Akreditasi Ristekdikti,
No: 30/E/KPT/2019 (Sinta 4)
DOI: https://doi.org/10.31294/p.v23i1.9781

Paradigma,
Vol. 23, No. 1 Maret 2021
P-ISSN 1410-5063, E-ISSN: 2579-3500

Analisis Runtun Waktu Untuk Memprediksi Jumlah Mahasiswa Baru Dengan Model Random Forest

Marchell Rianto [1], Roni Yunis [2]

[1]Program Studi Sistem Informasi, STMIK Mikrsokil Medan
e-mail: 172110954@students.mikroskil.ac.id

[2]Program Studi Sistem Informasi, STMIK Mikrsokil Medan
e-mail: roni@mikroskil.ac.id

Abstrak *- Penerimaan mahasiswa baru merupakan proses penting pada instansi pendidikan seperti perguruan tinggi yang berguna untuk menyaring calon mahasiswa yang diterima sesuai kriteria yang ditentukan oleh perguruan tinggi. Tujuan penelitian ini adalah melakukan prediksi jumlah mahasiswa baru menggunakan model Random Forest dengan dataset penerimaan mahasiswa baru Universitas XYZ periode 2010-2019. Model Random Forest adalah salah satu algoritma machine learning yang sangat baik dalam memecahkan masalah klasifikasi dan regresi. Berdasarkan hasil penelitian didapatkan bahwa model yang dihasilkan memiliki tingkat akurasi 99,8 % dengan nilai MSE dan MAE sebesar 0,02% dalam meramalkan mahasiswa baru. Parameter terbaik dari model dengan nilai maxnodes 100 dan ntree 900 serta adanya trend penurunan jumlah mahasiswa untuk beberapa tahun kedepan.*

Kata kunci: *random forest, jumlah mahasiswa baru, MSE, MAE.*

Abstract *- Admission of new students is an important process in educational institutions such as tertiary institutions which is useful for screening accepted prospective students according to the criteria determined by the college. The purpose of this study is to predict the number of new students using the Random Forest model with the new student admissions dataset of XYZ University. The Random Forest Model is a machine learning algorithm that is excellent at solving classification and regression problems. Based on the research results, it was found that the resulting model has an accuracy rate of 99.8% with MSE and MAE values of 0.02% in predicting new students. The best parameter of the model with a max nodes value of 100 and tree 900 and a decreasing trend in the number of students for the next few years.*

Keywords*: random forest, number of new students, MSE, MAE.*

PENDAHULUAN

Paragdima perguruan tinggi sepatutnya berfokus pada kemerdekaan dan kemandirian pelakon pelajar. Untuk mendapatkan nilai dan tujuan hidup sebagai manusia seutuhnya yang bermanfaat bagi keluarga, masyarakat dan bangsa perlunya meningkatkan kualitas pendidikan (Karmita et al., 2019). Strategi yang akan diambil oleh pihak manajemen suatu perguruan tinggi salah satunya berangkat dari banyaknya mahasiswa yang mendaftar pada perguruan tinggi tersebut. Banyak macam strategi yang perlu dilakukan perguruan tinggi seperti perencanaan keuangan, perencanaan pemasaran, dan salah satu yang paling penting adalah perencanaan akademik.

Perencanaan akademik merupakan bagian penting yang perlu dilakukan untuk merencanakan proses belajar mengajar dalam suatu kampus. Perencanaan membutuhkan data perkiraan (prediksi data) untuk menentukan rencana yang dilakukan atau persiapan apa yang harus dilakukan supaya kita lebih siap menghadapai yang akan terjadi (LTMPT, 2020). Salah satu perencanaan yang harus dilakukan oleh Perguruan Tinggi adalah jumlah penerimaan mahasiswa baru dan menyiapkan sarana dan prasarana.

Peramalan atau *forecasting* adalah memperkirakan atau memprediksi besarnya atau jumlah sesuatu kejadian pada waktu yang akan datang berdasarkan dataset yang terkumpul agar dapat merencanakan strategi di masa yang akan datang (Eka Chandra & Sarinem, 2015). Peramalan mahasiswa baru sangat dibutuhkan bagi perguruan tinggi guna menunjang keputusan bagi perguruan tinggi untuk merubah strategi pemasaran dalam meningkatan mahasiswa baru.

Dalam penelitian ini penerapan model *Random Forest* dipilih karena model tersebut secara khusus menekankan kemudahan aplikasi, biaya komputasi yang rendah, akurasi prediksi yang tinggi dan lebih

Naskah diterima: 2021-03-07, direvisi: 2021-03-19, disetujui: 2021-03-22

fleksibel. *Random Forest* merupakan teknik klasifikasi kelas menggunakan pohon keputusan. Seperti kombinasi prediktor pohon di mana setiap pohon bergantung pada kumpulan data acak sampel secara independen dan dengan distribusi yang sama pada pohon keputusan (Services, 2018).

METODOLOGI PENELITIAN

a. Bahan Penelitian

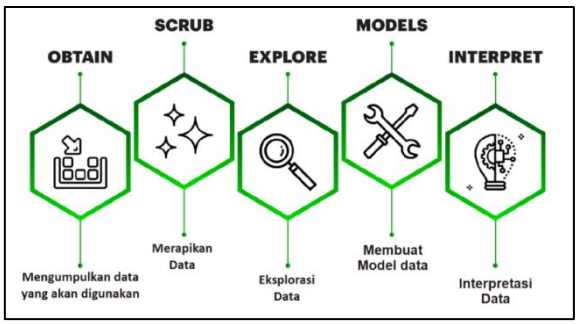
Data yang digunakan dalam penelitian ini adalah data penerimaan mahasiswa baru tahun ajaran 2010/2011 sampai 2019/2020 dari salah satu perguruan tinggi di Medan yaitu Universitas XYZ. Data tersebut memiliki 9613 *record* (baris) dan 11 *column* (atribut).

Tabel 1. Atribut dataset penerimaan mahasiswa baru

Nama Atribut	Contoh Data
Tahun.Akademik	2010/2011
Tanggal.Daftar	24.05.2010
Program.Studi	SISTEM INFORMASI
Jumlah.Mhs	8
Jumlah.Grade.A	2
Jumlah.Grade.B	1
Jumlah.Grade.C	4
Jumlah.Grade.D	2
Kode.Sekolah	10210905
Nama.Sekolah	SMA S METHODIST 7
Nama.Kota	MEDAN

b. Tahap Penelitian (OSEMN)

Dalam melakukan penelitian prediksi mahasiswa baru dengan *Random Forest* ada beberapa langkah-langkah yang dilakukan. Pada penelitian ini menggunakan metode OSEMN. OSEMN adalah plesetan kata yang kedengaran seperti "*Awesome*" atau "mengagumkan" dan merupakan singkatan dari *Obtain, Scrub, Explore, Model, dan iNterpret* (Accenture, 2018).



Gambar 1. Langkah-langkah OSEMN

1. *Obtaining Data (Mengumpulkan data)*

Pada proses pertama yaitu mengidentifikasi dan mengumpulkan data yang tepat dalam penelitian. Jika tidak memiliki data maka tidak ada yang dapat dilakukan. Banyak cara untuk mendapatkan dataset yaitu dari situs *Kaggle.com*, *data.fivethirtyeight.com, BuzzFeedNews* dan masih banyak repositori dataset lainnya (Interviewsqs, 2019). Dalam penelitian ini dataset sudah sajikan dalam bentuk .csv (*Comma Separated Value*) yang diberikan bagian PSI (Pusat Sistem Informasi) Universitas XYZ untuk mendukung program penelitian ini.

2. *Scrubbing Data (Pembersihan Data)*

Scrubbing Data adalah kegiatan melakukan konversi atau merapikan data dari satu format ke format lain dan menggabungkan semuanya ke dalam satu format standar di agar memberikan gambar yang akurat pada hasil akhir (Sharma, 2018). Pada tahap ini akan memeriksa data dan membersihkan data dari data yang tidak diperlukan dengan menggunakan bahasa pemrograman R Software. Proses pembersihan data juga bertujuan mengubah data mentah menjadi data konsisten yang dapat dianalisis. Ini bertujuan agar meningkatkan isi pernyataan statistik berdasarkan data serta keandalannya.

3. *Exploring Data*

Exploring Data adalah kegiatan melihat pola data, mengelompokkan dan melakukan beberapa jenis visualisasi dan pengujian statistik data (Brownlee, 2020). Clustering atau pengelompokkan ditujukan agar data yang dihasilkan tersusun dengan rapi dan dapat mengidentifikasi pola dan tren dalam data [33]. Dalam dataset ditemukan atribut seperti "tahun.akademik","tanggal.daftar", "program.studi", "jumlah.mhs", "jumlah.grade", "kode.sekolah", "nama.sekolah". Atribut tersebutlah akan dilakukan pengelompokkan dan pengkategorian data agar data tersusun dan mudah dipahami.

4. *Modelling Data*

Pada tahap *Modelling data* adalah proses memvisualisasikan, mengelompokkan, dan melakukan pengurangan dimensi model dari data. Dengan melihat pola dan tren data yang unik dapat dijadikan model terbaik di mana model terbaik adalah model yang prediktif hasilnya. Pada penelitian ini penulis dibantu aplikasi *R Software* dengan meng*install package Random*. *Random Forest* berfungsi untuk menampilkan hasil prediksi dari semua individu pohon, dengan melihat mana kelas yang mendapat suara terbanyak (Brownlee, 2020).

Metode *Random Forest* merupakan salah satu metode yang digunakan untuk klasifikasi dan regresi. *Random Forest* dapat digunakan untuk *time series* dengan cara meningkatkan keakuratan metode klasifikasi dengan menggabungkan metode

klasifikasi dalam kata lain *Random Forest* dilakukan secara *ensemble* (Muhammad et al., 2017).

$$m(\boldsymbol{u}) = \boldsymbol{u}\mathrm{E}[\vartheta|\boldsymbol{u} = \mathbf{u}] \quad (1)$$

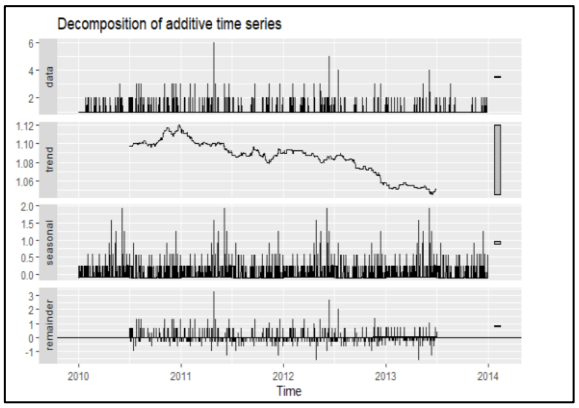
Dimana:
$\underset{\sim}{\boldsymbol{y}}$ = vektor acak dengan elemen *k*,
$m(\boldsymbol{u})$ = fungsi regresi

5. *Interpreting Data*
Dengan Tahap terakhir adalah menafsirkan atau mengkomunikasikan data dengan penyajian data, menyampaikan keefektifan hasil data prediksi dan dapat dipahami oleh orang lain merupakan tujuan dari tahap ini. Pada penelitian ini akan menggunakan *package visualisation tools yaitu* ggplot. Ggplot mampu menampilkan grafik yang merepresentasikan data numerik dan univariat secara simple dari data yang kompleks.

HASIL DAN PEMBAHASAN

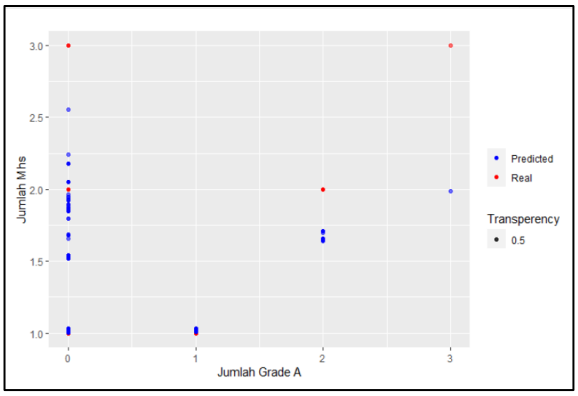
Dalam hasil penelitian jumlah mahasiswa baru dengan menggunakan dataset yang didapatkan, terdapat beberapa visualisasi data dengan R *Software*. Visualisasi data terdiri dari jumlah mahasiswa Universitas XYZ, plot *random forest*, plot *time series* pendaftaran mahasiswa, hingga dekomposi *time series.*

1. *Visualisasi Trend*
Di bawah merupakan dekomposi *time series*, yang membentuk *trend* dalam beberapa waktu yang lampau. Pola data *seasonal* dan *remainder* terjadi kenaikan dan penurunan yang tidak konsisten. Sedangkan pada pola data *trend* mengalamai penurunan dalam beberapa waktu ke depan seperti Gambar 2.



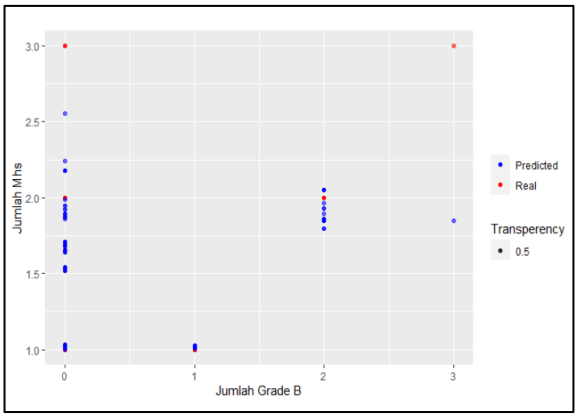
Gambar 2. Visualisasi Trend

2. *Visualisasi hasil prediksi jumlah Grade.A*
Pada gambar di bawah terdapat 2 mahasiswa yang tidak dapat nilai grade A dan 3 mahasiswa yang tidak mendapatkan grade A, untuk prediksinya akan ada beberapa mahasiswa yang akan menyebar secara tidak konstan. Jumlah grade A dengan jumlah mahasiswa 2 mendapatkan hasil prediksi lebih mengarah ke yang lebih rendah. Dan terdapat 3 mahasiswa dengan grade A yang akan diprediksi menurun. Sedangkan, pada prediksi yang tidak mendapatkan nilai A sama sekali beragam khususnya pada titik plot biru. Kesimpulan pada grafik jumlah mahasiswa yang *real* selalu mendapatkan nilai yang genap, sedangkan pada hasil prediksi mendapatkan nilai persebaran distributif dengan nilai desimal seperti pada Gambar 3.



Gambar 3. Hasil prediksi Jumlah Grade A

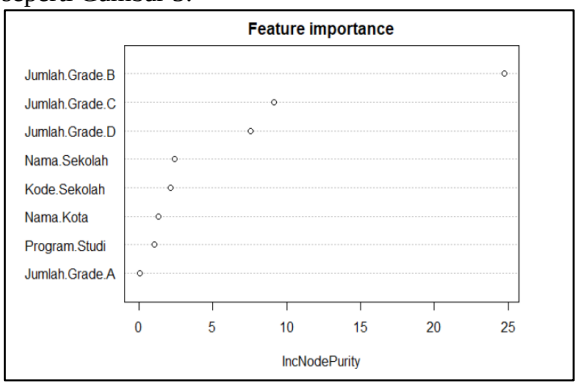
3. *Visualisasi hasil prediksi jumlah Grade.B*
Pada gambar di bawah menunjukkan hasil prediksi jumlah grade B pada jumlah mahasiswa dengan kuantitas sebanyak 2 akan menurun. Lalu, hasil prediksi jumlah grade B pada jumlah mahasiswa dengan kuantitas sebanyak 3 akan menurun dari 3 ke 1.75 poin. Hasil prediksi yang warna biru (*predicted*) sama dengan nilai yang sebenarnya, sehingga pada visualisasi dibawah terlihat tidak jauh berbeda seperti Gambar 4.



Gambar 4. Hasil prediksi Jumlah Grade B

4. *Visualisasi variabel penting*
Dibawah merupakan beberapa variabel-variabel penting yang dapat digunakan sebagai aspek dalam melakukan prediksi. Variabel dengan rata-rata penurunan impuritas node (*IncNodePurity*) yang lebih tinggi merupakan variabel yang lebih penting. Pada gambar di bawah *IncNodePurity* menunjukkan tertinggi yaitu 25 yang artinya pada variabel

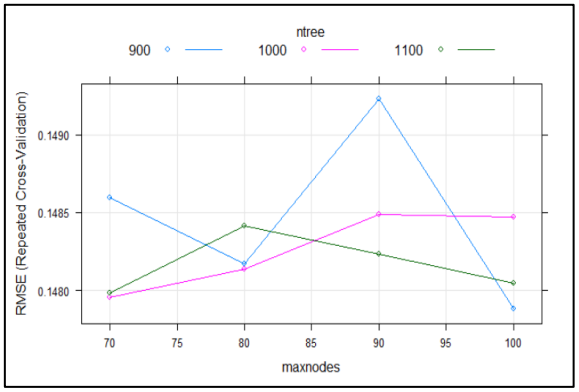
Jumlah.Grade.B merupakan variabel yang penting. Diikuti variabel lainnya seperti jumlah.Grade.C, jumlah.Grade.D, Nama.Sekolah, Kode.Sekolah, Nama.Kota, Program.Studi, dan Jumlah.Grade.A seperti Gambar 5.



Gambar 5. Variabel Penting

5.Hasil Plot

Pada hasil plot di bawah menunjukkan beberapa varian *ntree* dengan jumlah RMSE yang beraneka ragam serta pergerakan grafik yang fluktuaktif. RMSE (*Root Mean Square Error*) adalah besarnya tingkat kesalahan hasil prediksi, semakin kecil (mendekati angka 0) maka hasil prediksi akan semakin kuat dan akurat (Analytics, 2018). Parameter terbaik dari model di bawah terletak pada *maxnodes* 100 dan *ntree* 900 yang artinya parameter tersebut adalah parameter terbaik seperti Gambar 6 di bawah.



Gambar 6. Plot

a. Nilai MAE (*Mean Absolute Error*)

Didapatkan hasil dari MAE (*Mean Absolute Error*) sebesar 0.0251982062423021 atau 2%. Nilai MAE merepresentasikan rata – rata kesalahan (*error*) absolut antara hasil peramalan dengan nilai sebenarnya (y_test) seperti Gambar 7.

```
[1] "MAE: 0.0251982062423021"
```

Gambar 7. Nilai MAE

b. Nilai *R Squared*

R square bernilai antar 0 – 1 dengan ketentuan semakin mendekati angka satu berarti semakin baik. Pada gambar di bawah didapatkan nilai *r square* sebesar 0,93 seperti Gambar 8.

```
[1] "R2: 0.938069274265157"
```

Gambar 8. Nilai R *Squared*

c. *Tuning Parameter*

Nilai MSE dan MAE dari model sangat baik, yaitu di bawah 5%, dengan tingkat akurasi model hampir 98%. Maka disiapkan nilai N yang lebih sedikit dengan sebanyak 500. *Tuning Parameter* digunakan untuk mencari besaran *node* dan *tree* yang terbaik dari model. Lalu, akan dilakukan pembatasan *maxnodes* sebesar 70,80,90 dan 100. Sedangkan untuk *ntree* dilakukan pembatasan 900, 1000 dan 1100.

Dari hasil prediksi jumlah mahasiswa baru dalam beberapa tahun ke depan akan menurun dengan tingkat akurasi sebesar 98% karena nilai MSE (*Mean Square Error*) dan MAE (*Mean Absolute Error*) dari model sangat baik, yaitu dibawah 5%. Dengan beberapa hasil prediksi dan visualisasi yang telah diperoleh peneliti dapat memberikan beberapa rekomendasi strategi atau perencanaan dalam upaya untuk meningkatkan jumlah mahasiswa, karena Universitas XYZ merupakan salah satu bisnis di bidang pendidikan, untuk itu perlu dilakukannya kegiatan *marketing* atau *promotion*. Seperti melakukan promosi di beberapa sosial media yang banyak digunakan para calon mahasiswa seperti *Facebook*, *Instagram*, *YouTube*, *Twitter* dan *TikTok*. Selain itu pihak Universitas XYZ juga perlu menyusun strategi, hasil prediksi sudah menunjukkan jumlah mahasiswa dengan nilai ujian masuk dengan grade B paling banyak, maka bisa dilakukan strategi dalam proses penyusunan soal tes penerimaan mahasiswa baru.

KESIMPULAN

Berdasarkan hasil analisis peramalan yang telah dilakukan maka dapat diperoleh kesimpulan sebagai berikut: Berdasarkan hasil penelitian didapatkan hasil *forecasting* dengan model *random forest* 3 memiliki variabel penting dalam prediksi yaitu Jumlah.Grade.B, Lalu parameter terbaik dari model terdapat pada *maxnodes* 100 dan *ntree* 900 yang menunjukkan parameter tersebut adalah parameter yang sangat akurat. Peramalan jumlah mahasiswa baru dalam beberapa tahun ke depan akan menurun dengan tingkat akurasi sebesar 98% karena nilai MSE dan MAE dari model sangat baik, yaitu di bawah 5%. Dari hasil peramalan disarankan bagi penelitian selanjutnya dapat mengkombinasi model *Random Forest* dengan model lain yang mendukung agar tingkat keakuratan dan visualisasi data yang lebih baik dan jelas lagi.

REFERENSI

Accenture. (2018). *Business Analysis in the Data Science Age*. IIBA.

Aji Haristu, R. (2019). Penerapan Metode Random Forest untuk prediksi win ratio Pemain PUBG. In *Prodi TI USD Yogyakarta* (Issue 21 (1193)).

Analytics, P. (2018). *How to implement Random Forests in R*. Rbloggers. https://www.r-bloggers.com/2018/01/how-to-implement-random-forests-in-r/

Brownlee, J. (2020). *How To Work Through A Problem Like A Data Scientist*. Machine Learning Mastery.

Coghlan, A. (2018). *A Little Book of R For Time Series. Release 0.2*. 75.

Eka Chandra, N., & Sarinem. (2015). Peramalan Penyebaran Jumlah Kasus Virus Ebola Di Guinea Dengan Metode Arima. *UJMC*, *2*(November), 28–35.

Interviewsqs. (2019). *11 websites to find free, interesting datasets*. Interviewqs.Com. https://www.interviewqs.com/blog/free_online_data_sets

Karmita, S., Putra, A. B. W., Gaffar, A. F. O., & Wiguna, A. S. (2019). Prediksi Jumlah Calon Mahasiswa Baru Menggunakan Fuzzy Time Series-Time Invariant. *Prosiding SAKTI (Seminar Ilmu Komputer Dan Teknologi Informasi)*, *3*(1), 208–214.

Kenton, W. (2020). *What Is a Time Series?* 31 Maret.

LTMPT. (2020). *Prinsip Penerimaan Mahasiswa Baru*. https://ltmpt.ac.id/?mid=10

Muhammad, M., Harjono, & Akhsani, L. (2017). *Peramalan Mahasiswa Baru Ft Dan Fkip Um Purwokerto Dengan Model ArimA 1Malim*. *18*(2), 123–132.

Riadi, M. (2017). *Pengertian, Fungsi dan Jenis-Jenis Peramalan (Forecasting)*. 14 November.

Services, E. E. (2018). *Data Science and Big Data Analytics*. John Wiley.

SHARMA, M. (2018). *Data Cleaning Using R*. Dataanalyticsedge.Com. http://dataanalyticsedge.com/2018/05/02/data-cleaning-using-r

PROFIL PENULIS

Marchell Rianto, lahir di Jakarta pada 10 Agustus 1999. Mahasiswa semester VII Program Studi Sistem Informasi STMIK Mikroskil Medan.

Roni Yunis, lahir di Bukittinggi pada 19 April 1975. Dosen Tetap di Program Studi Sistem Informasi STMIK Mikroskil Medan.