**Journal of Information and Information Security (JIFORTY)**
**Vol. 2, No. 1, Juni 2021, 77-88**
**e-ISSN : 2722-4058**
**Available Online at http://ejurnal.ubharajaya.ac.id/index.php/jiforty**

# Implementasi Big Data Analytical Untuk Perguruan Tinggi Menggunakan Machine Learning

**Rakhmat Purnomo[1,*], Wowon Priatna[1], Tri Dharma Putra[1]**

[1] Fakultas Ilmu Komputer; Universitas Bhayangkara Jakarta Raya; Jl. Perjuangan 081, Marga Mulya, Bekasi Utara; 02188955882; e-mail: rakhmat.purnomo@dsn.ubharajaya.ac.id, wowon.priatna@dsn.ubharajaya.ac.id, tridharma.putra@dsn.ubharajaya.ac.id

* Korespondensi: e-mail: rakhmat.purnomo@dsn.ubharajaya.ac.id

Diterima: 10 Juni 2021; Review: 26 Juni 2021; Disetujui: 29 Juni 2021; Diterbitkan: 3 Juli 2021

## *Abstract*

*The dynamics of higher education are changing and emphasize the need to adapt quickly. Higher education is under the supervision of accreditation agencies, governments and other stakeholders to seek new ways to improve and monitor student success and other institutional policies. Many agencies fail to make efficient use of the large amounts of available data. With the use of big data analytics in higher education, it can be obtained more insight into students, academics, and the process in higher education so that it supports predictive analysis and improves decision making. The purpose of this research is to implement big data analytical to increase the decision making of the competent party. This research begins with the identification of process data based on analytical learning, academic and process in the campus environment. The data used in this study is a public dataset from UCI machine learning, from the 33 available varibales, 4 varibales are used to measure student performance. Big data analysis in this study uses spark apace as a library to operate pyspark so that python can process big data analysis. The data already in the master slave is grouped using k-mean clustering to get the best performing student group. The results of this study succeeded in grouping students into 5 clusters, cluster 1 including the best student performance and cluster 5 including the lowest student performance.*

***Keywords**: Machine Learning, Big Data Analitical, Learning Academic, k-mean Clustering Python, Hadoop, apache spark.*

## Abstrak

Dinamika pendidikan tinggi sedang berubah dan menekankan kebutuhan untuk beradaptasi dengan cepat. Pendidikan tinggi berada di bawah pengawasan lembaga akreditasi, pemerintah, dan pemegang kepentingan lainnya untuk mencari cara baru untuk meningkatkan dan memantau keberhasilan mahasiswa dan kebijakan kelembagaan lainnya. Banyak lembaga gagal memanfaatkan sejumlah besar data yang tersedia secara efisien. Dengan penggunaan big data analytic di perguruan tinggi maka dapat diperoleh wawasan yang lebih tentang mahasiswa, akademisi, dan proses di perguruan tinggi sehingga mendukung analisis prediksi dan peningkatan pengambilan keputusan. Tujuan penelitian ini adalah untuk impelementasi big data analytical untu menigkat pegambilan keputusan pihak kampu. Penelitian ini dimulai dengan identifikasi data-data proses berdasarkan learning analitycal, academic dan proses dilingkungan kampus. Data yang digunakan dalam penelitian ini adalah dataset public dari bersumber dari UCI machine learning, dari 33 varibale tersedia digunakan 4 varibale untuk mengukur kinerja mahasiswa. Big data analisis dalam penelitian ini menggunakan spark apace sebagai library untuk mengoperaiskan pyspark agar python dapat mengolah big data analisis. Data yang sudah ada di master slave dilakukan pengelompokan data menggunakan k-mean clustering untuk medapatkan kelompok mahasiswa yang kinerja terbaik. Hasil dari penelitian ini

berhasil mengelompokan mahasiswa menjadi 5 cluster, cluster 1 termasuk dalam kinerja mahasiswa terbaik dan cluster 5 termasuk kinerja mahasiswa yang terendah.

**Kata kunci:** Machine Learning, Big Data Analitical, Learning Academic, k-mean Clustering Python, Hadoop, apache spark.

## 1. Pendahuluan

Perkembangan Teknologi digital di berbagai ruang dan waktu terus menghasilkan data dalam jumlah besar. Big Data menggambarkan pertumbuhan yang signifikan dalam volume dan variasi data yang tidak dapat lagi dikelola menggunakan database tradisional. Dengan bantuan analitik, jumlah data yang tampaknya berbeda dan heterogen ini dapat diproses untuk pola, yang pada dapat menghasilkan wawasan penting untuk pengambilan keputusan (Daniel, 2015) Dampak Digitalisasi data telah membuka peluang penggunaan big data di perguruan tinggi. Nilai big data terletak pada hasil analisis dan prediksi atau tindakan yang diambil dari hasil analisis dan prediksi tersebut (ADMIN SEVIMA, 2019). Dinamika pendidikan tinggi sedang berubah dan menekankan kebutuhan untuk beradaptasi dengan cepat. Pendidikan tinggi berada di bawah pengawasan lembaga akreditasi, pemerintah, dan pemegang kepentingan lainnya untuk mencari cara baru untuk meningkatkan dan memantau keberhasilan siswa dan kebijakan kelembagaan lainnya (Tulasi, 2013)

Pendidikan masa depan lebih sering dihubungkan dengan teknologi baru, pendidikan tinggi beroperasi dengan lingkungan semakin komplek (Asniar, 2015) seperti perangkat komputasi di mana-mana, desain ruang kelas yang fleksibel. Karena institusi pendidikan tinggi mempunyai banyak data tentang mahasiswa, dan basis data catatan mahasiswa menjadi lebih kompleks untuk diakses, maka Perguruan Tinggi harus mendata semua data terkait akademik dari berbagai kegiatan seperti data mahasiswa, data registrasi, data asesmen dan lainya. untuk mempermudah pendidikan tinggi dalam mengakes yang dapat dihasilkan oleh mahasiswa adalah menerapkan analytical big data dalam e-learning analitycal, dimana institusi dapat menentukan mahasiswa yang sukses atau mahasiswa yang droup out (Asniar, 2015).

Penelitian (Murumba & Micheni, 2017) menerapakan penerapan big data mengeksplorasi atribut big data yang relevan dengan institusi pendidikan. Penelitian (Tulasi, 2014) menerapkan big data analisis dalam analisa pembelajaran dengan menganalisa data statis dan dinamis dilingkungan belajar-mengajar untuk memungkinkan intervensi tepat waktu oleh pendidik sedangkan (Kumar Sinha & Singh, 2019) rekomendasi analisis big data dalam mencapai keberhasilan mahasiswa diberbagai perguruan tinggi. Penelitian (Daniel, 2015) menerapakan big data untuk rekomendasi project untuk proyek big data untuk tingkat kelembangaan institusi.

Beberapa Penelitian menggunakan kombinasi big data dan machine learning dalam menyesaikan permasalah data besar diantaranya: menggunakan machine learning dalam mengolah data besar untuk sisi keamanan (Kaur, Sharma, & Mittal, 2018) untuk memperkirakan

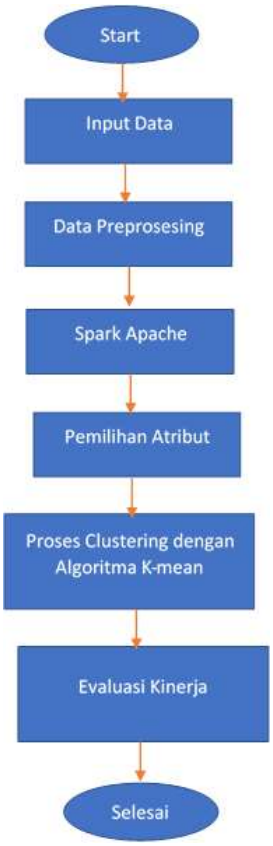
permintaan pariwisata kapal pesiar Tiongkok (Xie, Qian, & Wang, 2021) pengenalan wajah (Vinay et al., 2015) kerangka sistem untuk otonomi (Jamshidi, 2017)

Dari latar belakang dan penelitian tujuan penelitian ini adalah untuk implementasi big data analisis diperguruang tinggi dengan berfocus kepada learning academic dalam menentukan prediksi kinerja mahasiswa menggunakan k-mean clustering.

## 2. Metode Penelitian

Dalam Penelitian ini menggunakan dataset yang bersifat public dapat Students Academic Performance yang bersumber dari https://archive.ics.uci.edu jumlah 32 attribut diambil 4 yang akan digunakan dalam variable yang akan diolah oleh k-mean clustering.

Tahapan dalam penelitian ini dapat dilihat pada gambar 1.




Sumber : Hasil Penelitian (2021)

Gambar 1. Tahapan Penelitian

### 2.1. Input Data

Data yang telah diperoleh berupa format excel akan dirubah dan disesuaikan dengan format csv yang akan digunakan pada pemograman python, selanjutnya akan diinput kedalam directory folder yang nantinya akan ditampilkan pada directory Jupiter.

### 2.2. Data Preprosesing

Dalam tahapan persiapan pemprosesan data adalah mengecek data, memastikan tidak record, attribute yang kosong atau pun format sudah sesuai yang akan diolah python. Data yang akan dioleh adalah data student performace sebagai bagian dari academic analytical dalam penerapan big data analytical dalam perguruan tinggi.

### 2.3. Spark Apache

Dalam tahapan ini adalah instalasi spark apache dan perangkat lunak lainnya. Spark apache ini berfungsi untuk mengoperasikan big data analisis. Dimana beberapa software pendukung yang harus diinstall adalah: Anaconda ,Spark apache, Hadoop,Java development kit (JDK) dan skala. Dalam tahapan ini setelah semua perangkat lunak talah terinstall dan spark apache terlah berhasil digunakan. Selanjutnya akan membuat server dalam spark apache yang dinamakan master slave. Master slave yang telah berhasil dikoneksikan selanjutnya IP address akan digunakan untuk membuat koneksi worker (client).

### 2.4. Data Selection (pemilihan Attribut)

Data set dengan format csv dapat diproses dengan Bahasa pemrograman Python. Maka, rentang data yang akan diolah harus didefinisikan terlabih dahulu. melakukan seleksi rentang data adalah Impor fungsi arrange dari library numpy serta membuat sebuah variabel yang menampung fungsi arrange Definisikan secara numerik jumlah baris data yang ingin diseleksi Proses ini akan menghasilkan variabel yang menampung data yang telah diseleksi.

### 2.5. K-mean Clustering

Untuk menjalankan big data analisis dengan Algoritma k-mean Clustering dipemograman python dibutuhkan beberapa tahapan:

1. Mengimport pyspark (sebuah library big data menggunakan spark apache
2. Mengimport library machine learning (MLlib) yang dibutuhkan oleh python untuk memproses machine learning
3. Mengimport library k-mean clustering
4. Menentukan centroid (titik pusat cluster) menggunakan mean masing-masing kelompok data
5. Mealokasikan data kepada centroid terdekat
6. Proses berulang Melakukan iterasi untuk memastikan tidak ada data yang pindah cluster.

### 2.6. Evaluasi Kinerja

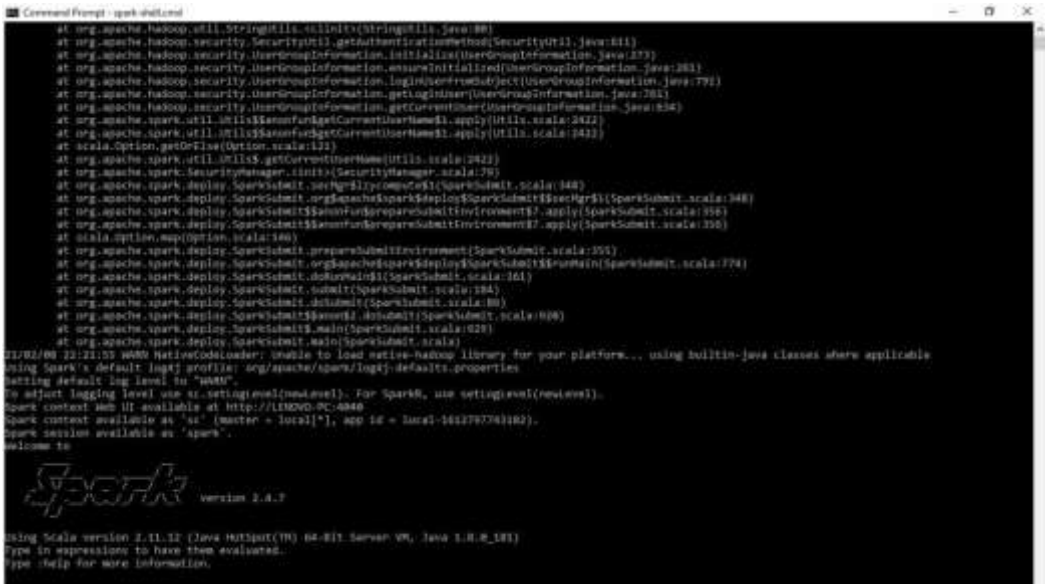
Pada tahapan ini akan menguji kinerja dari hasil prediksi k-mean clustering apakah sudah efekfit sehingga dapat digunakan sebagai rekomendasi model untuk digunakan.

## 3. Hasil dan Pembahasan

### 3.1. Perancangan Big Data

#### 3.1.1. Instalasi Spark apache

Dalam penelitian ini untuk menjalankan Big Data digunakan spark apache untuk memproses data berskala besar. Spark apache menyediakan library untuk mengoperasikan python, distribusi skala, dan menampung Hadoop. Spark apache yang digunakan adalah versi 2.47 digunakan untuk membuat master Komputer sebagai server dan mengoperasikan akses client menggunakan worker. Berikut tampilan spark apache yang sudah terinstall di computer dapat dilihat pada gambar 2.




Sumber : Hasil Penelitian (2021)

Gambar 2. Hasil instalasi Spark Apache

#### 3.1.2. Instalasi apache Hadoop

Apache Hadoop merupakan salah satu distribusi big data yang dikembangkan oleh Apache Software Foundation. Apache Hadoop ini dibangun dengan lisensi free dan open source. Dalam penelitian ini digunakan Apache Hadoop sebagai framework untuk Big Data dengan versi 2.7.0. Instalasi dan konfigurasi dilakukan pertama kali pada komputer yang bertindak sebagai hadoop-master (single node), untuk menjadikan multimode cluster akan dilakukan synkronisasi dengan komputer slave dan dengan melakukan perubahan konfigurasi yang minimal pada node. Secara garis besar tahapan instalasi dan konfigurasi Big Data Hadoop dengan Install Java Development Kit (JDK), Buat user dan group untuk Hadoop, Konfigurasi firewall dinonaktifkan dan Instalasi Hadoop dan konfigurasi Hadoop Environment pada node master dan slave.

### 3.1.3. Konfigurasi Environment

Pada tahapan ini adalah setting environment variable pada my computer agar big data dapat digunakan. Setting environment dapat dilihat pada tabel 1.

Tabel 1. Setting Environtment

| Variable | Fungsi |
|---|---|
| PYSPARK_DRIVER_PYTHON = E:\phyton\scripts\jupyter.exe | Untuk menjalankan python pada spark apache. Dimana editor menggunakan jupyter.exe |
| PYSPARK_DRIVER_PYTHON_OPTS = notebook | Tempat menulis code |
| PYSPARK_PYTHON = E:\phyton\python.exe | Agar psypark dapat dijalankan |
| C:\Program Files (x86)\Common Files\Oracle\Java\javapath | Akses JDK agar bisa mendukung spark apache |
| C:\spark\scala\bin | Untuk menjalankan scala untuk menjalankan data frame, dan SQL |
| C:\spark\spark\bin | Agar spark dapat dioperasikan |

Sumber : Hasil Penelitian (2021)

### 3.1.4. Membangun Cluster Komputer

Untuk menjalankan big data menggunakan spark apache adalah dengan mengakses pyspark. Untuk menjalankan pyspark. Berikut tahapan yang harus dilakukan:

1. Koneksi spark master

   Masuk ke anaconda command prompt kemudian arahkan ke directory dimana spark apache/bin berada. Jalankan koneksi ke master slave sebagai computer slave dengan menggunakan perintah: spark-class org.apache.spark.deploy.master.Master.
2. Koneksi computer worker

   Komputer worker digunakan sebagai client dimana data yang akan deprogram oleh python dilakukan di worker. Untuk masuk ke worker untuk koneksi ke worker:spark-class org.apache.spark.deploy.worker.Worker spark://192.168.43.168:7077. Ip 192.168.43.168:7077 adalah alamat dari master slave, dengan koneksi ke master slave big data yang disini menggunakan pyspark dapat dikoneksikan.
3. Menjalankan spark apache

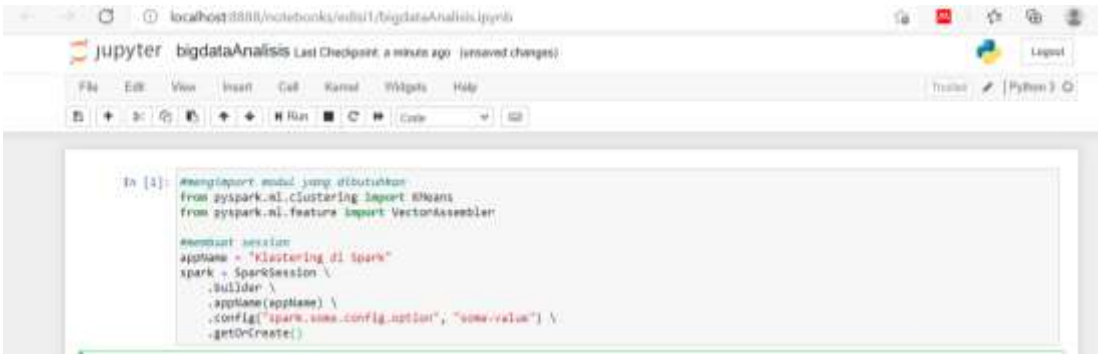
   Setelah master slave dan worker diajalankan maka pyspark dapat dijalankan. Memangil pyspark adalah untuk masuk ke directory dimana data-data atau menjalankan python dilakukan

### 3.2. Implementasi Kmean Clustering menggunakan Spark Apache

### 3.2.1. Preprosesing

Tahapan ini adalah mempersiapkan data yang akan diproses memeriksa memilih, menghapus atau menghapus data yang akan digunakan dalam penelitian ini. Data diolah menggunakan Microsoft excel dan simpan dengan format xls. Data yang telah diolah simpan

didalam directory riset internal yang selanjutnya akan tampil dalam perangkat lunak Jupiter. Berikut listing program untuk persiapan input data dalam dilihat pada Gambar 4.7.




Sumber : Hasil Penelitian (2021)

Gambar 2. Hasil instalasi Spark Apache

Untuk input data dari dataset yang sudah tersimpan di directory jupyter. Dimana ada beberapa dataset yang akan dikelompokan data menggunakan k-mean clustering. Berikut salah contoh salah satu input data dari excel kedalam jupyter menggunakan python, dapat dilihat pada gambar 3.


Sumber: Hasil Penelitian (2021)

Gambar 3. Proses Input Data

### 3.2.2. Data Training

Tahap selanjutnya adalah menyiapkan data training sebelum memproses K-mean clusteringnya. Dimana dalah tahap ini sebelum dimasukan ke model kmean data dirubah kedalam bentuk vector assembler. Berikut listing program tahap training menggunakan python dapat dilihat pada gambar 4.


Sumber : Hasil Penelitian (2021)

Gambar 4. List Program Untuk Data Training

### 3.2.3. Membuat Model K-Mean

Tahapan ini adalah proses membangun model k-mean clustering, menentukan jumlah cluster (kelompok) data yang akan digunakan untuk proses prediksi. Berikut listing program dapat dilihat pada gambar 5.


Sumber : Hasil Penelitian (2021)

Gambar 5. Membangun Model K-mean

### 3.2.4. Menentukan Titik Cluster

Pada tahapan ini adalah menentukan titik pusat cluster (centroid), dimana dalam penelitian ini untuk menentukan nilai awal centroid dilakukan secara acak. Dalam penelitian ini ditentukan kluster (kelompok) sejumlah 5 cluster. Berikut implementasi penentuan titik awal cluster menggunakan Bahasa pemograman phyton, ditampilkan pada gambar 6.


Sumber : Hasil Penelitian (2021)

Gambar 6. Listing Program Untuk Menentukan Centroid

### 3.2.5. Prediksi Cluster

Hasil terakhir dari tahapan pengelompokan data (clustering) dalam algoritma k-Mean adalah menentukan tiap record termasuk dalam cluster berapa. Dimana dalam penelitian ini sudah ditentukan jumlah cluster adalah 5. Berikut implementasi python untuk memprediksi cluster, dapat dilihat pada gambar 7.

```
In [30]: 1 prediction = model.transform(train)#melakukan prediksi kloster
         2 prediction.groupBy("prediction cluster").count().orderBy("prediction cluster").show()
         3 prediction.select('no', 'prediction cluster').show(395)#menampilkan 5 data hasil prediksi
```

Sumber : Hasil Penelitian (2021)

Gambar 7. Listing Program Untuk Prediksi Cluster

Hasil yang didapatkan dari data record berjumlah 395 menunjukan cluster 0 berjumlah 44, cluster 1 berjumlah 147, cluster 2 berjumlah 127, cluster 3 berjumlah 5 dan cluster 4 berjumlah 72. Lengkap dapat dilihat pada gambar 8.

| prediction cluster | count |
|---|---|
| 0 | 44 |
| 1 | 147 |
| 2 | 127 |
| 3 | 5 |
| 4 | 72 |

Sumber : Hasil Penelitian (2021)

Gambar 8. Hasil prediksi cluster

Hasil prediksi dari setiap record yang telah dilakukan proses clustering dengan k-mean dapat dilihat dari gambar 9.

```
In [8]:  1 df_pred = no.join(transformed, 'no')
         2 df_pred.show()
```

| No | absences | G1 | G2 | G3 | prediction cluster |
|---|---|---|---|---|---|
| 1 | 6 | 5 | 6 | 6 | 1 |
| 2 | 4 | 5 | 5 | 6 | 0 |
| 3 | 10 | 7 | 8 | 10 | 4 |
| 4 | 2 | 15 | 14 | 15 | 2 |
| 5 | 4 | 6 | 10 | 10 | 1 |
| 6 | 10 | 15 | 15 | 15 | 2 |
| 7 | 0 | 12 | 12 | 11 | 1 |
| 8 | 6 | 6 | 5 | 6 | 1 |
| 9 | 0 | 16 | 18 | 19 | 2 |
| 10 | 0 | 14 | 15 | 15 | 2 |
| 11 | 0 | 10 | 8 | 9 | 1 |
| 12 | 4 | 10 | 12 | 12 | 1 |
| 13 | 2 | 14 | 14 | 14 | 2 |
| 14 | 2 | 10 | 10 | 11 | 1 |
| 15 | 0 | 14 | 16 | 16 | 2 |
| 16 | 4 | 14 | 14 | 14 | 2 |

Sumber : Hasil Penelitian (2021)

Gambar 8. Hasil prediksi cluster

### 3.3. Hasil Analisis implementasi Big Data

Dalam penelitian ini menghasilkan pengelompokan data dari dataset student performace yang termasuk bagian dari analytical academic untuk analisis big data di perguran tinggi. Apache Spark merupakan salah satu bagian dari proyek Big Data Apache Software Foundation yang menangani analisis data pada Big Data. Dimana untuk menggunakan k-mean clustering dalam memprediksi data membutuhkan library pyspark.ml.clustering import KMeans. Apache spark digunakan oleh python agar ada computer yang dapat digunakan sebagai server (Master Slave) dan sebagai client (worker) serta untuk menjalankan library Machine Learning atau disingkat dengan MLlib di python.

Hasil yang didapat dalam analisis big data ini, k-mean clustering berhasil melakukan prediksi menjadi 5 cluster sesuai dengan yang ditentukan pada centroid dan tidak terjadi kendala.

## 4. Kesimpulan

Dari hasil penelitian impelementasi analisis big data pada perguruan tinggi menggunakan machine learning disimpulkan bahwa Analisis big data pada perguruan tinggi meliputi learning analytical, academic analytical dan analytical staff. Dalam penelitian ini dilakukan pada bidang analytical academic dengan memprediksi student performance dengan dataset yang didapatkan dataset public. Dalam peneltian Analisis big data dioperasikan menggunakan apache spark selanjutnya untuk proses pengelompokan data menggunakan algoritma k-mean clustering bagian dari algoritma machine learning. Hasil kinerja mahasiswa didapat dari 4 variable yaitu nilai abseces, grade1,grade 2 dan grade 3 dalam dikelompokan kedalam 5 cluster. Dimulai dari cluster 1 termasuk nilai rata-rata yang paling besar, dan cluser ini adalah kelompok dari mahasiswa yang nilai terkecil.

Untuk pengembangan penelitian selanjutnya kriteria learning analytical, academic analytical, dan staff aniytical harus diolah dan dilakukan analis bersama-sama. Untuk hasil

kinerja big data lebih baik, sebaiknya menggunakan computer server atau cloud computing. Sebaiknya menggunakan dataset yang langung diambil dari system informasi akademik tiap perguruan tinggi untuk mendapatkan nilai maximal dalam penelitian

**Ucapan Terima Kasih**

Terima Kasih Untuk LPPM yang sudah Mendanai Penelitian internal sehingga penelitain big data ini bisa diselesaikan.

**Daftar Pustaka**

ADMIN SEVIMA. (2019). Manfaat dan Penggunaan Big Data Analytic untuk Perguruan Tinggi. Retrieved from Sevima website: https://sevima.com/manfaat-dan-penggunaan-big-data-analytic-untuk-perguruan-tinggi/

Asniar. (2015). *Penggunaan Big Data Analytic di Perguruan Tinggi*. (June), 1–5. https://doi.org/10.13140/RG.2.1.4581.9046

Aurelia, R. (2017). *IMPLEMENTASI METODE K-MEANS CLUSTERING DALAM MENGELOMPOKKAN EMOSI SENANG, MARAH, DAN NETRAL BERDASARKAN VOKAL MANUSIA*. Universitas Multimedia Nusantara.

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, *46*(5), 904–920. https://doi.org/10.1111/bjet.12230

Jamshidi, M. M. (2017). A system of systems framework for autonomy with big data analytic and machine learning. *Procedia Computer Science*, *120*, 6. https://doi.org/10.1016/j.procs.2017.11.202

Kaur, P., Sharma, M., & Mittal, M. (2018). Big Data and Machine Learning Based Secure Healthcare Framework. *Procedia Computer Science*, *132*, 1049–1059. https://doi.org/10.1016/j.procs.2018.05.020

Kumar Sinha, S., & Singh, H. (2019). Significance of Big Data and Analytics of Student Success in Higher Education. *Journal of Computer Science and Information Technology*, *8*(11), 7–12. Retrieved from www.ijcsmc.com

Murumba, J., & Micheni, E. (2017). Big Data Analytics in Higher Education: A Review. *The International Journal of Engineering and Science*, *06*(06), 14–21. https://doi.org/10.9790/1813-0606021421

Pratiwi, E. S., & Herlawati, H. (2019). Sistem Informasi Penjualan Katering Berbasis Web Pada CV. Saung Alit Telaga Murni Cikarang Barat Eka. *INFORMATION SYSTEM FOR EDUCATORS AND PROFESSIONALS*, *03*(2), 177–188.

Tulasi, B. (2013). Significance of Big Data and Analytics in Higher Education. *International Journal of Computer Applications*, *68*(14), 21–23. https://doi.org/10.5120/11648-7142

Tulasi, B. (2014). Learning Analytics and Big Data in Higher Education. *International Journal of Engineering Research & Technology (IJERT)*, *3*(1), 3377–3383.

Vinay, A., Shekhara, V. S., Rituparna, J., Aggrawal, T., Balasubramanya Murthy, K. N., & Natarajan, S. (2015). Cloud based big data analytics framework for face recognition in social networks using machine learning. *Procedia Computer Science*, *50*, 623–630. https://doi.org/10.1016/j.procs.2015.04.095

Xie, G., Qian, Y., & Wang, S. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management*, *82*(October 2019), 104208. https://doi.org/10.1016/j.tourman.2020.104208