akurasi 0.76. Model ini tidak overfitting karena nilai Cross Validationnya adalah 0.75

Feature yang paling penting secara berurutan adalah

```
Feature: 0, Score: 0.19752
Feature: 1, Score: 0.10149
Feature: 2, Score: 0.10165
Feature: 3, Score: 0.27712
Feature: 4, Score: 0.27851
Feature: 5, Score: 0.04372
```



Dengan feature 0:X1, 1 : X2, 2 : X3, 3 : X4, 4: X5, 5, X6
Artinya feature yang paling mempengaruhi label/target variable (kategori keaktifan customer) berdasarkan analisis random forest adalah feature X4 dan X5 yaitu jumlah unggahan di internet dalam 2 bulan terakhir dan jumlah unduhan di internet dalam 2 bulan terakhir. Untuk meningkatkan jumlah kategori pelanggan yang aktif maka dapat memperhatikan 2 variabel tersebut.

4. Kesimpulan

1. Informasi dan insight dari data dapat dilihat dalam EDA dan hasil analisis
2. Model klasifikasi terbaik adalah Decision Tree. Hasil prediksi *flag customer churn or active* terlampir dalam file xlsx
3. Feature yang paling penting secara berurutan adalah

Maka strategi untuk memaksimumkan jumlah customer yang aktif adalah dengan fokus pada feature tersebut secara berurutan. Hal ini sejalan dengan prinsip Pareto 80/20 yang mengungkapkan bahwa 20% input mempengaruhi 80% output, jadi perusahaan bisa melakukan efisiensi dengan fokus pada sedikit feature namun memperoleh hasil yang lebih besar

Daftar Pustaka

https://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_linear_regression, Diakses pada 8 November 2020
https://iq.opengenus.org/gaussian-naive-bayes/, Diakses pada 8 November 2020
https://www.advernesia.com/blog/data-science/pengertian-dan-cara-kerja-algoritma-k-nearest-neighbours-knn/, Diakses pada 8 November 2020
https://id.wikipedia.org/wiki/Random_forest, Diakses pada 8 November 2020
https://ichi.pro/id/normalisasi-vs-standardisasi-mana-yang-lebih-baik-2667604896, Diakses pada 8 November 2020
https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/, Diakses pada 8 November 2020