MENINGKATKAN KINERJA *K-NEAREST NEIGHBOR* MENGGUNAKAN *PERMUTATION FEATURE IMPORTANCE* UNTUK PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA UNIVERSITAS PENDIDIKAN GANESHA

TESIS

Oleh

GD. ADITYA JANA SATVIKA

NIM 2029101022

PROGRAM STUDI ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS PENDIDIKAN GANESHA
2024

UNIVERSITAS PENDIDIKAN GANESHA
UNDIKSHA

MENINGKATKAN KINERJA *K-NEAREST NEIGHBOR* MENGGUNAKAN *PERMUTATION FEATURE IMPORTANCE* UNTUK PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA UNIVERSITAS PENDIDIKAN GANESHA

TESIS

Diajukan kepada

Universitas Pendidikan Ganesha

Untuk Memenuhi Sebagian Persyaratan

Memperoleh Gelar Magister Komputer

Program Studi Ilmu Komputer

Oleh

GD. ADITYA JANA SATVIKA

NIM 2029101022

PROGRAM STUDI ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS PENDIDIKAN GANESHA
2024

Tesis oleh Gd. Aditya Jana Satvika ini telah diperiksa dan disetujui untuk Mengikuti Ujian Tesis.

Singaraja, 24 November 2023

Pembimbing I

DR. I NYOMAN SUKAJAYA, M.T.

.....

NIP 196711151993031001

Pembimbing II

DR. I GEDE ARIS GUNADI, S.SI., M.KOM.

.....

NIP 197703182008121004

Tesis oleh Gd. Aditya Jana Satvika ini telah dipertahankan di depan tim penguji dan dinyatakan diterima sebagai salah satu persyaratan untuk memperoleh gelar Magister Komputer di Program Studi Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha.

Disetujui pada tanggal: 30 Januari 2024

oleh

Tim Penguji

....., Ketua (Dr. I Made Gede Sunarya, S.Kom., M.Cs.)
NIP 198307252008011008

....., Anggota (Dr. I Made Agus Wirawan, S.Kom., M.Cs.)
NIP 198408272008121001

....., Anggota (Prof. Drs. Sariyasa, M.Sc., Ph.D.)
NIP 196406151989021001

....., Anggota (Dr. I Nyoman Sukajaya, M.T.)
NIP 196711151993031001

....., Anggota (Dr. I Gede Aris Gunadi, S.Si., M.Kom.)
NIP 197703182008121004

Mengetahui Direktur
Program Pascasarjana Undiksha,

Prof. Dr. I Nyoman Jampel, M.Pd.
NIP 195910101986031003

LEMBAR PERNYATAAN

Saya menyatakan dengan sesungguhnya bahwa tesis yang saya susun sebagai syarat untuk memperoleh gelar Magister Komputer dari Program Pascasarjana Universitas Pendidikan Ganesha seluruhnya merupakan hasil karya saya sendiri. Bagian-bagian tertentu dalam penulisan tesis yang saya kutip dari hasil karya orang lain telah dituliskan sumbernya secara jelas dan sesuai dengan norma, kaidah, serta etika akademis.

Apabila di kemudian hari ditemukan seluruh atau sebagian tesis ini bukan hasil karya saya sendiri atau adanya plagiat dalam bagian-bagian tertentu, saya bersedia menerima sanksi pencabutan gelar akademik yang saya sandang dan sanksi-sanksi lainnya sesuai dengan peraturan perundang-undangan yang berlaku di wilayah Negara Kesatuan Republik Indonesia.

Gianyar, 08 Februari 2024

Yang memberi pernyataan,

(Gd. Aditya Jana Satvika)

PRAKATA

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas anugerah-Nya, sehingga tesis yang berjudul: "Meningkatkan Kinerja *K-Nearest Neighbor* Menggunakan *Permutation Feature Importance* Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Universitas Pendidikan Ganesha", dapat diselesaikan sesuai dengan yang direncanakan.

Tesis ini ditulis untuk memenuhi salah satu persyaratan dalam menyelesaikan studi di Program Pascasarjana Universitas Pendidikan Ganesha, Program Studi Ilmu Komputer. Terselesaikannya tesis ini telah banyak memperoleh uluran tangan dari berbagai pihak. Untuk itu, ijinkan penulis menyampaikan terima kasih dan penghargaan kepada pihak-pihak berikut.

1. Dr. I Nyoman Sukajaya, M.T, sebagai pembimbing I yang telah dengan sabar membimbing, mengarahkan, dan memberikan motivasi yang demikian bermakna, sehingga penulis mampu melewati berbagai hambatan dalam perjalanan studi dan penyelesaian tesis ini;
2. Dr. I Gede Aris Gunadi, S.Si., M.Kom, sebagai pembimbing II yang juga senantiasa bersabar memberikan arahan, bimbingan, nasehat kepada penulis, sehingga penulis mampu melewati berbagai hambatan dalam perjalanan studi dan penyelesaian tesis ini;
3. Dr. I Made Agus Wirawan, S.Kom., M.Cs. dan Prof.Drs. Sariyasa, M.Sc., Ph.D. sebagai penguji yang telah banyak memberikan masukan-masukan yang bermanfaat guna menghasilkan tesis yang lebih baik lagi;
4. Wakil Rektor I Universitas Pendidikan Ganesha yang telah memberikan ijin untuk mengadakan observasi guna mendapatkan data/informasi yang dibutuhkan dalam penelitian ini;
5. Kepada Unit Penunjang Akademik Teknologi, Informasi dan Komunikasi (UPA-TIK Undiksha) yang telah memberikan bantuan dalam menyediakan data/informasi mahasiswa program studi pendidikan teknik informatika

Undiksha sesuai dengan kebutuhan penulis, sehingga penelitian ini dapat berlangsung sesuai dengan yang diharapkan;

6. Bapak I Wayan Widiarta dan Ibu Ni Luh Sutris Maryani selaku kedua orang tua tercinta dari penulis yang tiada henti memberikan dukungan moril ataupun material, memberikan dorongan yang positif, dan memberikan doa restu sehingga penulis dapat menyelesaikan studi. Serta terimakasih telah mempercayai penulis untuk menimba ilmu lebih tinggi lagi di Universitas Pendidikan Ganesha, semoga apa yang diharapkan oleh bapak dan ibu kepada penulis dapat dirasakan;
7. I Made Dwitya Sananda Nugraha selaku adik kandung dari penulis yang sangat luar biasa, selalu memberikan dorongan, doa dan semangat yang tiada henti diberikan kepada penulis untuk dapat menyelesaikan pendidikan.

Semoga semua bantuan yang telah diberikan dalam menyelesaikan studi ini, senantiasa dibalas oleh Tuhan Yang Maha Esa.

Penulis menyadari bahwa tesis ini belum sempurna. Namun, kehadirannya dalam konstelasi masyarakat akademis akan menambah perbendaharaan ilmu dalam perkembangan ilmu pengetahuan. Semoga tesis ini dapat memberikan manfaat bagi masyarakat akademis.

Singaraja, 24 November 2023

Penulis

DAFTAR ISI

DAFTAR ISI iii

DAFTAR TABEL vii

DAFTAR GAMBAR x

DAFTAR LAMPIRAN xii

ABSTRAK xiii

ABSTRACT xiv

BAB I PENDAHULUAN 1

1.1 Latar Belakang 1

1.2 Identifikasi Masalah 5

1.3 Batasan Masalah 5

1.4 Rumusan Masalah 6

1.5 Tujuan Penelitian 6

1.6 Manfaat Penelitian 7

1.6.1 Manfaat teoritis (akademis) 7

1.6.2 Manfaat praktis 7

BAB II KAJIAN PUSTAKA 8

2.1 Prestasi Belajar 8

2.2 Faktor Hasil Belajar 9

2.2.1 Faktor jenis kelamin 9
2.2.2 Faktor status sosial dan ekonomi 12
2.2.3 Faktor jumlah mata kuliah mengulang 15
2.2.4 Faktor indeks prestasi mahasiswa 16
2.3 *Machine Learning Approach* 18
2.3.1 *Supervised learning* 18
2.3.2 *Unsupervised learning* 19
2.3.3 *Reinforcement learning* 20
2.4 *Machine Learning Life Cycle* 20
2.5 *Data Mining* 21
2.6 *Data Acquisition* 22
2.7 *Data Pre-processing* 26
2.8 *Data Cleaning* 27
2.8.1. Missing values 27
2.8.2. Outlier data 33
2.9 *Data Integration* 36
2.10 *Data Transformation* 36
2.11 *Feature Encoding* 36
2.12 *Data Normalization* 40
2.13 *Imbalanced Dataset* 43

2.14 Klasifikasi 48
2.15 *K-Nearest Neighbor* (KNN) 49
2.16 *Feature Importance* 52
2.17 *Underfitting & Overfitting* 55
2.18 *Hyperparameter Tuning* 57
2.19 Evaluasi Model 59
2.20 Kajian Hasil Penelitian Relevan 59
2.21 Kerangka Berpikir 60
BAB III METODE PENELITIAN 64
3.1 Data Penelitian 64
3.2 Rancangan Penelitian 67
3.3 *Data Preprocessing* 72
3.4 Split Dataset 74
3.5 *Oversampling* Menggunakan K-means SMOTE 74
3.6 *Hyperparameter Tuning* 75
3.7 Implementasi *K-Nearest Neighbor* (KNN) 76
3.8 Pengujian 85
3.8.1 *Permutation Feature Importance* (PFI) 85
3.8.2 *Confusion matrix* 88
3.9 Skema Pengujian 91

3.10 Keluaran Yang Diharapkan 94

BAB IV HASIL PENELITIAN DAN PEMBAHASAN 95

4.1 Pengumpulan Data 95

4.2 Data Preprocessing 96

4.2.1 Seleksi fitur secara manual 96

4.2.2 Handle missing values 97

4.2.3 Deteksi *outlier* menggunakan *z-score* 99

4.2.4 Feature encoding 106

4.2.5 Data normalization 108

4.3 Implementasi & Evaluasi 110

4.3.1 Implementasi KNN (tanpa teknik *oversampling*) 110

4.3.2 Implementasi KNN + PFI (tanpa teknik *oversampling*) 113

4.3.3 Implementasi KNN (dengan teknik *oversampling*) 125

4.3.4 Implementasi KNN + PFI (dengan teknik *oversampling*) 128

BAB V PENUTUP 132

5.1 Rangkuman 132

5.2 Simpulan 134

5.3 Saran 135

DAFTAR PUSTAKA 138

LAMPIRAN-LAMPIRAN 145

DAFTAR TABEL

Tabel 2.1 Contoh data *outlier* 33

Tabel 2.2 Proses One-Hot Encoding pada tipe data kategori..... 38

Tabel 2.3 Proses Binary Encoding pada data kategori..... 40

Tabel 3.1 Fitur data yang digunakan dalam penelitian 65

Tabel 3.2 Parameter uji pada proses *hyperparameter tuning* 76

Tabel 3.3 Contoh data *training* untuk perhitungan KNN secara manual..... 78

Tabel 3.4 Contoh data *testing* untuk perhitungan KNN secara manual..... 78

Tabel 3.5 Hasil penghitungan *Euclidean distance* secara manual 81

Tabel 3.6 Hasil prediksi KNN menggunakan *Euclidean distance* manual..... 82

Tabel 3.7 Hasil penghitungan *Manhattan distance* secara manual..... 83

Tabel 3.8 Hasil prediksi KNN menggunakan *Manhattan distance* manual..... 84

Tabel 3.9 Ilustrasi hasil prediksi pada data *testing* 86

Tabel 3.10 Ilustrasi pengacakan nilai pada sebuah fitur 88

Tabel 3.11 Tabel *Confusion Matrix* 89

Tabel 4.1 Dataset hasil seleksi fitur manual..... 97

Tabel 4.2 Nilai minimum & maksimum setiap fitur numerikal..... 100

Tabel 4.3 Data *outlier* pada kolom UKT 102

Tabel 4.4 Data pendapatan orang tua mahasiswa 104

Tabel 4.5 Data *outlier* pada kolom Pendapatan Orang tua 104

Tabel 4.6 Data *outlier* pada kolom MK Mengulang (Hingga Smt 4) 105

Tabel 4.7 Transformasi kolom Jk menggunakan *One-hot encoding* 107

Tabel 4.8 Transformasi kolom Keterangan menggunakan *label encoding*..... 108

Tabel 4.9 Normalisasi data menggunakan *MinMax Normalization*..... 109

Tabel 4.10 Detail pembagian *training data* dan *testing data*..... 110

Tabel 4.11 *Confusion Matrix* dari implementasi KNN..... 112

Tabel 4.12 *Performance score* dari implementasi KNN..... 112

Tabel 4.13 Rangking kontribusi fitur terhadap model KNN 115

Tabel 4.14 *Confusion Matrix* dari implementasi KNN + PFI (Tanpa Jk_P, Jk_L, dan IPS 4)..... 116

Tabel 4.15 *Performance score* dari implementasi KNN + PFI (Tanpa Jk_P, Jk_L, dan IPS 4)..... 116

Tabel 4.16 *Confusion Matrix* dari implementasi KNN + PFI (Tanpa fitur Jk_P dan Jk_L) 117

Tabel 4.17 *Performance score* dari implementasi KNN + PFI (Tanpa fitur Jk_P dan Jk_L)..... 117

Tabel 4.18 *Confusion Matrix* dari implementasi KNN + PFI (Tanpa fitur Jk_P dan IPS 4)..... 118

Tabel 4.19 *Performance score* dari implementasi KNN + PFI (Tanpa fitur Jk_P dan IPS 4)..... 119

Tabel 4.20 *Confusion Matrix* dari implementasi KNN + PFI (Tanpa fitur Jk_L dan IPS 4)..... 119

Tabel 4.21 *Performance score* dari implementasi KNN + PFI (Tanpa fitur Jk_L dan IPS 4)..... 120

Tabel 4.22 *Confusion Matrix* dari implementasi KNN + PFI (Tanpa fitur Jk_P)121

Tabel 4.23 *Performance score* dari implementasi KNN + PFI (Tanpa fitur Jk_P) 121

Tabel 4.24 *Confusion Matrix* dari implementasi KNN + PFI (Tanpa fitur Jk_L) 122

Tabel 4.25 *Performance score* dari implementasi KNN + PFI (Tanpa fitur Jk_L) 122

Tabel 4.26 *Confusion Matrix* dari implementasi KNN + PFI (Tanpa fitur IPS 4) 123

Tabel 4.27 *Performance score* dari implementasi KNN + PFI (Tanpa fitur IPS 4) 123

Tabel 4.28 Perbandingan *performance score* KNN & KNN + PFI (Tanpa menggunakan fitur Jk_P, Jk_L, dan IPS 4)..... 124

Tabel 4.29 Hasil optimasi KNN dengan *training data* yang sudah di*oversampling* 127

Tabel 4.30 Hasil implementasi PFI pada KNN dengan *training data* yang sudah di*oversampling*..... 129

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi prediksi menggunakan metode *supervised learning* 19

Gambar 2.2 Persentase tugas seorang *data scientists* 21

Gambar 2.3 Jenis data penelitian atau statistik 24

Gambar 2.4 Gambaran data yang berdistribusi normal 35

Gambar 2.5 Ilustrasi metode *undersampling* dan *oversampling*..... 44

Gambar 2.6 Ilustrasi *Synthetic Minority Oversampling Technique* (SMOTE)..... 45

Gambar 2.7 SMOTE dalam data *noise* 47

Gambar 2.8 Ilustrasi proses klasifikasi 48

Gambar 2.9 Ilustrasi proses klasifikasi menggunakan KNN 52

Gambar 2.10 Ilustrasi proses *K-Fold Cross validation*..... 58

Gambar 2.11 Proses PFI..... 61

Gambar 3.1 Rancangan umum penelitian 68

Gambar 3.2 Flowchart KNN 77

Gambar 3.3 Skema pengujian PFI 93

Gambar 4.1 Status *missing value* setiap fitur 98

Gambar 4.2 Distribusi *z-score* kolom UKT 101

Gambar 4.3 Distribusi *z-score* kolom Pendapatan Orang tua 103

Gambar 4.4 Distribusi *z-score* kolom MK Mengulang (Hingga Smt 4)..... 105

Gambar 4.5 Perubahan *baseline accuracy* dan *baseline recall* berdasarkan rasio antara kelas mayoritas dengan kelas minoritas 125

Gambar 4.6 Perubahan *baseline recall* dan *baseline precision* berdasarkan rasio antara kelas mayoritas dengan kelas minoritas 126

DAFTAR LAMPIRAN

Lampiran 1. Surat Permohonan Izin Observasi Data 145

Lampiran 2. Surat Izin Observasi 146

Lampiran 3. Dataset penelitian 147

Lampiran 4. Hasil uji kombinasi KNN dan PFI 159

Lampiran 5. Kode program PFI 169