IMPLEMENTASI *DATA MINING* UNTUK PREDIKSI *DROP-OUT* DENGAN MENGGUNAKAN *RANDOM FOREST METHOD*

Tesis
untuk memenuhi sebagian persyaratan
mencapai derajat Sarjana S-2 Program Studi
Magister Sistem Informasi

Meylani Utari
30000318410029

SEKOLAH PASCASARJANA
UNIVERSITAS DIPONEGORO
SEMARANG
2020

HALAMAN PENGESAHAN

TESIS

IMPLEMENTASI *DATA MINING* UNTUK PREDIKSI *DROP-OUT* DENGAN MENGGUNAKAN *RANDOM FOREST METHOD*

Oleh:
Meylani Utari
30000318410029

Telah diujikan dan dinyatakan lulus ujian tesis pada tanggal 23 Juni 2020 oleh tim penguji Program Studi Magister Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro.

Semarang, 29 Juni 2020
Mengetahui,

Penguji I

Dr. Suryono, S.Si., M.Si
NIP. 197306301998021001

Penguji II

Farikhin, M.Si., Ph.D
NIP. 197312202000121001

Pembimbing I

Dr. Budi Warsito, S.Si.,M.Si
NIP. 197508241999031003

Pembimbing II

Dr. Retno Kusumaningrum,S.Si.,M.Kom
NIP. 198104202005012001

Mengetahui :
Dekan Sekolah Pascasarjana
Universitas Diponegoro

Dr. R.B. Sularto, S.H., M.Hum
NIP. 196701011991031005

Ketua Program Studi
Magister Sistem Informasi

Dr. Suryono, S.Si., M.Si
NIP. 197306301998021001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam tesis ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar akademik di suatu perguruan tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Semarang, 26 Juni 2020

Meylani Utari

PERNYATAAN PERSETUJUAN PUBLIKASI TESIS UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Diponegoro, saya yang bertanda tangan di bawah ini :

Nama	: Meylani Utari
NIM	: 30000318410029
Program Studi	: Magister Sistem Informasi
Program	: Pascasarjana
Jenis Karya	: Tesis

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Diponegoro Hak Bebas Royalti Noneksklusif atas karya ilmiah saya yang berjudul :

IMPLEMENTASI *DATA MINING* UNTUK PREDIKSI *DROP-OUT* DENGAN MENGGUNAKAN *RANDOM FOREST METHOD*

beserta perangkat yang ada. Dengan Hak Bebas Royalti Noneksklusif ini Program Studi Magister Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tesis saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik hak cipta.

Dibuat di : Semarang
Pada Tanggal : 26 Juni 2020
Yang menyatakan

Meylani Utari
NIM. 30000318410029

KATA PENGANTAR

Puji dan syukur kepada Tuhan yang Maha Esa, atas segala berkat, rahmat, dan karunia yang dicurahkan, sehingga tesis dengan judul Implementasi *Data Mining* untuk Prediksi *Drop-Out* dengan Menggunakan *Random Forest Method* ini dapat diselesaikan. Tesis ini disusun untuk memenuhi salah satu persyaratan memperoleh gelar Magister Komputer (M.Kom) pada Program Studi Magister Sistem Informasi Universitas Diponegoro. Pada kesempatan ini penulis menyampaikan terima kasih yang sebesar – besarnya kepada :

1. Dr. Budi Warsito, S.Si., M.Si selaku Pembimbing I dan Dr. Retno Kusumaningrum, S.Si, M.Kom selaku Pembimbing II yang telah memberikan waktu, ilmu, saran, kritik, semangat, dan nasihat selama penulisan tesis ini.
2. Dr. Suryono, S.Si, M.Si selaku Ketua Program Studi Magister Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro Semarang.
3. Program Beasiswa PasTi Tahun 2018 yang diselengarakan oleh Kemeristekdikti dan Kemendikbud yang telah memberikan kesempatan kepada Penulis sehingga dapat melanjutkan pendidikan Magister di Universitas Diponegoro Semarang dan menyelesaikan penelitian ini.
4. Universitas Sriwijaya dan Fakultas Ilmu Komputer sebagai tempat Penulis bekerja yang telah memberikan dukungan kepada Penulis dengan memberikan izin untuk melanjutkan Pendidikan Magister di Universitas Diponegoro Semarang.
5. Suami, kedua orangtua, sahabat, teman-teman, serta pihak-pihak lain yang tidak dapat disebutkan satu per-satu, yang telah memberikan kontribusi baik moral maupun material sehingga tesis ini dapat terselesaikan dengan baik.

Penulis menyadari bahwa dalam penyusunan tesis ini masih jauh dari sempurna. Oleh karena itu, saran dan kritik yang sifatnya membangun sangat diharapkan. Akhirnya, penulis berharap semoga tulisan ini bermanfaat. Amin.

Semarang, 26 Juni 2020

Penulis,

Meylani Utari

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

ABSTRAK

Drop-out merupakan salah satu isu yang paling umum pada dunia pendidikan, tingkat *drop-out* yang tinggi memberikan dampak yang buruk terhadap perguruan tinggi seperti reputasi dan akreditasi yang kurang baik. Prediksi awal mahasiswa yang berisiko *drop-out* sangat penting dilakukan untuk membantu mengurangi persentase tingkat *drop-out*. Metode klasifikasi statistik telah banyak diimplementasikan untuk mengetahui persentase tingkat *drop-out*, namun hasil akhir pengklasifikasian yang diberikan tidak signifikan serta tingkat ketepatan dugaan yang lebih rendah dibandingkan dengan penggunaan metode seperti pohon *ensemble*. Pada penelitian ini, saya menerapkan teknik *data mining* dengan menggunakan algoritma *random forest* dan *decision tree* yang disertai dengan teknik *imbalance dataset* yaitu *synthetic minority oversampling technique* (SMOTE) untuk melakukan prediksi *drop-out* terhadap mahasiswa sarjana dan diploma pada Fakultas ABC di Perguruan Tinggi XYZ yang mendaftar pada tahun pendaftaran 2008 hingga 2012 dengan konsep *business intelligence*. Dalam penelitian ini, saya membandingkan kinerja dari algoritma *random forest* dan *decision tree* dengan SMOTE dan tanpa SMOTE. Dari hasil percobaan pada penelitian ini, *random forest* tanpa SMOTE menghasilkan nilai akurasi sebesar 92.27%, *random forest* dengan SMOTE menghasilkan nilai akurasi terbaik pada saat nilai k = 2 sebesar 94.72%, *decision tree* tanpa SMOTE menghasilkan nilai akurasi sebesar 91.31% dan *decision tree* dengan SMOTE menghasilkan nilai akurasi terbaik pada saat nilai k = 10. Hasil percobaan menunjukkan bahwa algoritma *random forest* dengan SMOTE yang memiliki nilai akurasi terbaik dibandingkan dengan ketiga algoritma lainnya. Dimasa depan, hasil penelitian ini dapat digunakan sebagai acuan untuk membangun model prediksi dini terhadap mahasiswa yang berpotensi tinggi untuk *drop-out* serta memungkinkan untuk mengetahui variabel-variabel yang menyebabkan mahasiswa tersebut dapat *drop-out*.

Kata Kunci : *drop out*; *random forest*; *synthetic minority over sampling*; SMOTE; data pendidikan; *data mining*; *classification; prediction; imbalance dataset; business intelligence.*

ABSTRACT

Drop-out is one of the most common issues in education, high drop-out rates have a bad impact on universities such as bad reputation and low accreditation. Prediction of students who are at risk of drop-out is very important to be done in order to help reduce the percentage of students drop-out rates. Statistical classification methods have been widely implemented to determine the percentage drop-out rate, but the final results of the classification given are insignificant and the level of accuracy of prediction is lower compared to the ensemble trees method. In this research, the author applies data mining technique using a random forest algorithm and decision tree, accompanied by an imbalance dataset technique, namely synthetic minority oversampling technique (SMOTE) to predict drop-out of undergraduate and diploma students at the ABC Faculty at XYZ University, which registered between 2008 to 2012 along with business intelligence concept. In this research, the author also compared the performance of the random forest algorithm and decision tree with SMOTE and without SMOTE. From the experimental results in this research, random forest without SMOTE produces an accuracy value of 92.27%, random forest with SMOTE produces the best accuracy value when the k = 2 value is 94.72%, decision tree without SMOTE produces an accuracy value of 91.31% and decision tree with SMOTE produces the best accuracy value when the value k = 10. The experimental results show that the random forest algorithm with SMOTE has the best accuracy compared to the other three algorithms. In the future, the results of this research can be used as a reference to build early prediction models of students with high potential for drop-out and make it possible to find out the variables that cause students to drop out.

Keywords: drop out; random forest; synthetic minority over-sampling; SMOTE; educational data; data mining; classification; prediction; imbalance dataset; business intelligence.