

KATA PENGANTAR

Puji syukur penulis naikkan kepada Tuhan Yesus Kristus karena hanya oleh kasih dan karunia-Nya, Skripsi ini dapat diselesaikan dengan baik tepat pada waktunya.

Skripsi dengan judul “Implementasi *Machine Learning* dengan Algoritma *Logistic Regression* dan *Random Forest* untuk Prediksi Performa Calon Mahasiswa Baru” ini ditujukan untuk memenuhi sebagian persyaratan akademik guna memperoleh gelar Sarjana Teknik Strata Satu di Universitas Pelita Harapan, Tangerang.

Penulis menyadari bahwa tanpa adanya dukungan dari berbagai pihak, Skripsi ini tidak akan dapat diselesaikan dengan baik seperti sekarang ini. Oleh karena itu, penulis bermaksud untuk mengucapkan terima kasih kepada pihak-pihak yang mendukung proses penyelesaian Skripsi ini, yaitu kepada:

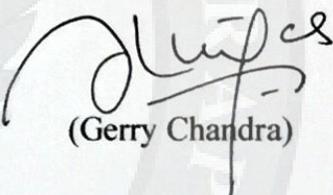
1. Bapak Eric Jobiliong, Ph.D., selaku Dekan Fakultas Sains dan Teknologi Universitas Pelita Harapan.
2. Bapak Dr. Henri P. Uranus, selaku Ketua Program Studi Teknik Elektro Universitas Pelita Harapan.
3. Bapak Dr.-Ing. Ihan Martoyo, MTS selaku pembimbing yang telah banyak meluangkan waktu dan tenaga untuk memberikan bimbingan, masukan, dan motivasi dalam pengerjaan Skripsi ini.
4. Seluruh dosen dan staf Universitas Pelita Harapan, khususnya Program Studi Teknik Elektro, yang telah memberikan ilmu-ilmu dan bantuan kepada penulis sebagai bekal dalam pengerjaan Skripsi ini.
5. Brian Lee, M. Amil Busthon dan Sari Rahmawati yang telah banyak meluangkan waktu untuk berdiskusi, memberikan ilmu-ilmu dan solusi yang membantu penulis dalam Skripsi ini.
6. Orang tua penulis yang selalu memberikan dukungan moril maupun materiil kepada penulis.
7. Seluruh rekan mahasiswa Program Studi Teknik Elektro angkatan 2016 yang telah membantu penulis dalam proses pengerjaan Skripsi ini serta

menjadi teman seperjuangan di Laboratorium Penelitian Teknik Elektro Universitas Pelita Harapan.

8. Seluruh rekan mahasiswa aktif Program Studi Teknik Elektro Universitas Pelita Harapan yang terus memberikan dukungan dan motivasi kepada penulis.
9. Pihak-pihak lain yang tidak dapat disebutkan satu per satu.

Akhir kata, penulis menyadari bahwa masih terdapat banyak kekurangan dalam Skripsi ini. Oleh karena itu, kritik dan saran dari pembaca akan sangat bermanfaat bagi penulis untuk dapat berkembang lebih lanjut. Semoga Skripsi ini dapat bermanfaat bagi setiap pihak yang membacanya, Tuhan memberkati.

Tangerang, 20 Januari 2020



(Gerry Chandra)

DAFTAR ISI

	halaman
HALAMAN JUDUL	
PERNYATAAN TENTANG TUGAS AKHIR DAN PENYERAHAN HAK NONEKSKLUSIF TANPA ROYALTI.....	
PERSETUJUAN DOSEN PEMBIMBING SKRIPSI	
PERSETUJUAN TIM PENGUJI SKRIPSI.....	
ABSTRAK	v
<i>ABSTRACT</i>	vi
KATA PENGANTAR	vii
DAFTAR ISI.....	ix
DAFTAR GAMBAR	xi
DAFTAR TABEL.....	xiii
DAFTAR LAMPIRAN.....	xiv
BAB I PENDAHULUAN	
1.1 Latar Belakang	1
1.2 Maksud dan Tujuan	3
1.3 Batasan Masalah.....	3
1.4 Metode Penelitian.....	3
1.5 Sistematika Penulisan.....	4
BAB II LANDASAN TEORETIS	
2.1 <i>Python</i>	5
2.2 <i>Jupyter Notebook</i>	5
2.3 <i>Anaconda</i>	6
2.4 <i>Scikit-learn</i>	6
2.5 <i>Data Science</i>	7
2.6 <i>Educational Data Mining</i>	7
2.7 <i>Predictive Analytics</i>	7
2.8 <i>Machine Learning</i>	8
2.9 <i>Supervised Learning</i>	8
2.9 Klasifikasi.....	8
2.10 <i>Logistic Regression</i>	9
2.11 <i>Decision Tree</i>	10
2.11 <i>Random Forest</i>	11
2.12 <i>Model Evaluation Metric</i>	12
2.12.1 <i>Confusion Matrix</i>	13
2.12.2 <i>Accuracy</i>	14
2.12.3 <i>Recall</i>	14
2.12.4 <i>Precision</i>	14
2.12.5 <i>F1-score</i>	15
2.12.6 Kurva AUC-ROC.....	15

BAB III METODE PENELITIAN	
3.1	Deskripsi Sistem.....17
3.1.1	<i>Datasets</i>17
3.1.2	Perangkat.....20
3.2	Alir Kerja.....21
3.2.1	Akuisisi Data.....22
3.2.2	Pembersihan Data.....22
3.2.4	<i>Exploratory Data Analysis</i>31
3.2.5	Pengembangan Model.....31
3.2.6	Evaluasi Model.....37
BAB IV EXPLORATORY DATA ANALYSIS	
4.1	Informasi Sekolah.....38
4.1.1	Nilai SMA.....38
4.1.2	Kurikulum.....40
4.1.3	Letak Geografis Sekolah.....41
4.2	Informasi Universitas.....43
4.2.1	Jumlah Aplikasi.....43
4.2.2	Program Studi.....44
4.2.3	Fakultas.....46
4.3	Informasi Profesi Orang Tua.....46
4.3.1	Profesi Ayah.....47
4.3.2	Profesi Ibu.....50
BAB V HASIL DAN ANALISIS PERFORMA MODEL	
5.1	Hasil Evaluasi Performa Model.....53
5.1.1	Kurva <i>Probability vs. Various Metrics</i>53
5.1.2	<i>Confusion Matrix</i>55
5.1.3	Kurva AUC-ROC.....57
5.2	Analisis Performa Model.....58
5.3	Eksplorasi Model.....60
5.3.1	Model <i>Balanced</i>60
5.3.2	Model Sederhana.....62
5.4	Analisis <i>Feature Importances</i>63
BAB VI PENUTUP	
6.1	Kesimpulan.....68
6.2	Saran.....69
DAFTAR PUSTAKA	
DAFTAR SINGKATAN	
LAMPIRAN	

DAFTAR GAMBAR

	halaman
Gambar 2.1	Ilustrasi struktur <i>Decision Tree</i> 10
Gambar 2.2	Struktur <i>Confusion Matrix</i> 13
Gambar 2.3	Kurva AUC-ROC 16
Gambar 3.1	Diagram alir penelitian 21
Gambar 3.2	Diagram proses pembersihan data 22
Gambar 3.3	<i>Countplot</i> kolom <i>school_prop</i> sebelum pembersihan 23
Gambar 3.4	<i>Countplot</i> kolom <i>school_prop</i> setelah pembersihan 24
Gambar 3.5	Format tabel <i>dataset</i> OPCS sebelum normalisasi semester 25
Gambar 3.6	Format tabel <i>dataset</i> OPCS setelah normalisasi semester 26
Gambar 3.7	Format tabel <i>dataset</i> OPCS setelah pivot 26
Gambar 3.8	Format tabel <i>dataset high school</i> setelah pembersihan 28
Gambar 3.9	Diagram populasi data dalam proses integrasi <i>dataset</i> 29
Gambar 3.10	Jumlah masing-masing label variabel target 31
Gambar 3.11	Diagram proses pengembangan model 32
Gambar 3.12	<i>Mean Categorical Encoder</i> 34
Gambar 3.13	Informasi kolom pada data final 35
Gambar 3.14	<i>Train-test split</i> 35
Gambar 3.15	Ilustrasi <i>SMOTE</i> 36
Gambar 4.1	Distribusi nilai SMA. Sumbu <i>x</i> menunjukkan nilai mata pelajaran yang bersangkutan 39
Gambar 4.2	<i>Scatterplot</i> nilai-nilai SMA dengan variabel target. Sumbu <i>x</i> menunjukkan nilai mata pelajaran yang bersangkutan 40
Gambar 4.3	(a) <i>Countplot</i> kurikulum. (b) Persentase kegagalan berdasarkan kurikulum 41
Gambar 4.4	<i>Countplot</i> provinsi sekolah 42
Gambar 4.5	Persentase kegagalan berdasarkan provinsi sekolah 42
Gambar 4.6	(a) <i>Countplot</i> unit geografis sekolah. (b) Persentase kegagalan berdasarkan unit geografis sekolah 43
Gambar 4.7	(a) <i>Countplot</i> jumlah aplikasi. (b) Persentase kegagalan berdasarkan jumlah aplikasi 44
Gambar 4.8	<i>Countplot</i> populasi program studi 44
Gambar 4.9	Persentase kegagalan berdasarkan program studi 45
Gambar 4.10	(a) <i>Countplot</i> status pindah program studi. (b) Persentase kegagalan berdasarkan status pindah program studi 45
Gambar 4.11	(a) <i>Countplot</i> fakultas. (b) Persentase kegagalan berdasarkan fakultas 46
Gambar 4.12	(a) <i>Countplot</i> status hubungan program studi dengan profesi orang tua. (b) Persentase kegagalan berdasarkan status hubungan program studi dengan profesi orang tua 47
Gambar 4.13	<i>Countplot</i> program studi yang berhubungan dengan profesi orang tua 47
Gambar 4.14	<i>Countplot</i> profesi ayah 48

Gambar 4.15	<i>Countplot</i> program studi berdasarkan profesi ayah sebagai wiraswasta	48
Gambar 4.16	<i>Countplot</i> program studi berdasarkan profesi ayah sebagai kontraktor	49
Gambar 4.17	<i>Countplot</i> program studi berdasarkan profesi ayah sebagai dokter	49
Gambar 4.18	<i>Countplot</i> profesi ibu	50
Gambar 4.19	<i>Countplot</i> program studi berdasarkan profesi ibu sebagai wiraswasta	51
Gambar 4.20	<i>Countplot</i> program studi berdasarkan profesi ibu sebagai dokter	51
Gambar 4.21	<i>Countplot</i> program studi berdasarkan profesi ibu sebagai notaris	52
Gambar 4.22	<i>Countplot</i> program studi berdasarkan profesi ibu sebagai pengacara	52
Gambar 5.1	Kurva <i>probability vs. various metrics (Logistic Regression)</i>	54
Gambar 5.2	Kurva <i>probability vs. various metrics (Random Forest)</i>	54
Gambar 5.3	<i>Confusion matrix (Logistic Regression)</i>	55
Gambar 5.4	<i>Confusion matrix (Random Forest)</i>	56
Gambar 5.5	Kurva AUC-ROC (<i>Logistic Regression</i>)	57
Gambar 5.6	Kurva AUC-ROC (<i>Random Forest</i>)	58
Gambar 5.7	Kurva <i>probability vs. various metrics model balanced (Logistic Regression)</i>	62
Gambar 5.8	<i>Feature importances</i> model utama	64
Gambar 5.9	<i>Feature importances</i> model utama, tanpa variabel <i>faculty</i> dan <i>major_name</i>	65
Gambar 5.10	<i>Feature importances</i> model <i>balanced</i>	66
Gambar 5.11	<i>Feature importances</i> model <i>balanced</i> , tanpa variabel <i>faculty</i> , <i>major_name</i> dan <i>hs_final</i>	67

DAFTAR TABEL

	halaman
Tabel 3.1 Detail kolom <i>dataset online admission</i>	17
Tabel 3.2 Detail kolom <i>dataset OPCS</i>	19
Tabel 3.3 Detail kolom <i>dataset high school</i>	20
Tabel 3.4 Versi perangkat	21
Tabel 5.1 <i>Classification report</i>	57
Tabel 5.2 <i>Classification report model balanced</i>	61
Tabel 5.3 <i>Classification report model sederhana</i>	63



DAFTAR LAMPIRAN

	halaman
Lampiran A	
<i>Environment Setup</i>	A-1
Lampiran B	
<i>OA Dataset Cleansing Notebook</i>	B-1
<i>OPCS Dataset Cleansing Notebook</i>	B-8
<i>HS Dataset Cleansing Notebook</i>	B-12
<i>Data Integration and EDA Notebook</i>	B-14
<i>Model Development: Feature Engineering Notebook</i>	B-23
<i>Model Development: Train Notebook</i>	B-28
<i>Model Development: Evaluation Notebook</i>	B-33
Lampiran C	
<i>Python script: evaluate.py</i>	C-1
<i>Python script: utils.py</i>	C-2
<i>Python script: visualize.py</i>	C-4
Lampiran D	
<i>Form Similarity Check Clearance</i>	D-1
<i>Originality Report: BAB I</i>	D-2
<i>Originality Report: BAB II</i>	D-3
<i>Originality Report: BAB III</i>	D-4
<i>Originality Report: BAB IV</i>	D-5
<i>Originality Report: BAB V</i>	D-6
<i>Originality Report: BAB VI</i>	D-7
<i>Originality Report: FULL</i>	D-8
Lampiran E	
<i>Paper Seminar Nasional Sains, Rekayasa dan Teknologi 2019</i>	E-1