Indonesian Journal of Data and Science (IJODAS) ISSN: 2715-9930
Vol 4, No 1, March 2023, pp. 31-37

Prediksi potensi donatur menggunakan model *Logistic Regression*

Sitti Rahmah Jabir[a,1], Huzain Azis[a,2], Dewi Widyawati[a,3], A. Ulfah Tenripada[a,4]

[a] *Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM.05, Makassar dan 90231, Indonesia*
[1] *rahmahjabir@umi.ac.id;* [2] *huzain.azis@umi.ac.id;* [3] *dewiwidyawati@umi.ac.id;* [4] *a.ulfah@umi.ac.id;*

INFORMASI ARTIKEL

Diterima : 02 – 01 – 2023
Direvisi : 28 – 02 – 2023
Diterbitkan : 31 – 03 – 2023

Kata Kunci:
Prediksi Donatur,
Missing Value,
Single Imputation.

ABSTRAK

GRDS menghadapi kelangkaan dana, ketika diperlukan untuk merawat para korban Gaja. Gaja adalah topan bernama kelima dari musim siklon Samudra Hindia Utara 2018 yang mempengaruhi sebagian besar tempat di Tamil Nadu, India selama bulan November 2018. Tujuan dari penelitian ini adalah untuk menggunakan riwayat donasi untuk menganalisis apakah donator akan menyumbang atau tidak menggunakan metode klasifikasi data mining yaitu regresi logistik. Data Tamil Nadu diberikan untuk menerapkan model yang dibangun untuk memprediksi donator yang paling mungkin menjadi korban topan Gaja. Pada tahap pengumpulkan data seringkali terjadi hambatan, salah satu hambatannya yaitu fenomena *missing data* atau data hilang. Akibat dari adanya *missing data* adalah pendugaan parameter menjadi tidak efisien. Ukuran data yang berkurang dapat mengakibatkan kesulitan dalam menganalisis, sehingga hasil yang didapatkan menjadi tidak valid dan tujuan dari penelitian tidak tercapai. Data yang hilang akan diisi menggunakan metode single imputation. Data yang telah diimputasi menggunakan beberapa metode akan membantu dalam melakukan prediksi. Dimana algoritma yang digunakan untuk melakukan prediksi ialah logistic regression. Beberapa data dihilangkan setelah melihat multikolinearitas. Dalam tahap pemodelan, data dibagi menjadi 2 yaitu 70% untuk data pelatihan dan 30% untuk data tes. Dimana hasil perhitungan akurasi dari model ialah 0,6129 yang menunjukkan bahwa model tidak melakukan prediksi dengan baik menggunakan metode tersebut.

I. Pendahuluan

GRDS atau yang dikenal sebagai Glenegals *Disaster Relief Service* (GRDS) menghadapi kelangkaan dana, ketika diperlukan untuk merawat para korban Gaja. Gaja adalah topan bernama kelima dari musim siklon Samudra Hindia Utara 2018 yang mempengaruhi sebagian besar tempat di Tamil Nadu, India selama bulan November 2018. Oleh karena itu, GRDS telah merencanakan kampanye penggalangan dana untuk melayani para korban. Data didapatkan dari histori donator sebelumnya, dimana data tersebut terdiri dari 27 variabel yang 4849 observasi. Dimana data tersebut diambil dari Kaggle.com [1].

Pada tahap proses pengumpulkan data seringkali terjadi hambatan, salah satu hambatannya yaitu fenomena *missing data* atau data hilang. *Missing data* adalah hilangnya sebagian informasi atau data pada suatu penelitian. Akibat dari adanya *missing data* adalah pendugaan parameter menjadi tidak efisien. Oleh karena itu, perlu dilakukan estimasi untuk mengisi data yang hilang tersebut agar hasil dari pengolahan data nantinya memiliki hasil yang maksimum[2].

Pada penelitian ini akan dilakukan imputasi data yang hilang dengan menggunakan metode mean imputation. *Mean imputation* merupakan salah satu metode imputasi yang paling umum digunakan. Imputasi dengan metode mean mengisi missing data dalam suatu variable dengan nilai rata-rata dari semua nilai yang diketahui pada suatu variable[3]. Setelah data yang hilang telah diisi, data tersebut akan dibangunkan sebuah model menggunakan metode regresi logistic. Berdasarkan penelitian yang dilakukan Wan Hanieza (2019), *logistic regression* dianggap model yang memberikan hasil yang baik dalam melakukan prediksi[4].

II. Metode

A. *Data Cleaning*

Singkatan didefinisikan pada penggunaan pertama di bagian isi meskipun telah didefinisikan pada Abstrak. Penggunaan singkatan judul tidak diperkenankan. Contoh penulisan singkatan yang benar pada IJODAS adalah *Artificial Intelligence* (AI) bukan ditulis AI (*Artificial Intelli*Pembersihan data atau yang dikenal sebagai data cleaning merupakan proses kompleks dan terdiri dari beberapa tahap yang meliputi penentuan aturan kualitas data, mendeteksi eror/kesalahan data, dan memperbaiki kesalahan[5].

Terdapat beberapa cara untuk menginput data yang hilang salah satunya menggunakan nilai rata-rata (*mean*). Imputasi menggunakan nilai mean merupakan salah satu metode imputasi yang paling umum digunakan. Imputasi dengan metode mean mengisi missing data dalam suatu variable dengan nilai rata-rata dari semua nilai yang diketahui pada suatu variable[3].

B. *Multicollinearity*

Multikolinearitas sering digambarkan sebagai fenomena statistik di mana ada hubungan yang sempurna atau tepat antara variabel prediktor. Dalam kejadian multikolinearitas, sulit untuk menghasilkan perkiraan koefisien individu yang dapat diandalkan untuk variabel prediktor dalam model yang menghasilkan kesimpulan yang salah tentang hubungan antara hasil dan variabel prediktor. Setelah metode imputasi, maka akan beralih ke langkah selanjutnya yaitu memeriksa kolinearitas antar variabel[6].

Multikolinearitas dikenal sebagai suatu kondisi dimana terjadi korelasi antara variable bebas atau antar variable bebas tidak bersifat saling bebas. Dimana besaran yang dapat digunakan untuk mendeteksi adanya multikolinearitas merupakan faktor inflasi ragam atau yang dikenal sebagai variance inflation factor (VIF). Faktor ini digunakan sebagai kriteria untuk mendeteksi multikolinearitas pada regresi linier yang melibatkan lebih dari dua variable bebas[7].

C. *Prediciton Modelling*

Pemodelan prediktif adalah metode untuk memprediksi masa depan dan untuk melakukan pengambilan keputusan dengan cepat di tingkat pelanggan, klien, dan lainnya. Untuk memprediksi masa depan, data disambung menjadi dua bagian yaitu pelatihan atau validasi dan set pengujian. Data pelatihan digunakan untuk pemodelan dan membandingkan juga memilih validasi dan diuji pada set pengujian di masa mendatan[8].

III. Hasil dan Pembahasan

A. *Eksplorasi Data*

Pada penelitian ini digunakan data dengan jumlah 4.849 observasi yang terdiri dari 27 variabel. **Tabel 1**. memperlihatkan informasi tentang variabel yang terdiri dari jenis, panjang, format, format dan label variabel. Beberapa data memiliki jenis yang tepat tetapi beberapa di antaranya tidak. Dalam penelitian ini, data akan diubah bentuknya menjadi numerik kecuali D_id untuk melakukan regresi logistik menggunakan nilai numerik.

Tabel 1. Metadata Dataset

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
23	Age	Num	8	BEST.		Age
26	AreaHomeValue	Num	8	NLMNY15.2		AreaHomeValue
27	AreaMedIncome	Num	8	NLMNY15.2		AreaMedIncome
16	CallCntAll	Num	8	BEST.		CallCntAll
19	CallCntCardAll	Num	8	BEST.		CallCntCardAll
18	CallCntCardP1	Num	8	BEST.		CallCntCardP1
17	CallCntCardP2	Num	8	BEST.		CallCntCardP2
15	CallCntP1	Num	8	BEST.		CallCntP1
14	CallCntP2	Num	8	BEST.		CallCntP2
10	DONAvgAll	Num	8	NLMNY15.2		DONAvgAll
11	DONAvgCardP1	Num	8	NLMNY15.2		DONAvgCardP1
8	DONAvgLast	Num	8	NLMNY15.2		DONAvgLast
9	DONAvgP1	Num	8	NLMNY15.2		DONAvgP1
5	DONCntAll	Num	8	BEST.		DONCntAll
7	DONCntCardAll	Num	8	BEST.		DONCntCardAll
6	DONCntCardP1	Num	8	BEST.		DONCntCardP1
13	DONTimeFirst	Num	8	BEST.		DONTimeFirst
12	DONTimeLast	Num	8	BEST.		DONTimeLast
2	D_ID	Char	8	$8.	$8.	D_ID
22	DemArea	Char	2	$2.	$2.	DemArea
25	DemHomeOwner	Char	1	$1.	$1.	DemHomeOwner
4	DonCntP1	Num	8	BEST.		DonCntP1
1	Donor	Num	8	BEST.		Donor
3	Donor_D	Num	8	NLMNY15.2		Donor_D
20	Donor_Status	Char	1	$1.	$1.	Donor_Status
21	Donor_Status_Prev_Camp	Num	8	BEST.		Donor_Status_Prev_Camp
24	Gender	Char	1	$1.	$1.	Gender

Pada dataset yang dimiliki dilakukan eksplorasi data, yang menunjukkan terdapat beberapa variable yang memiliki nilai yang hilang (*missing values*).

Tabel 2. *Missing Values* pada Dataset

The MEANS Procedure

Variable	Label	N Miss
Donor	Donor	0
Donor_D	Donor_D	2498
DonCntP1	DonCntP1	0
DONCntAll	DONCntAll	0
DONCntCardP1	DONCntCardP1	0
DONCntCardAll	DONCntCardAll	0
DONAvgLast	DONAvgLast	0
DONAvgP1	DONAvgP1	0
DONAvgAll	DONAvgAll	0
DONAvgCardP1	DONAvgCardP1	893
DONTimeLast	DONTimeLast	0
DONTimeFirst	DONTimeFirst	0
CallCntP2	CallCntP2	0
CallCntP1	CallCntP1	0
CallCntAll	CallCntAll	0
CallCntCardP2	CallCntCardP2	0
CallCntCardP1	CallCntCardP1	0
CallCntCardAll	CallCntCardAll	0
Age	Age	1128
AreaHomeValue	AreaHomeValue	0
AreaMedIncome	AreaMedIncome	0
DemArea		0

Dari pengamatan di atas, terdeteksi terdapat tiga variabel yang mengandung missing value yaitu Donor_D yang memiliki 2498 baris data yang hilang, 893 untuk DONAvgCardP1 dan 1128 untuk Age. Variabel yang mengandung nilai hilang tinggi adalah Donor_D. Ini hampir 52% dari total pengamatan. Pada gambar.., variable yang ditampilkan yang memiliki data yang hilang ialah variable yang memiliki nilai numerik. Untuk menginput data, jumlah nilai yang hilang pada variabel Donor_D adalah 2345, 887 untuk variable DONAvgCardP1, dan 1279 untuk variable Age.

B. Imputasi Data

Pada proses imputasi, data yang diimputasi menggunakan nilai rata-rata (*mean*). Nilai rata-rata akan diambil dan diisi ke dalam nilai yang hilang. Untuk Donor_D, nilai meannya ialah 14.86, DONAvgP1 adalah 14.36, DONAvgAll adalah 11.95, DONAvgCardP1 adalah 13.59 dan Age ialah menjadi 58.797. Di sisi lain, cara lain untuk menghitung data tidak hanya menggunakan metode imputasi rata-rata, tetapi juga imputasi tunggal dan metode imputasi ganda. Untuk langkah selanjutnya, ia akan mencoba mengisi nilai yang hilang dengan metode single dan multiple imputasi.

C. Multicollinearity Test

Pada penelitian ini akan dilakukan pengecekan terhadap korelasi antara variable bebas atau antar variable bebas tidak bersifat saling bebas. Dimana besaran yang dapat digunakan untuk mendeteksi adanya multikolinearitas merupakan faktor inflasi ragam atau yang dikenal sebagai variance inflation factor (VIF). Ketika nilai VIF > 2 maka akan dilakukan penghapusan terhadap variable yang memiliki korelasi yang tinggi.

Tabel 3. Nilai Estimasi Parameter pada Dataset

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	0.52361	0.08977	5.83	<.0001	0
Donor_D	Donor_D	1	0.00515	0.00094354	5.46	<.0001	1.28965
DonCntP1	DonCntP1	1	0.01021	0.00789	1.29	0.1958	5.94283
DONCntAll	DONCntAll	1	-0.00296	0.00240	-1.23	0.2169	9.96318
DONCntCardP1	DONCntCardP1	1	0.01957	0.01124	1.74	0.0817	6.70227
DONCntCardAll	DONCntCardAll	1	0.00008971	0.00520	0.02	0.9862	12.81493
DONAvgLast	DONAvgLast	1	-0.00615	0.00114	-5.38	<.0001	2.49041
DONAvgP1	DONAvgP1	1	0.00123	0.00218	0.56	0.5724	9.14752
DONAvgAll	DONAvgAll	1	0.00020811	0.00228	0.09	0.9274	6.39385
DONAvgCardP1	DONAvgCardP1	1	-0.00187	0.00166	-1.13	0.2593	4.09692
DONTimeLast	DONTimeLast	1	-0.00933	0.00224	-4.17	<.0001	1.70409
DONTimeFirst	DONTimeFirst	1	0.00160	0.00064268	2.50	0.0126	12.08376
CallCntP2	CallCntP2	1	-0.00854	0.00447	-1.91	0.0560	9.30194
CallCntP1	CallCntP1	1	0.00130	0.00315	0.41	0.6790	11.90799
CallCntAll	CallCntAll	1	0.00352	0.00188	1.87	0.0619	37.54330
CallCntCardP2	CallCntCardP2	1	0.01182	0.01058	1.12	0.2640	4.18473
CallCntCardP1	CallCntCardP1	1	0.00211	0.00304	0.69	0.4873	3.91690
CallCntCardAll	CallCntCardAll	1	-0.01306	0.00494	-2.64	0.0082	36.18186
Donor_Status		1	-0.00078572	0.00865	-0.09	0.9277	1.89617
Donor_Status_Prev_Camp	Donor_Status_Prev_Camp	1	0.04306	0.01957	2.20	0.0279	1.91013
DemArea		1	-0.00031562	0.00053477	-0.59	0.5551	1.18374
Age	Age	1	0.00067642	0.00049811	1.36	0.1745	1.05839
Gender		1	-0.00579	0.01204	-0.48	0.6304	1.01917
DemHomeOwner		1	0.02666	0.01646	1.62	0.1054	1.35805
AreaHomeValue	AreaHomeValue	1	2.913491E-7	1.572668E-7	1.85	0.0640	1.34891
AreaMedIncome	AreaMedIncome	1	-6.02981E-7	3.370863E-7	-1.79	0.0737	1.56757

Untuk mengetahui variabel berkorelasi tinggi, langkah pertama adalah melihat nilai VIF dari mana ditunjukkan pada tabel. Berdasarkan data, variabel CallCntAll memiliki nilai VIF tertinggi dari variabel lainnya. Langkah selanjutnya yaitu melihat tabel diagnostik kolinearitas untuk melihat nilai tinggi lainnya secara berurutan.

Tabel 4. Nilai Variance dari Dataset

Collinearity Diagnostics						
Proportion of Variation						
imeFirst	CallCntP2	CallCntP1	CallCntAll	CallCntCardP2	CallCntCardP1	CallCntCardAll
0003569	0.00002786	0.00001232	0.00001016	0.00003234	0.00007080	0.00000979
0022712	3.110759E-9	0.00000213	0.00005185	1.936065E-7	0.00012499	0.00005051
0080439	0.00000283	8.058136E-7	0.00017806	0.00004152	0.00008404	0.00013347
0.00290	0.00000355	0.00000646	0.00047533	0.00008204	0.00077661	0.00039102
0.00103	0.00000187	3.348839E-7	0.00001144	0.00001053	0.00179	0.00002870
0012947	0.00065719	0.00053501	0.00001622	0.00136	0.00402	0.00002128
0002276	0.00144	0.00058512	0.00011287	0.00134	0.00150	0.00005888
0038314	0.00214	0.00010293	0.00037545	0.00025938	0.00001678	0.00009257
0037076	0.00382	0.00034521	0.00008140	0.00121	0.00053905	0.00000227
0.00160	0.01535	0.00093997	0.00038267	0.00281	0.00007044	0.00000127
0.00182	0.00317	0.00013862	0.00003149	0.00067469	0.00005025	0.00002533
0006522	0.00004402	0.00004017	0.00000337	0.00016108	0.00005652	0.00001697
0.00174	0.00064405	0.00004591	0.00000632	0.00130	0.01008	0.00044922
0.00726	0.00566	0.00000122	0.00000158	0.00000592	0.00550	0.00058149
0032949	0.00040556	0.00006625	0.00006418	0.00001597	0.04154	0.00011792
0.00725	0.00075807	0.00000537	0.00115	0.00283	0.02264	0.00405
0.00437	0.01241	0.00028921	0.00004268	0.00166	0.00562	0.00229
0.07282	0.00000309	0.00039044	0.00162	0.00542	0.06589	0.00284
0.00246	0.04404	0.00181	0.00300	0.00752	0.29373	0.00073719
0.01901	0.00004087	0.00390	0.00289	0.02857	0.01067	0.00019396
0.00625	0.00027426	0.00228	0.00975	0.16740	0.12507	0.00081070
0.04888	0.02462	0.02885	0.00679	0.34904	0.08722	0.00137
0.41039	0.06883	0.00163	0.06675	0.00647	0.12930	0.08128
0.05919	0.26071	0.20303	0.00053347	0.02198	0.08116	0.00924
0.27623	0.04573	0.71038	0.23174	0.01125	0.11118	0.00307
0.07444	0.50923	0.04460	0.67393	0.38855	0.00132	0.89213

Di coloumn CallCntCall, terdapat nilai tertinggi yang berada di garis bawah. Nilainya dapat dilihat pada tabel diagnostik kolinearitas. Nilai tertinggi ada di CallCntAll dengan kolinearitas adalah 0,67393. Setelah memperoleh nilai tertinggi, kita merujuk pada coloumn lain dalam baris yang sama yang juga memiliki hasil tertinggi. Dari hasilnya, CallCntCardAll lebih tinggi dari CallCntAll. Kolinearitas CallCntCardAll sebesar 0.89213 yaitu hampir 90%. Kami mengambil 2 nilai yang tinggi dan menghapus salah satunya. Data yang akan dijatuhkan adalah yang memiliki chi-square rendah.

Tabel 5. Nilai Chi-Square

Analysis of Effects Eligible for Entry			
Effect	DF	Score Chi-Square	Pr > ChiSq
Donor_D	1	0.0000	1.0000
DonCntP1	1	94.4879	<.0001
DONCntAll	1	51.7054	<.0001
DONCntCardP1	1	83.3925	<.0001
DONCntCardAll	1	58.1784	<.0001
DONAvgLast	1	91.1013	<.0001
DONAvgP1	1	49.2144	<.0001
DONAvgAll	1	57.1820	<.0001
DONAvgCardP1	1	54.2319	<.0001
DONTimeLast	1	42.7723	<.0001
DONTimeFirst	1	28.3144	<.0001
CallCntP2	1	9.4585	0.0021
CallCntP1	1	22.7242	<.0001
CallCntAll	1	30.9555	<.0001
CallCntCardP2	1	11.1105	0.0009
CallCntCardP1	1	41.1470	<.0001
CallCntCardAll	1	28.9216	<.0001
Donor_Status	1	3.1831	0.0744
Donor_Status_Prev_Ca	1	68.3170	<.0001
DemArea	1	0.3803	0.5374
Age	1	7.7500	0.0054
Gender	1	0.0065	0.9359
DemHomeOwner	1	0.3266	0.5677
AreaHomeValue	1	0.6670	0.4141
AreaMedIncome	1	1.4893	0.2223

Karena nilai chi-square CallCntCardAll kurang dari dari CallCntAll, maka diputuskan untuk menghilangkan variabel CallCntCardAll. Dimana langkah ini akan diulang sampai semua VIF < 2. Setelah melakukan rotasi, didapatkan beberapa tersisa yang dapat dilihat pada tabel di bawah ini:

Tabel 6. Hasil Pengurangan Variabel yang memiliki multikolinearitas

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.50804	0.06210	8.18	<.0001	.	0
Donor_D	Donor_D	1	0.00507	0.00093159	5.44	<.0001	0.79691	1.25484
DonCntP1	DonCntP1	1	0.01678	0.00408	4.11	<.0001	0.63011	1.58704
DONCntCardAll	DONCntCardAll	1	0.00132	0.00204	0.65	0.5168	0.50677	1.97328
DONAvgLast	DONAvgLast	1	-0.00644	0.00085040	-7.57	<.0001	0.72521	1.37891
DONTimeLast	DONTimeLast	1	-0.00704	0.00179	-3.93	<.0001	0.91395	1.09415
Donor_Status		1	-0.00693	0.00642	-1.08	0.2800	0.96060	1.04102
Donor_Status_Prev_Camp	Donor_Status_Prev_Camp	1	0.04608	0.01897	2.43	0.0152	0.55858	1.79026
DemArea		1	-0.00025795	0.00053365	-0.48	0.6289	0.84993	1.17657
Age	Age	1	0.00074715	0.00049433	1.51	0.1307	0.96114	1.04044
Gender		1	-0.00236	0.01200	-0.20	0.8442	0.98956	1.01055
DemHomeOwner		1	0.02874	0.01645	1.75	0.0807	0.73873	1.35367
AreaHomeValue	AreaHomeValue	1	3.101209E-7	1.571545E-7	1.97	0.0485	0.74377	1.34449
AreaMedIncome	AreaMedIncome	1	-6.08307E-7	3.360451E-7	-1.81	0.0703	0.64308	1.55502

Berdasarkan hasil pada tabel 6, maka ditemukan inflasi varians dari semua variabel <2. Diasumsikan bahwa tidak ada variabel yang memiliki multikolinearitas. Ketika data bersih dari multikolinearitas, data siap digunakan untuk pemodelan. Didapatkan variable yang tersisa sebanyak 12 dari 25 variabel sebelumnya.

D. Logistic Regression

Setelah multikolinearitas, model dapat dibangun untuk memprediksi donator akan donor atau tidak dari variabel lain. Sebelum dilakukan pemodelan, data perlu dibagi menjadi pelatihan dan set tes. Set pelatihan akan digunakan untuk membangun model dan set tes akan digunakan untuk prediksi.

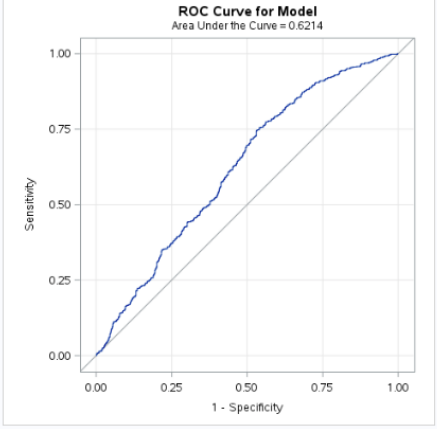
Tabel 7. Nilai C-Statistics data *mean imputation*

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	62.4	Somers' D	0.249
Percent Discordant	37.6	Gamma	0.249
Percent Tied	0.0	Tau-a	0.124
Pairs	2876784	c	0.624

Berdasarkan tabel, statistik C dapat memprediksi apakah modelnya bagus atau tidak. Ketika statistik C lebih besar dari 0,5 maka model dapat diterima. Hasilnya, C adalah 0,624, itu berarti model data imputasi menggunakan rata-rata dapat diterima.

Saat memprediksi donor, data pelatihan yang telah selesai dengan imputasi rata-rata digunakan dalam set tes. Data dibagi menjadi 70% dan 30% untuk data pelatihan dan data tes.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	62.1	Somers' D	0.243
Percent Discordant	37.9	Gamma	0.243
Percent Tied	0.0	Tau-a	0.121
Pairs	528914	c	0.621



Pada hasil pertama, model menggunakan pelatihan yang ditetapkan dalam 70% dan 30% dari tes. ROC menunjukkan 62,1% data diprediksi benar..

IV. Kesimpulan

Berdasarkan hasil pada bab sebelumnya, kesimpulan yang dapat dirangkum ialah setiap data yang hilang dapat diisi dengan menggunakan nilai rata-rata (*mean*). Untuk imputasi rata-rata, nilai yang hilang diisi dengan dengan nilai rata-rata di masing-masing variabel. Dalam perlakuan multikolinearitas menggunakan dua data yang diperhitungkan yaitu data yang telah dilakukan *mean imputation*. Kedua data tersebut memiliki nilai multikolinearitas yang berbeda. Menggunakan data *mean imputation*, 12 area variabel yang dihilangkan yaitu CallCntCardAll, CallCntP1, CallCntP2, DonCntAll, CallCntCardP1, DonAvgP1, DonTimeFirst, CallCntAll, CallCntCardP2, DonCntCardP1, DonCntCardP1, DonAvgCardP1 dan DonAvgAll.

Setelah data bersih dari multikolinearitas, diterapkan model logistic regression. Data *mean imputation* memperoleh C-statistik tinggi di 0,621. Akurasi model memiliki 62,1% dari model prediksi. Ini memiliki persentase yang sama ketika menggunakan imputasi rata-rata dalam set pelatihan.

Daftar Pustaka

[1] Katil, "Predict Donations Using Donors' Past Behaviour," 2018. https://www.kaggle.com/code/gauravsalaskar/predict-donations-using-donors-past-behaviour/data (accessed May 12, 2022).

[2] I. Eldiyana, E. Nurlaelah, and N. Herrhyanto, "Estimasi Missing Data Dengan Metode Multivariate Imputation By Chained Equations (Mice) Untuk Membentuk Persamaan Regresi Linear Berganda," *J. EurekaMatika*, vol. 8, no. 1, pp. 97–107, 2020.

[3] Mukarromah, S. Martha, and Ilhamsyah, "Perbandingan Imputasi Missing Data Menggunakan Metode Mean Dan Metode Algoritma K-Means," *Bul. Ilm. Mat. Stat. dan Ter.*, vol. 04, no. 3, pp. 305–312, 2015.

[4] W. Hanieza, H. M. Sarkan, N. N. A. Sjarif, and Y. Yahya, "A Prediction Model for Blood Donation Using Multiple Logistic Regression," *Open Int. J. Informatics*, vol. 7, no. 2, pp. 147–157, 2019.

[5] N. P. A. Widiari, I. M. A. D. Suarjaya, and D. P. Githa, "Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 8, no. 2, p. 137, 2020, doi: 10.24843/jim.2020.v08.i02.p07.

[6] D. N. Schreiber-Gregory, "Multicollinearity: What Is It, Why Should We Care, and How Can It Be Controlled?," *SAS Inst. INC*, pp. 1404–2017, 2017, [Online].

[7] M. Sriningsih, D. Hatidja, and J. D. Prang, "Penanganan Multikolinearitas Dengan Menggunakan

Analisis Regresi Komponen Utama Pada Kasus Impor Beras Di Provinsi Sulut," *J. Ilm. Sains*, vol. 18, no. 1, p. 18, 2018, doi: 10.35799/jis.18.1.2018.19396.

[8] SAS, "Predictive Modeling using SAS Purpose of Predictive Modeling," 2006.