Final Project Presentation

Nomor Kelompok: Kelompok Stumble
Nama Mentor: Erwin
Nama:

- Dominicus Christian Bagus Susanto
- Rizka Mahdalela

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Petunjuk

- Waktu presentasi adalah 5 menit (tentatif, tergantung dari banyaknya kelompok yang mendaftarkan diri)
- Waktu tanya jawab adalah 5 menit
- Silakan menambahkan gambar/visualisasi pada slide presentasi
- Upayakan agar tetap dalam format poin-poin (ingat, ini presentasi, bukan esai)
- Jangan masukkan *code* ke dalam slide presentasi (tidak usah memasukan screenshot jupyter notebook)

Latar Belakang

Latar Belakang Project

Sumber Data:
https://www.kaggle.com/datasets/blastchar/telco-customer-churn?resource=download

Problem: **classification**

Tujuan:

- Memprediksi dan menganalisis perilaku pelanggan yang mempertahankan ataupun meninggalkan langganan dan mengembangkan program retensi

Explorasi Data dan Visualisasi

Business Understanding

(Amaresan, 2021) Customer churn adalah persentase pelanggan yang berhenti berlangganan suatu bisnis tertentu.

(Yunita, 2019) Churn dihitung dari berapa banyak pelanggan yang meninggalkan bisnis dalam waktu tertentu.

Customer churn harus diminimalisasi karena bisnis akan mengalami kerugian besar jika kehilangan pelanggan, karena faktanya mendapat pelanggan baru 5 kali lebih mahal daripada mempertahankan pelanggan yang sudah ada.

Data Cleansing

Data terdiri dari 7042 baris dan 21 kolom

Tidak terdapat missing value ketika pertama kali pengecekan

Namun ketika mengubah data bertipe object ke float dan

dan integer, ditemukan beberapa missing value berjumlah 11.

11 row data tersebut dihapus karena tidak terlalu berpengaruh

Terhadap 7000 data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        7043 non-null   object 
 1   gender            7043 non-null   object 
 2   SeniorCitizen     7043 non-null   int64  
 3   Partner           7043 non-null   object 
 4   Dependents        7043 non-null   object 
 5   tenure            7043 non-null   int64  
 6   PhoneService      7043 non-null   object 
 7   MultipleLines     7043 non-null   object 
 8   InternetService   7043 non-null   object 
 9   OnlineSecurity    7043 non-null   object 
 10  OnlineBackup      7043 non-null   object 
 11  DeviceProtection  7043 non-null   object 
 12  TechSupport       7043 non-null   object 
 13  StreamingTV       7043 non-null   object 
 14  StreamingMovies   7043 non-null   object 
 15  Contract          7043 non-null   object 
 16  PaperlessBilling  7043 non-null   object 
 17  PaymentMethod     7043 non-null   object 
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object 
 20  Churn             7043 non-null   object 
```

Data Cleansing

Melakukan encoding terhadap data kategorikal menjadi data numerikal agar dapat dikalkulasi

```
{'Contract': {0: 'Month-to-month', 1: 'One year', 2: 'Two year'},
 'Dependents': {0: 'No', 1: 'Yes'},
 'DeviceProtection': {0: 'No', 1: 'Yes', 2: 'No internet service'},
 'InternetService': {0: 'DSL', 1: 'Fiber optic', 2: 'No'},
 'MultipleLines': {0: 'No phone service', 1: 'No', 2: 'Yes'},
 'OnlineBackup': {0: 'Yes', 1: 'No', 2: 'No internet service'},
 'OnlineSecurity': {0: 'No', 1: 'Yes', 2: 'No internet service'},
 'PaperlessBilling': {0: 'Yes', 1: 'No'},
 'Partner': {0: 'Yes', 1: 'No'},
 'PaymentMethod': {0: 'Electronic check',
                   1: 'Mailed check',
                   2: 'Bank transfer (automatic)',
                   3: 'Credit card (automatic)'},
 'PhoneService': {0: 'No', 1: 'Yes'},
 'StreamingMovies': {0: 'No', 1: 'Yes', 2: 'No internet service'},
 'StreamingTV': {0: 'No', 1: 'Yes', 2: 'No internet service'},
 'TechSupport': {0: 'No', 1: 'Yes', 2: 'No internet service'}}
```

Data Cleansing

Melakukan feature selection berdasarkan Variance Threshold sebesar 0.1.

(Bex, 2021) melakukan VT dapat menambah performa atau setidaknya mengurangi kompleksitas model

Hasil dari feature selection ini adalah fitur PhoneService dihapus

Exploratory Data Analysis

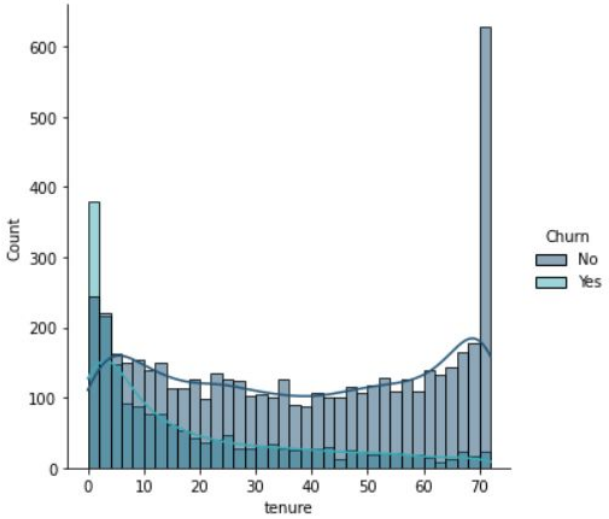
Insight 1

Pelanggan memiliki presentase churn tinggi

Ketika masa berlangganan atau tenure masih rendah,

Sedangkan ketika pelanggan sudah lama berlangganan,

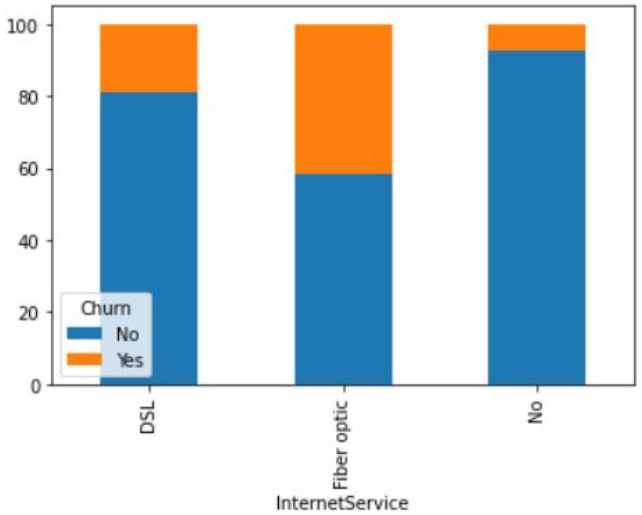
Presentase churn kecil atau rendah.



Exploratory Data Analysis

Insight 2

Pelanggan yang berlangganan menggunakan internet service fiber optic lebih banyak yang churn daripada pelanggan yang menggunakan internet service lain.



Exploratory Data Analysis

Insight 3

Pelanggan yang berlangganan dengan kontrak

Month-to-month lebih berpotensi untuk churn daripada

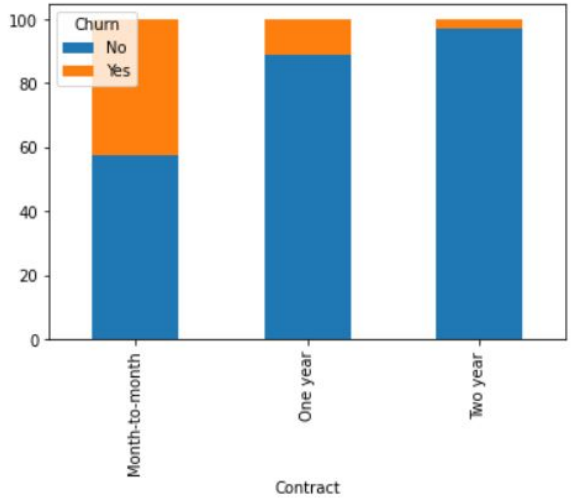
Kontrak one year dan two year. Hal ini dapat berkaitan

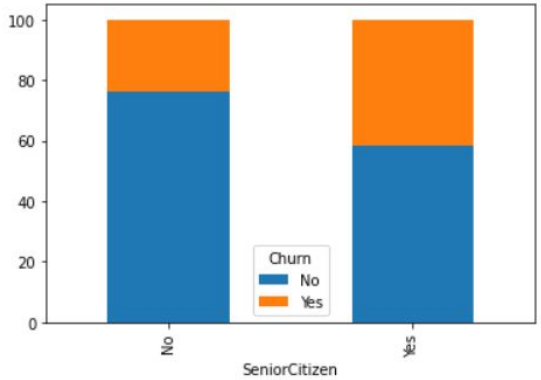
Dengan presentase senior citizen yang lebih banyak churn

Daripada yang bukan senior citizen.

Dapat disimpulkan bahwa senior citizen yang berlangganan

Month-to-month memiliki presentase churn yang tinggi.





Modelling

Modelling

Menggunakan train test split dengan rasio 80:20 dengan random state 101
Tidak melakukan rescaling data dan teknik PCA karena performa akurasi yang lebih buruk

```
KNN: 0.29 akurasi dengan standar deviasi 0.003900
Decision Tree: 0.57 akurasi dengan standar deviasi 0.005802
```

Model awal yang digunakan KNN, Multinomial Naive Bayes, DT

```
KNN: 0.76 akurasi dengan standar deviasi 0.008139
NB: 0.67 akurasi dengan standar deviasi 0.009082
Decision Tree: 0.73 akurasi dengan standar deviasi 0.014526
```

K-Nearest Neighbor

Parameters
N_neighbors = 3, 5, 10
Algorithm = 'ball_tree', 'kd_tree', 'brute'
Weights = 'uniform', 'distance'

	algorithm	n_neighbors	weights	Training accuracy	Testing accuracy	Ranking
4	ball_tree	10	uniform	0.808089	0.776711	1
10	kd_tree	10	uniform	0.808178	0.776533	2
16	brute	10	uniform	0.808222	0.776356	3

Multinomial Naive Bayes

Parameters
Alpha = 0.1, 1.0, 2.0, 3.0, 1.5
fit_prior = True, False

	alpha	fit_prior	Training accuracy	Testing accuracy	Ranking
0	0.1	True	0.673289	0.673956	1
4	2.0	True	0.673067	0.673956	1
2	1.0	True	0.673156	0.673778	3

Decision Tree

Parameters
splitter = ‘best’, ‘random’
criterion = ‘gini’, ‘entropy’
random _state = ‘1, 51, 101

	criterion	random_state	splitter	Training accuracy	Testing accuracy	Ranking
6	entropy	1	best	0.998622	0.734222	1
8	entropy	51	best	0.998622	0.733867	2
10	entropy	101	best	0.998622	0.733333	3

Model Akhir

Hasil K-Nearest Neighbour:

	precision	recall	f1-score	support
0	0.82	0.93	0.87	1052
1	0.65	0.40	0.50	355
accuracy			0.79	1407
macro avg	0.74	0.66	0.68	1407
weighted avg	0.78	0.79	0.78	1407

Hasil Naive Bayes:

	precision	recall	f1-score	support
0	0.88	0.65	0.75	1052
1	0.41	0.74	0.53	355
accuracy			0.67	1407
macro avg	0.65	0.69	0.64	1407
weighted avg	0.76	0.67	0.69	1407

Hasil Decision Tree:

	precision	recall	f1-score	support
0	0.83	0.80	0.81	1052
1	0.46	0.50	0.48	355
accuracy			0.72	1407
macro avg	0.64	0.65	0.65	1407
weighted avg	0.73	0.72	0.73	1407

Conclusion

Kesimpulan

Model paling baik untuk memprediksikan churn adalah K-Nearest Neighbor dengan tingkat akurasi mencapai 80%

Faktor yang harus diperhatikan stakeholder untuk menurunkan churn adalah memperbaiki kualitas produk internet service yang menggunakan fiber optic. Kemudian, stakeholder juga harus mampu untuk menarik perhatian customer pada usia remaja hingga bekerja (non-senior citizen) yang terbukti memiliki presentase churn rendah

Referensi

https://blog.hubspot.com/service/what-is-customer-churn

https://towardsdatascience.com/how-to-use-variance-thresholding-for-robust-feature-selection-a4503f2b5c3f

https://scialert.net/fulltext/?doi=jas.2014.171.176. **How the Parameters of K-nearest Neighbor Algorithm Impact on the Best Classification Accuracy: In Case of Parkinson Dataset**

Terima kasih!

Ada pertanyaan?