Final Project – Implementasi Metode Random Forest Pada Prediksi Customer Churn

By : Siti Uyun Mubarak

1. Background

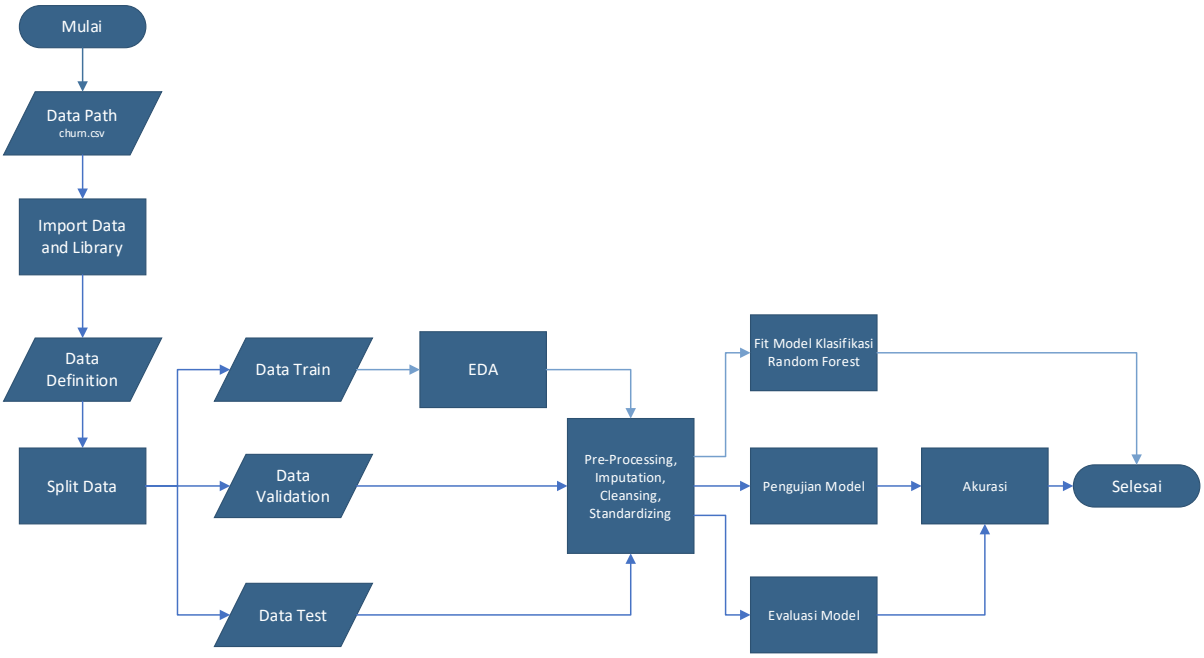
Banyaknya persaingan antar perusahaan saat ini menjadi tantangan utama yang harus di hadapi oleh sebuah perusahaan. Berbagai inovasi dan strategi khusus perlu dilakukan agar perusahaan memiliki langkah yang lebih maju dibandingkan perusahaan lainnya dan tetap bisa bertahan. Ada banyak faktor yang dapat meningkatkan penjualan produk suatu perusahaan, salah satunya yaitu pelanggan. Akan tetapi untuk memperoleh pelanggan baru memerlukan biaya hingga 10 kali lipat lebih mahal dibandingkan biaya untuk mempertahankan pelanggan yang ada [1]. Maka dari itu, perusahaan perlu untuk memprediksi pelanggannya agar dapat mengetahui tingkat loyalitas pelanggan dengan melakukan prediksi *customer churn*.

2. Objectives

Dari permasalahan yang telah dijabarkan diatas, dibuat sebuah solusi untuk dapat melakukan prediksi *customer churn* menggunakan teknik *machine learning*. Dimana data yang diperoleh pada *project* ini berasal dari sebuah situs web berdasarkan fitur yang telah disediakan. Setiap pengguna diberikan nilai prediksi yang akan memperkirakan status *churn* mereka pada waktu tertentu. Nilai ini didasarkan pada informasi demografis pengguna, perilaku saat menjelajahi situs web, data pembelian historis dan lainnya. Dengan adanya prediksi *customer churn* menggunakan teknik *machine learning* diharapkan dapat membantu perusahaan dalam mengambil keputusan atau tindakan saat adanya pelanggan yang terdeteksi *churn*.

3. Project Process

Pada *project* ini, terdapat beberapa informasi seperti umur pelanggan, media yang digunakan pelanggan untuk bertransaksi, wilayah tempat pelanggan, kategori keanggotaan pelanggan, jenis penawaran yang disukai pelanggan, umpan balik pelanggan dan lain-lain. Model yang akan digunakan adalah Random Forest. Untuk alur dari proses yang akan dilakukan dapat dilihat pada gambar berikut.



Gambar 1. Flowchart Project Process

4. Data Insight

Dataset pada *project* ini memiliki 22 fitur dan 1 target. Pendeskripsian data dapat dilihat pada tabel di bawah ini:

No.	Nama Fitur	Penjelasan
1.	*age*	umur pelanggan
2.	*gender*	jenis kelamin pelanggan
3.	*security_no*	nomor keamanan yang unik untuk mengidentifikasi seseorang
4.	*region_category*	wilayah tempat pelanggan
5.	*membership_category*	kategori keanggotaan yang digunakan pelanggan
6.	*joining_date*	tanggal ketika pelanggan menjadi anggota
7.	*joined_through_referral*	apakah pelanggan saat bergabung menggunakan kode atau ID rujukan
8.	*referral_id*	id referral
9.	*preferred_offer_types*	jenis penawaran yang disukai pelanggan
10.	*medium_of_operation*	media yang digunakan pelanggan untuk bertransaksi
11.	*internet_option*	layanan internet yang digunakan pelanggan
12.	*last_visit_time*	terakhir kali pelanggan mengunjungi website
13.	*days_since_last_login*	merepresentasikan hari sejak pelanggan terakhir mengunjungi situs web

No.	Nama Fitur	Penjelasan
14.	*avg_time_spent*	rata-rata waktu yang dihabiskan oleh pelanggan di situs web
15	*avg_transaction_value*	rata-rata transaksi yang dilakukan oleh pelanggan
16.	*avg_frequency_login_days*	rata-rata pelanggan *login* di situs web
17.	*points_in_wallet*	poin yang diberikan pelanggan pada setiap transaksi
18.	*used_special_discount*	apakah pelanggan menggunakan diskon yang ditawarkan
19.	*offer_application_preference*	apakah pelanggan suka penawaran yang direkomendasikan
20.	*past_complaint*	apakah pelanggan mengajukan keluhan
21.	*complaint_status*	keluhan yang diajukan oleh pelanggan diselesaikan
22.	*feedback*	umpan balik yang diberikan pelanggan
23.	*churn_risk_score*	merepresentasikan churn/no churn

Selain itu, terdapat 36.992 baris dan tidak adanya duplikat pada dataset. Gambaran mengenai dataset dapat dilihat di bawah ini.

	age	gender	security_no	region_category	membership_category	joining_date	joined_through_referral	referral_id	preferred_offer_types	medium_of_operatic
0	18	F	XW0DQ7H	Village	Platinum Membership	2017-08-17	No	xxxxxxxx	Gift Vouchers/Coupons	
1	32	F	5K0N3X1	City	Premium Membership	2017-08-28	?	CID21329	Gift Vouchers/Coupons	Deskt
2	44	F	1F2TCL3	Town	No Membership	2016-11-11	Yes	CID12313	Gift Vouchers/Coupons	Deskt
3	37	M	VJGJ33N	City	No Membership	2016-10-29	Yes	CID3793	Gift Vouchers/Coupons	Deskt
4	31	F	SVZXCWB	City	No Membership	2017-09-12	No	xxxxxxxx	Credit/Debit Card Offers	Smartpho

Gambar 2. Dataset

Selanjutnya, dilakukan pengecekan *missing value* pada setiap kolom. Untuk kolom numerik terdapat *missing value* pada "days_since_last_login" dan "points_in_wallet" sedangkan untuk kolom kategorikal *missing value* terdapat pada "gender", "region_category", "joined_through_referral", dan "medium_of_operation".

	age	days_since_last_login	points_in_wallet
0	40	NaN	548.87
1	49	17.0	773.76

Gambar 3. Missing Value pada Kolom Numerik

	gender	region_category	membership_category	joined_through_referral	preferred_offer_types	medium_of_operation	internet_option	used_special_discount
0	M	NaN	Basic Membership	No	Credit/Debit Card Offers	Smartphone	Fiber_Optic	Yes
1	M	NaN	Gold Membership	Yes	Gift Vouchers/Coupons	NaN	Mobile_Data	Yes

Gambar 4. Missing Value pada Kolom Kategorikal

Nilai median digunakan untuk imputasi kolom numerik sedangkan imputasi kolom kategorikal menggunakan nilai "KOSONG". Beberapa kolom yang berkaitan dengan waktu atau tanggal akan dilakukan penghapusan, hal tersebut dikarenakan permasalahan ini tidak berfokus pada data *time series*.

Setelah melakukan imputasi pada data, dilakukan *standarization* data agar setiap data yang ada tidak memiliki jarak yang terlalu tinggi. Berikut ini hasil perbandingan data sebelum dilakukan *standarization* dan sesudah dilakukan *standarization* data.

	age	days_since_last_login	points_in_wallet	gender_F	gender_M	region_category_City	region_category_KOSONG	region_category_Town	region_category_V
0	41.0	14.0	793.811069	1.0	0.0	1.0	0.0	0.0	
1	46.0	6.0	768.130000	0.0	1.0	0.0	0.0	1.0	
2	51.0	19.0	774.780000	1.0	0.0	1.0	0.0	0.0	
3	40.0	16.0	504.670000	0.0	1.0	1.0	0.0	0.0	
4	18.0	9.0	755.690000	1.0	0.0	0.0	0.0	0.0	

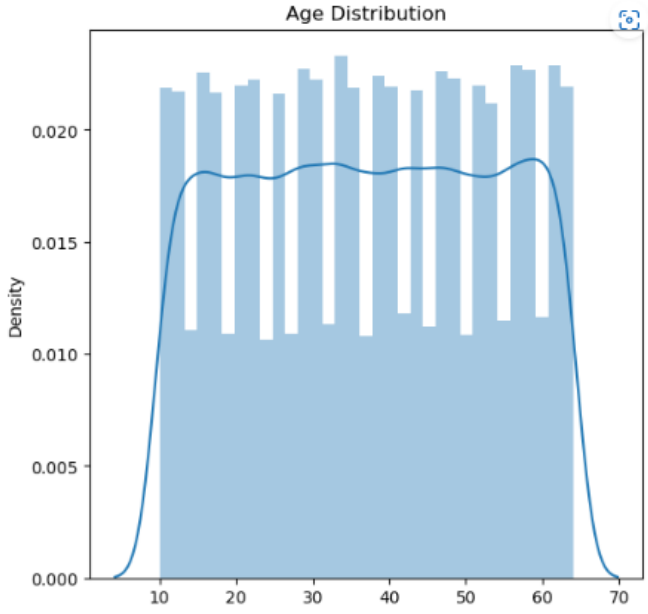
Gambar 5. Sebelum *Standardization*

	age	days_since_last_login	points_in_wallet	gender_F	gender_M	region_category_City	region_category_KOSONG	region_category_Town	region_categ
0	0.242058	0.223612	0.579109	1.002518	-1.002518	1.385093	-0.41431	-0.786961	
1	0.557015	-1.250490	0.435060	-0.997489	0.997489	-0.721973	-0.41431	1.270711	
2	0.871973	1.144925	0.472361	1.002518	-1.002518	1.385093	-0.41431	-0.786961	
3	0.179066	0.592137	-1.042726	-0.997489	0.997489	1.385093	-0.41431	-0.786961	
4	-1.206747	-0.697702	0.365283	1.002518	-1.002518	-0.721973	-0.41431	-0.786961	

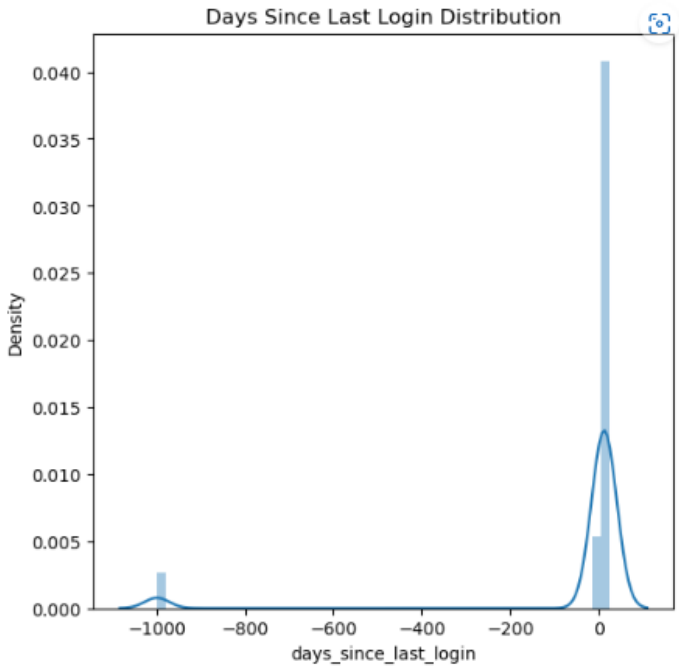
Gambar 6. Sesudah *Standardization*

5. Exploratory Data Analysis (EDA)

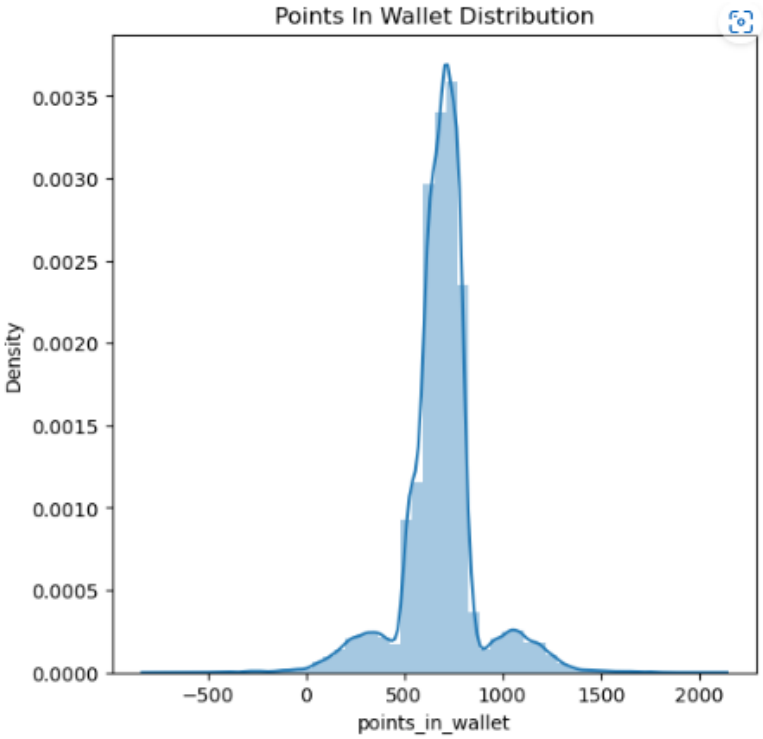
Pada tahap Exploratory Data Analysis (EDA) dilakukan pendistribusian pada data numerik. Kolom "days_last_login" memiliki *left skewed* yang paling tinggi. Untuk detail distribusi dari setiap kolom dapat dilihat pada gambar berikut.



Gambar 7. Distribusi Kolom Umur

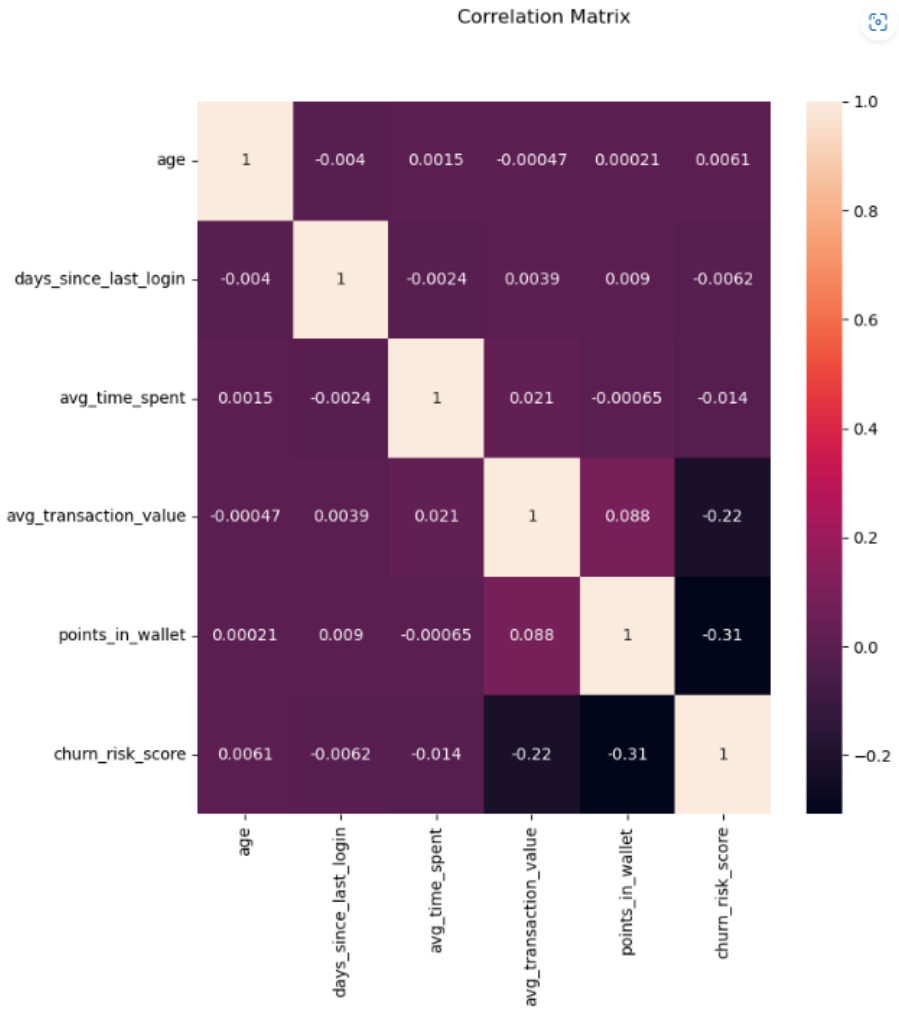


Gambar 8. Distribusi Kolom Days_Since_Last_login



Gambar 9. Distribusi Kolom Points_in_Wallet

Selanjutnya dilakukan pembuatan *correlation heat map* untuk melihat korelasi antar variabel yang dapat dilihat pada gambar berikut.



Gambar 10. Korelasi Antar Variabel

Dari gambar diatas dapat diketahui bahwa setiap kolom memiliki korelasi yang lemah terhadap variabel "churn_risk_score". Korelasi paling tinggi hanya sebesar 0.0061 yaitu pada kolom "age" dan kolom "churn_risk_score".

Pendeskripsian data pada kolom numerikal dan kategorikal juga dilakukan untuk melihat nilai min, max dan lainnya. Untuk melihat lebih detail perhatikan gambar berikut.

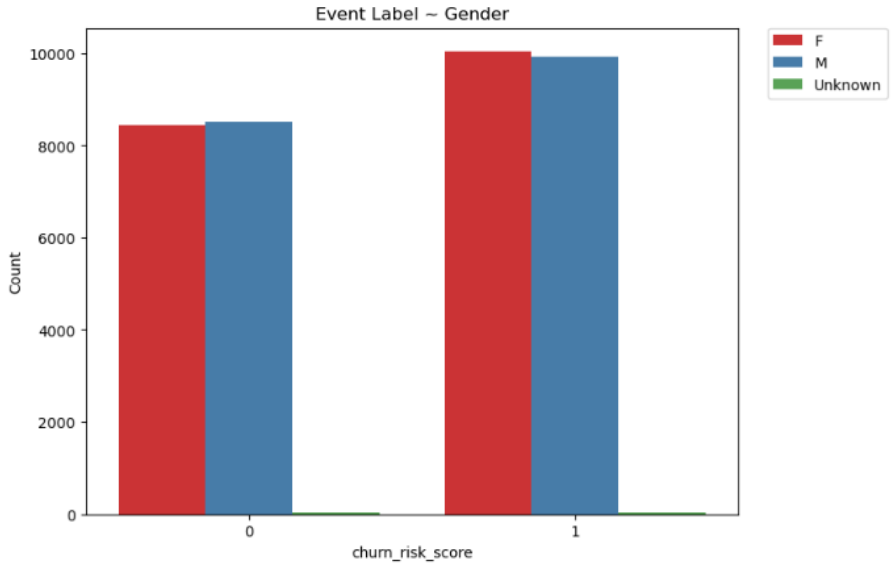
	age	days_since_last_login	avg_time_spent	avg_transaction_value	points_in_wallet	churn_risk_score
count	36992.000000	36992.000000	36992.000000	36992.000000	33549.000000	36992.000000
mean	37.118161	-41.915576	243.472334	29271.194003	686.882199	0.540982
std	15.867412	228.819900	398.289149	19444.806226	194.063624	0.498324
min	10.000000	-999.000000	-2814.109110	800.460000	-760.661236	0.000000
25%	23.000000	8.000000	60.102500	14177.540000	616.150000	0.000000
50%	37.000000	12.000000	161.765000	27554.485000	697.620000	1.000000
75%	51.000000	16.000000	356.515000	40855.110000	763.950000	1.000000
max	64.000000	26.000000	3235.578521	99914.050000	2069.069761	1.000000

Gambar 11. Deskripsi Kolom Numerik

	gender	security_no	region_category	membership_category	joining_date	joined_through_referral	referral_id	preferred_offer_types	medium_of_operatio
count	36992	36992	31564	36992	36992	36992	36992	36704	3699
unique	3	36992	3	6	1096	3	11359	3	
top	F	XW0DQ7H	Town	Basic Membership	2015-06-02	No	xxxxxxxx	Gift Vouchers/Coupons	Deskto
freq	18490	1	14128	7724	55	15839	17846	12349	1391

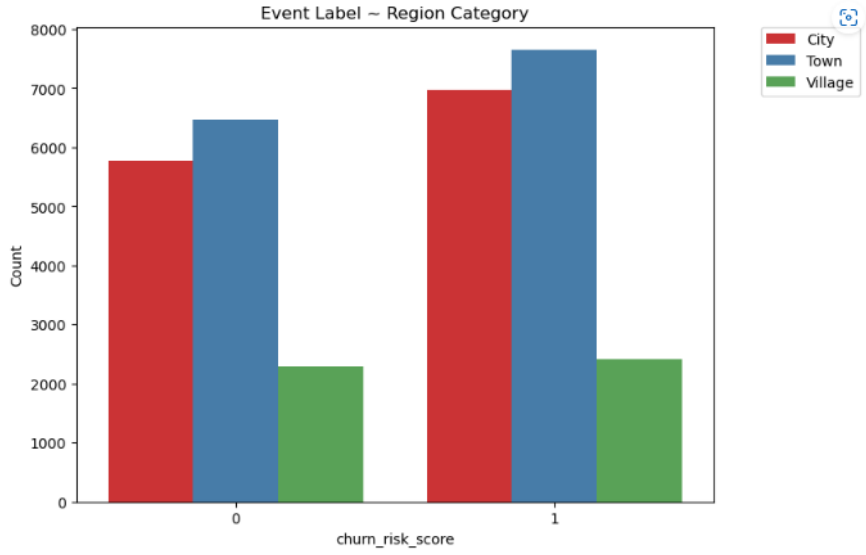
Gambar 12. Deskripsi Kolom Kategorik

Agar dapat mengetahui banyaknya nilai kolom kategorikal yang ada pada kolom target "churn_risk_score" dilakukan visualisasi *bar plot* seperti berikut.



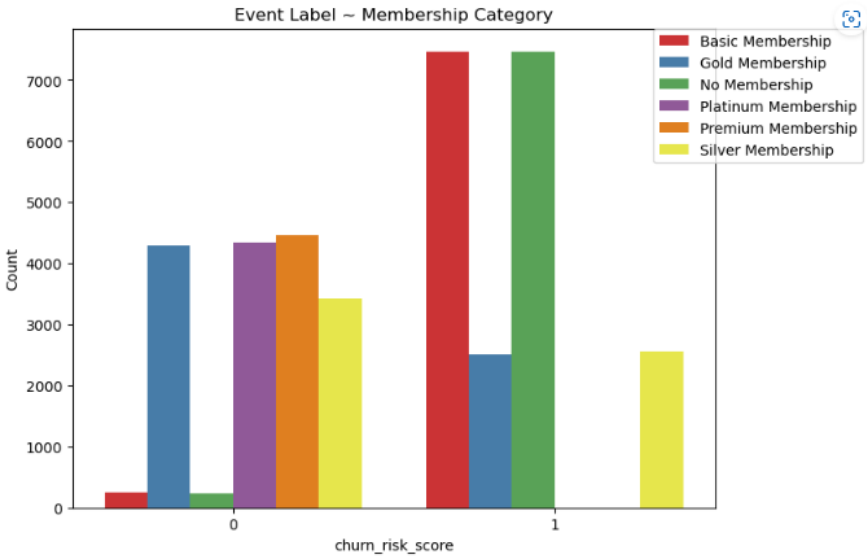
Gambar 13. Kolom Gender terhadap Churn Risk Score

Dari gambar diatas terlihat bahwa lebih banyak perempuan yang mendominasi dibandingkan laki-laki saat kondisi *churn*. Selanjutnya untuk "*region_category*" lebih banyak terjadi di wilayah *town* dibandingkan *city* atau *village*. Visualisasi tersebut dapat dilihat oleh *bar plot* dibawah ini.



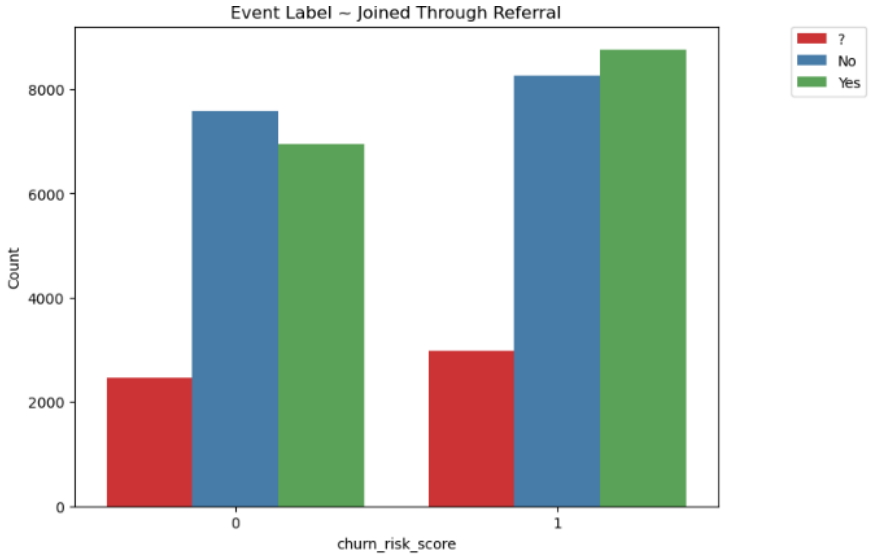
Gambar 14. Kolom Region Category terhadap Churn Risk Score

Kemudian untuk "event_label" pada saat kondisi *churn* di dominasikan oleh "No Membership" dan juga "Basic Membership". Sedangkan pada kondisi *no churn* didominasi oleh "Premium Membership". Detail visualisasi dapat dilihat pada gambar berikut.



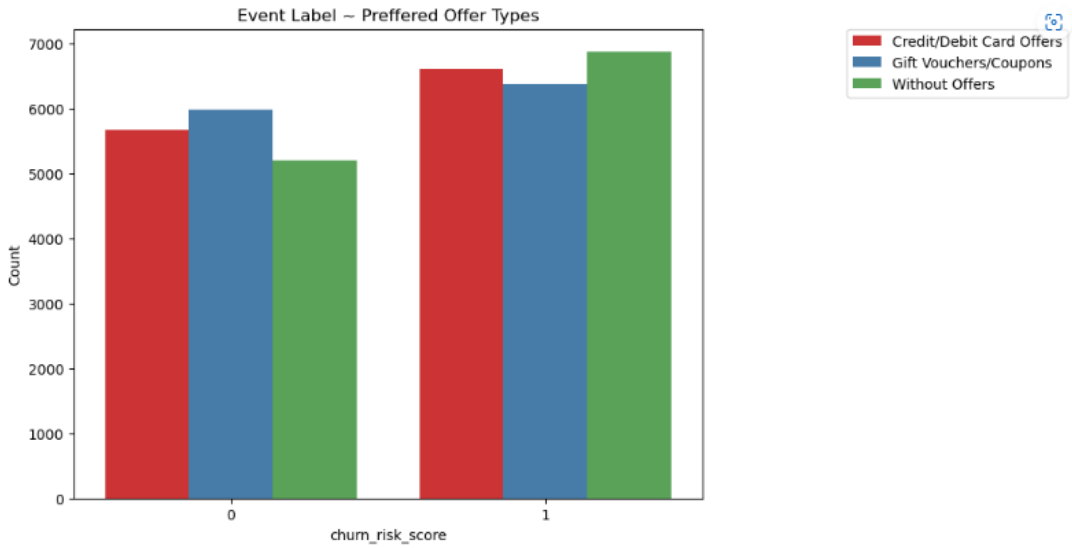
Gambar 15. Kolom Membership Category terhadap Churn Risk Score

Selanjutnya kondisi *churn* ketika *join* melalui *kode referral* lebih banyak terjadi dibandingkan dengan tidak melalui kode *referral*. Visualisasi tersebut dapat dilihat pada gambar berikut.



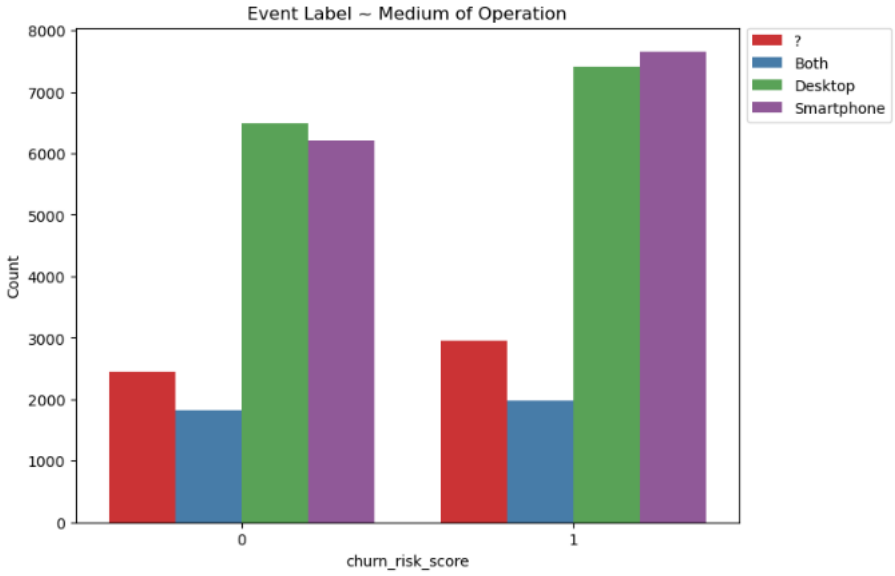
Gambar 16. Kolom Joined Through Referral terhadap Churn Risk Score

Kemudian untuk jenis penawaran yang disukai pelanggan, "without_offers" paling banyak terjadi pada saat kondisi *churn* sedangkan *no churn* di dominasi oleh "gifts voucher/coupons". Visualisasi pada kolom tersebut dapat dilihat pada gambar berikut.



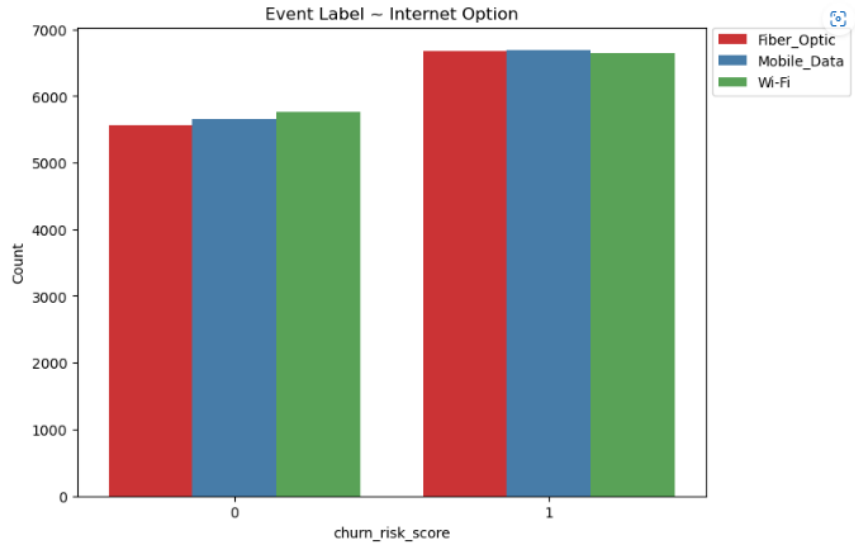
Gambar 17. Kolom Preffered Offer Types terhadap Churn Risk Score

Untuk medium yang paling banyak digunakan pelanggan saat kondisi *churn* adalah *smartphone* sedangkan untuk kondisi *no churn*, penggunaan medium paling banyak menggunakan *desktop*. Detail dari visualisasi kolom tersebut dapat dilihat pada gambar berikut.



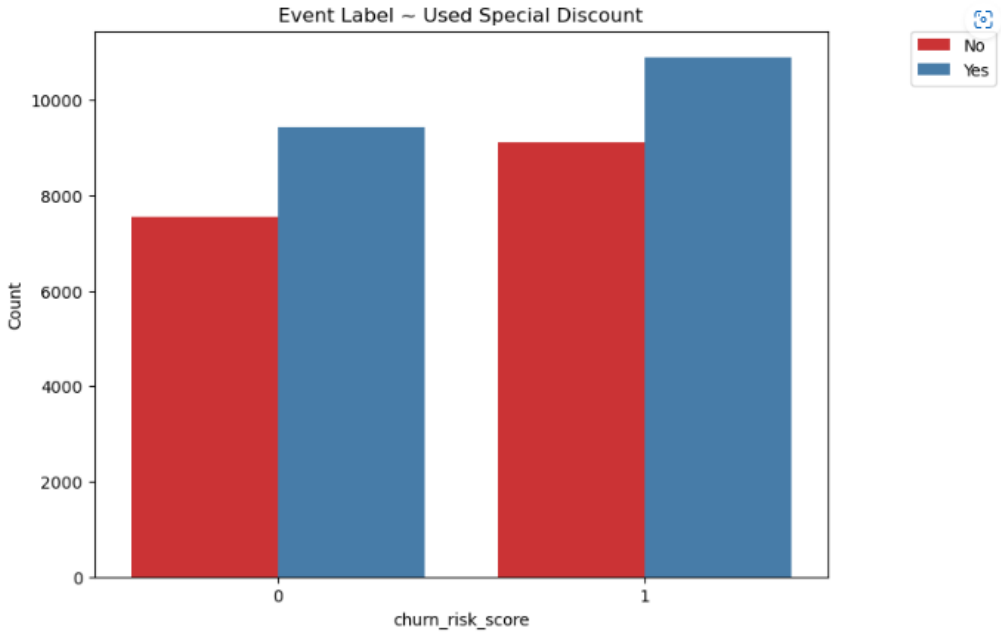
Gambar 18. Kolom Medium of Operation terhadap Churn Risk Score

Selanjutnya untuk *internet option* pada kondisi *churn*, penggunaan *mobile data, wifi* dan *fiber optic* ketiganya memiliki rentang nilai yang tidak jauh berbeda. Visualisasi pada kolom *internet option* dapat dilihat pada gambar berikut.

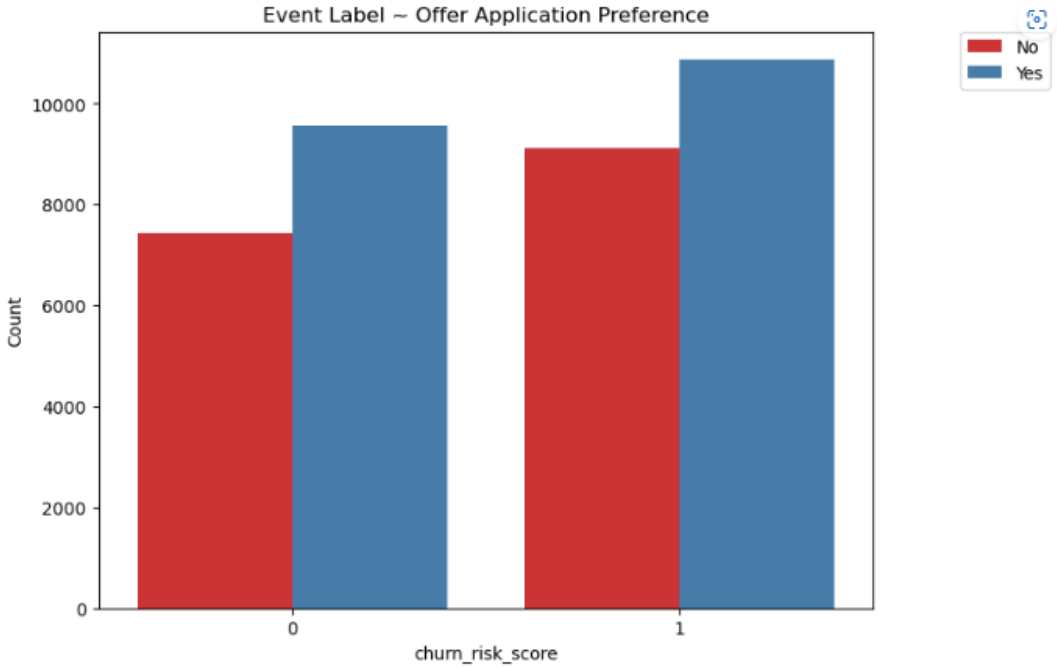


Gambar 19. Kolom Internet Option terhadap Churn Risk Score

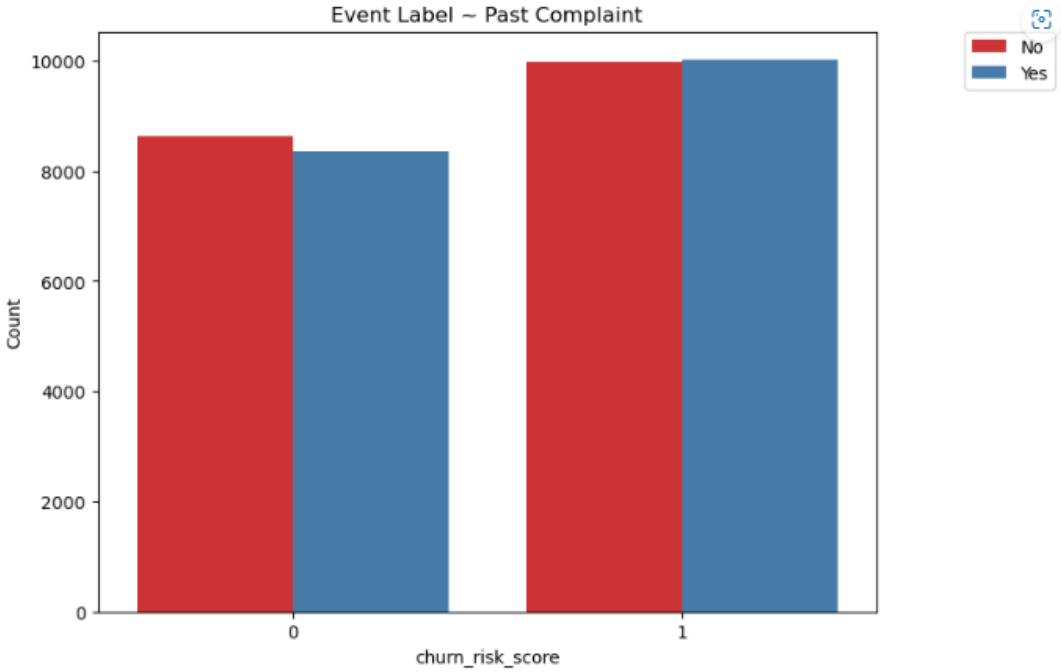
Kemudian untuk kolom *used special discount*, baik kondisi *churn* atau *no churn* keduanya paling banyak terjadi saat pelanggan menggunakan diskon spesial. Begitu pula dengan kolom *offer application preference*, baik kondisi *churn* atau *no churn* keduanya paling banyak terjadi saat pelanggan menyukai penawaran yang direkomendasikan. Hal yang sama juga terjadi pada kolom *complain status*, baik kondisi *churn* atau *no churn* paling banyak terjadi saat "complain not applicable". Lalu untuk kolom *past complain*, paling banyak terjadi ketika pelanggan mengajukan *complain* saat kondisi *churn*. Selanjutnya pada kolom *feedback* saat kondisi *churn* terjadi dinominasikan oleh "poor product quality" dan untuk "no churn" di dominasikan oleh "too many ads". Visualisasi dari lima kolom tersebut terhadap *churn risk score* dapat dilihat pada gambar dibawah ini.



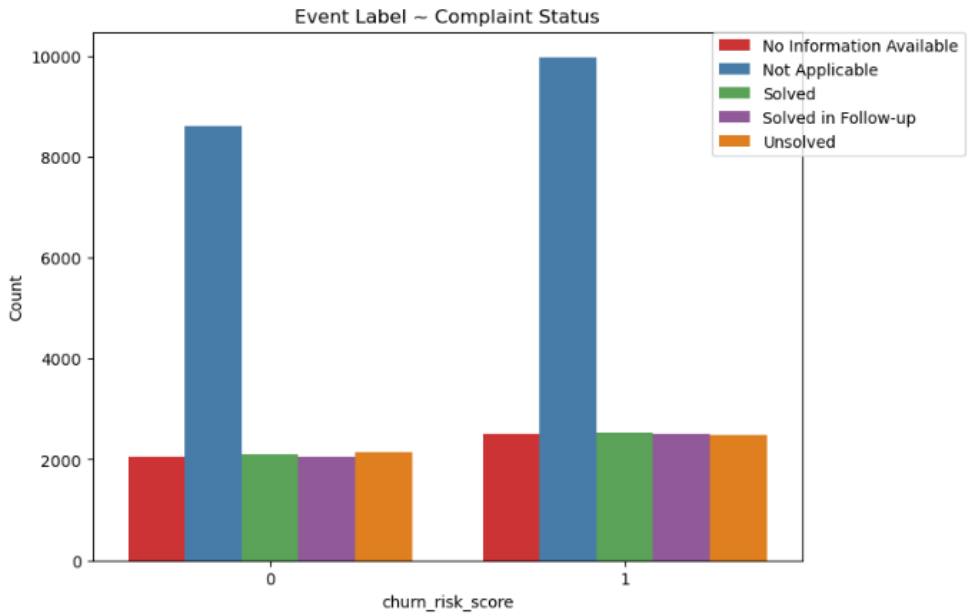
Gambar 20. Kolom Used Special Discount terhadap Churn Risk Score



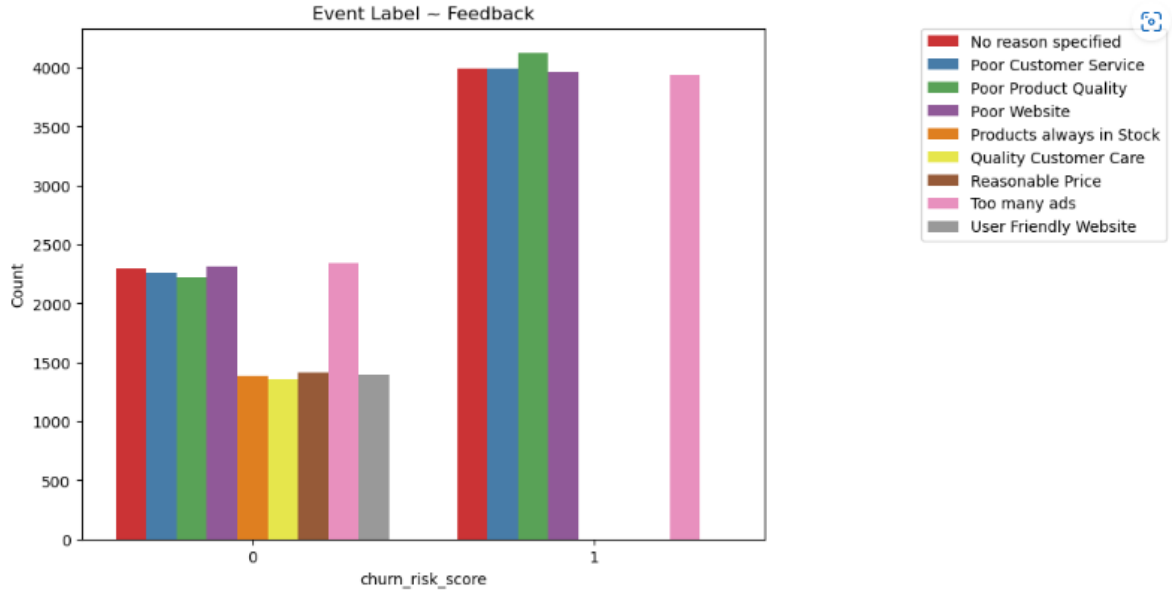
Gambar 21. Kolom Offer Application Preference terhadap Churn Risk Score



Gambar 22. Kolom Past Complaint terhadap Churn Risk Score



Gambar 23. Kolom Complaint Status terhadap Churn Risk Score



Gambar 24. Kolom Feedback terhadap Churn Risk Score

6. Modelling Process

Pada saat melakukan pemodelan pada metode *random forest*, dilakukan *hyperparameter* pada "max_*depth*", "criterion", "max_features", dan "n_estimators". Selanjutnya didapatkan hasil parameter terbaik yang dapat dilihat pada gambar berikut.

```
{'criterion': 'gini',
 'max_depth': 5,
 'max_features': 'log2',
 'n_estimators': 100}
```

Gambar 25. Best Parameter

Dari pemilihan parameter tebaik yang telah berhasil didapatkan, selanjutnya dilakukan *fit* ke dalam *data train*, dan dilakukan prediksi pada data validasi. Untuk melihat ketepatan dari sebuah prediksi digunakan *classification report* untuk melihat nilai akurasi, precision, recall dan f1 score.

Nilai akurasi didapatkan dari prediksi benar untuk data positif dan negatif dari keseluruhan data. Akurasi dapat menggambarkan keakuratan model klasifikasi yang digunakan. Nilai akurasi dapat diperoleh menggunakan persamaan berikut ini.

$$accuracy\ (\%) = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Precision adalah rasio yang memiliki prediksi nilai benar positif jika dibandingkan dengan keseluruhan hasil yang diprediksi positif. *Precision* dapat menggambarkan keakuratan data yang diinginkan dengan hasil prediksi yang diperoleh model klasifikasi. Nilai *precision* dapat diperoleh menggunakan persamaan berikut ini.

$$precision = \frac{TP}{TP + FP}$$

Recall adalah efektivitas dari pengklasifikasi dalam mengidentifikasi label positif. Nilai *Recall* dapat diperoleh menggunakan persamaan berikut ini.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score adalah nilai yang menandakan jika model yang dibangun memiliki nilai *precision* dan *recall* yang baik. Nilai *F1 score* dapat diperoleh menggunakan persamaan berikut ini.

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right)$$

Hasil dari data valid dapat dilihat pada gambar berikut.

	precision	recall	f1-score	support
0	0.83	0.80	0.82	2544
1	0.84	0.86	0.85	2997
accuracy			0.83	5541
macro avg	0.83	0.83	0.83	5541
weighted avg	0.83	0.83	0.83	5541

Gambar 26. Hasil Classification Report pada Validation Data

Sedangkan untuk melihat hasil *classification report* pada data test dapat dilihat pada gambar berikut.

	precision	recall	f1-score	support
0	0.82	0.80	0.81	2544
1	0.83	0.85	0.84	2999
accuracy			0.83	5543
macro avg	0.83	0.82	0.83	5543
weighted avg	0.83	0.83	0.83	5543

Gambar 26. Hasil Classification Report pada Data Testing

7. Kesimpulan

Model terbaik untuk melakukan prediksi cutomer churn adalah menggunakan *random forest* dengan *hyperparameter* pada "max_*depth"* 5, "criterion" bernilai gini, "max_features" log2 dan "n_estimators" 100. Pada project ini fitur menjadi semakin banyak setelah dilakukan *one hot encoding*. Sebaiknya dilakukan seleksi dan pemilihan fitur agar metode lain bisa diterapkan seperti metode KNN. Teknik *boosting* atau *stacking* juga dapat digunakan pada eksplorasi selanjutnya. Kemudian untuk *hyperparameter tuning* dapat dilakukan pengimplementasian dengan metode lain seperti randomsearchCV.

Reference

S. KhakAbi, M. R. Gholamian, and M. Namvar, "Data Mining Applications in Customer Churn Management," 2010 International Conference on Intelligent Systems, Modelling and Simulation, pp. 220–225, Jan. 2010

Masarifoglu, M., & Buyuklu, A. H. (2019). Applying Survival Analysis to 71 Telecom Churn Data. American Journal of Theoretical and Applied Statistics, 8(6), 261–275.

Cici Olivia, Indwiarti, Sibaroni Yulian (2015). Analisis Prediksi Churn Menggunakan Metode Logistic Regression dan Algoritma Decision Tree.Jurnal e-Proceeding of Engineering, Volume 2, Nomor 2.

Adhelia Nurfira Rachmi, (2020). Implementasi Metode Random Forest Dan Xgboost Pada Klasifikasi Customer Churn.

Dhea Laksnu Prianto, IIn Ernawati, Nurul Chamidah, (2022). Implementasi Churn Prediction Di Industri Telekomunikasi Dengan Metode Logistic Regression Dan Correlation-Based Feature Selection. Seminar Nasional Informatika, Sistem Informasi dan Keamanan Siber (SEINASI-KESI). Jakarta-Indonesia, Januari 2022.

Amanda Velia, R. S. Theodorus, N. S. Sandi, A. P. Anindya, Fajar Indrayatna. (2022). Klasifikasi Customer Churn Pada Perusahaan Telekomunikasi Menggunakan Support Vector Machine. Seminar Nasional Statistika Aktuaria I (2022).