ABSTRAK

Caesilia Apri Purwanti, 2024. *Data Mining* Berbantuan Python dengan Regresi Logistik untuk Prediksi Hasil PPDB SMA Negeri 1 Yogyakarta Berdasarkan Seleksi Tahun 2023. Skripsi. Program Studi Pendidikan Matematika, Jurusan Pendidikan Matematika dan Ilmu Pengetahuan Alam, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sanata Dharma.

Penelitian ini bertujuan untuk membentuk model prediksi hasil PPDB di SMA Negeri 1 Yogyakarta dan mengetahui faktor-faktor yang paling mempengarui penerimaan pada setiap jalur pendaftaran berdasarkan seleksi tahun 2023. Penelitian yang dilakukan dengan pendekatan kuantitatif serta jenis penelitian analisis data sekunder ini memanfaatkan Python sebagai alat bantu untuk membuat model prediksi. Dalam penelitian ini data seleksi PPDB tahun 2023 diproses menggunakan teknik *data mining* dengan Regresi Logistik, meliputi: (1) *data cleaning*, (2) *data integration*, (3) *data selection*, (4) *data transformation*, (5) *data mining*, (6) *pattern evaluation*, dan (7) *knowledge presentation*. Pembuatan model prediksi menggunakan Pipeline dan GridSearchCV untuk mencari parameter terbaik dari model dan terdapat 4 model di setiap jalur pendaftaran. Pada Jalur Zonasi, Afirmasi, Prestasi, dan PTO masing-masing memiliki 4 model yang dibedakan berdasarkan *feature selection* yang digunakan, yaitu (1) tanpa *feature selection*, (2) berdasarkan *feature selection* SelectKBest, (3) berdasarkan *feature selection* RFECV, dan (4) berdasarkan *feature selection* Lasso.

Model prediksi yang dihasilkan bergantung pada karakteristik data dan ketentuan yang berlaku pada setiap jalur pendaftaran. Adapun model terbaik dari setiap jalur pendaftaran, meliputi: (1) Model pada Jalur Zonasi berdasarkan *feature selection* SelectKBest dengan *accuracy* 97% dan faktor yang mempengaruhi adalah Nilai Prestasi, Nilai Literasi Numerasi, Nilai Literasi Baca Bahasa Indonesia, Nilai Literasi Baca Bahasa Inggris, Nilai Literasi Sains, Nilai Akreditasi, Rerata Nilai Rapor, serta Zona, (2) Model pada Jalur Afirmasi berdasarkan *feature selection* SelectKBest dengan *accuracy* 91% dan faktor yang mempengaruhi adalah Nilai Prestasi, (3) Model pada Jalur Prestasi berdasarkan *feature selection* SelectKBest dengan *accuracy* 74% dan faktor yang mempengaruhi adalah Nilai Prestasi, Nilai Literasi Numerasi, Nilai Literasi Baca Bahasa Indonesia, Nilai Literasi Baca Bahasa Inggris, Nilai Literasi Sains, serta Nilai Akreditasi, dan (4) Model pada Jalur PTO berdasarkan *feature selection* SelectKBest dengan *accuracy* 100% dan faktor yang mempengaruhi adalah Nilai Prestasi serta Nilai Literasi Numerasi.

Kata kunci: *data mining*, penerimaan peserta didik baru, python, regresi logistik

ABSTRACT

Caesilia Apri Purwanti, 2024. Python-assisted Data Mining with Logistic Regression to Predict PPDB Results of SMA Negeri 1 Yogyakarta Based on Selection in 2023. Thesis. Mathematics Education Study Program, Department of Mathematics and Natural Sciences Education, Faculty of Teacher Training and Education, Sanata Dharma University.

This research aims to form a prediction model of PPDB results at SMA Negeri 1 Yogyakarta and find the factors that most affect acceptance in each registration path based on selection in 2023. This research, which was conducted with a quantitative approach and secondary data analysis research type, utilized Python as a tool to create a prediction model. In this study, PPDB selection data in 2023 is processed using data mining techniques with Logistic Regression, including (1) data cleaning, (2) data integration, (3) data selection, (4) data transformation, (5) data mining, (6) pattern evaluation, and (7) knowledge presentation. Prediction modeling uses Pipeline and GridSearchCV to find the best parameters of the model and there are 4 models in each registration path. The Zoning, Affirmation, Achievement, and PTO pathways each have 4 models that are distinguished based on the feature selection used, namely (1) without feature selection, (2) based on SelectKBest feature selection, (3) based on RFECV feature selection, and (4) based on Lasso feature selection.

The resulting prediction model depends on the characteristics of the data and the condition that apply to each registration path. The best models for each enrollment pathway included: (1) Model on Zoning pathway based on SelectKbest feature selection with 97% accuracy and the influencing factors are Achievement Score, Numeracy Literacy Score, Indonesian Reading Literacy Score, English Reading Literacy Score, Science Literacy Score, Accreditation Score, Average Report Card Score, and Zone, (2) Model on Affirmation pathway based on SelectKbest feature selection with 91% accuracy and the influencing factor is Achievement Score, (3) Model on Achievement pathway based on SelectKbest feature selection with 74% accuracy and the influencing factors are Achievement Score, Numeracy Literacy Score, Indonesian Reading Literacy Score, English Reading Literacy Score, Science Literacy Score, and Accreditation Score, and (4) Model on PTO pathway based on SelectKbest feature selection with 100% accuracy and the influencing factors are Achievement Score and Numeracy Literacy Score.

Keywords: *data mining, logistic regression, new student admissions, python*