SIGMA
JURNAL TEKNOLOGI PELITA BANGSA
Vol.12 No.1 Maret 2020 p-ISSN: 2407-3903 e-ISSN: 2407-3903
Diterima XX XXX 2023 | Direvisi XX XXX 2023 | Dipublikasikan XX XXX 2023

Analisis Data Jumlah Pegawai Negri Sipil di Indonesia Menggunakan Pyspark dengan Metode Regresi Linear

Rani Sarifah faujiah[1], Ananto Tri Sasongko[2]
[1]Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[1]ranisrfh36@gmail.com, [2] ananto@pelitabangsa.ac.id

Abstract

Growth and changes in the number of Civil Servants (PNS) in Indonesia is a crucial issue in human resource management in government. In an effort to understand and forecast staffing dynamics, data analysis becomes a very important approach. Using PySpark and linear regression methods, this study aims to develop a predictive model that can provide deep insights into the future development of the number of civil servants. A linear regression approach will help identify factors that are significant in influencing personnel growth, thus enabling governments to take more informed decisions in human resource planning and management. PySpark, as a powerful data analysis tool, will be used to efficiently address large-scale staffing datasets. These analysis steps will include data loading, preprocessing to address anomalies or missing values, and training linear regression models. The trained model will be used to make predictions of the number of civil servants in the future. Thus, this study not only provides an overview of civil servant staffing projections, but also provides a strong analytical foundation for sustainable human resource policies at the government level. This analysis is expected to make a positive contribution in strategic planning and personnel development of civil servants in Indonesia.

Keywords*: Civil Servant, PySpark, Linear Regression, Data Analysis, Indonesia.*

Abastrak

Pertumbuhan dan perubahan jumlah Pegawai Negeri Sipil (PNS) di Indonesia merupakan isu krusial dalam pengelolaan sumber daya manusia di pemerintahan. Dalam upaya untuk memahami dan meramalkan dinamika kepegawaian, analisis data menjadi pendekatan yang sangat penting. Menggunakan metode PySpark dan regresi linier, penelitian ini bertujuan untuk mengembangkan model prediktif yang dapat memberikan wawasan mendalam tentang perkembangan jumlah PNS di masa depan. Pendekatan regresi linier akan membantu mengidentifikasi faktor-faktor yang signifikan dalam mempengaruhi pertumbuhan personel, sehingga memungkinkan pemerintah untuk mengambil keputusan yang lebih tepat dalam perencanaan dan manajemen sumber daya manusia. PySpark, sebagai alat analisis data yang kuat, akan digunakan untuk menangani kumpulan data kepegawaian skala besar secara efisien. Langkah-langkah analisis ini akan mencakup pemuatan data, prapemrosesan untuk mengatasi anomali atau nilai yang hilang, dan melatih model regresi linier. Model terlatih akan digunakan untuk membuat prediksi jumlah pegawai negeri sipil di masa depan. Dengan demikian, studi ini tidak hanya memberikan gambaran proyeksi kepegawaian PNS, tetapi juga memberikan landasan analisis yang kuat untuk kebijakan sumber daya manusia yang berkelanjutan di tingkat pemerintah. Analisis ini diharapkan dapat memberikan kontribusi positif dalam perencanaan strategis dan pengembangan kepegawaian PNS di Indonesia.

Kata kunci: PNS, PySpark, Regresi Linear, Analisis Data, Indonesia.

1. Pendahuluan

PNS adalah orang-orang yang dipilih dalam pemilihan berdasarkan peraturan-peraturan umum dan mereka yang bukan dipilih tetapi diangkat menjadi anggota dewan perwakilan rakyat dan anggora dewan daerah serta kepala desa dan sebagainya. Pengertian PNS menurut KUHP sangatlah luas, namun pengertian tersebut hanya berlaku dalam hal orang-orang yang melakukan kejahatan atau pelanggaran jabatan dan tindak pidana lain yang disebutkan dalam KUHP. Jadi pengertian ini tidak termasuk dalam hukum kepegawaian[1]. Pyspark merupakan library Python yang menggunakan bantuan tools berupa Apache Spark Pyspark merupakan modul Python untuk mengakses, memproses, dan menganalisis data dengan bantuan Apache Spark. Apache Spark menggunakan model pemrosesan in-memory dan menggunakan cluster komputasi untuk memproses data secara parallel, memungkinkan tugas pemrosesan data besar diselesaikan dengan cepat[2].Library tersebut digunakan untuk meningkatkan performa dalam melakukan pembersihan data. Hal ini disebabkan oleh besarnya data yang akan digunakan dan proses eksekusi akan memakan waktu yang sangat lama apabila hanya menggunakan bahasa Python saja. Proses pembersihan data ini meliputi penghilangan data yang kosong (missing data), penghapusan tanda baca di kalimat, case folding (pengubahan huruf kapital menjadi huruf kecil), penghapusan stopwords pada kalimat (misalnya of, or, dll), stemming (pengubahan suatu kata menjadi kata dasar), dan proses tokenization (pemecahan kalimat menjadi kata per kata)[3].Metode regresi linear dipilih sebagai alat analisis utama untuk memodelkan hubungan antara variabel independent seperti tahun dan jumlah PNS. Dengan menggunakan regresi linear, kita dapat meramalkan tren pertumbuhan PNS di masa depan berdasarkan pola historis dan faktor-faktor yang mempengaruhi. Analisis prediksi ini memiliki dampak strategis dalam perencanaan sumber daya manusia, memungkinkan pemerintah untuk mengantisipasi kebutuhan tenaga kerja PNS di berbagai sektor atau wilayah. Selain itu, dapat menjadi landasan untuk menyusun kebijakan yang lebih adaptif dan responsif terhadap dinamika perubahan jumlah PNS[4].Dengan menggunakan Pyspark dan regresi linear, analisis prediksi jumlah PNS terbanyak di Indonesia bukan hanya memberikan gambaran masa lalu, tetapi juga menjadi alat proaktif dalam pengambilan keputusan yang terinformasi, memastikan efisiensi dan efektivitas dalam manajemen sumber daya manusia di sektor pemerintahan.

2. Metode Penelitian

Metode penelitian yang digunakan yaitu metode penelitian kuantitatif yaitu pendekatan penelitian yang mengumpulkan dan menganalisis data berupa angka atau statistic. Variabel yang digunakan yaitu variable independent Adapun data yang digunakan menggunakan data time series data yang didapatkan dari hasil perhitungan waktu ke waktu seperti jumlah pns se indonesia pertahun. Tahapan metodologi yang dilakukan pada penelitian ini terdiri dari beberapa proses yaitu pengumpulan data, preprocessing data, pemodelan topik, data prosessing dan analisis dari topik yang dihasilkan[5].

1.1. Data Penelitian

Data yang digunakan pada penelitian ini yaitu jumlah pns menurut Kota dan Kabupaten di Indonesia selama 3 Tahun mulai Tahun 2017 sampai Tahun 2019 yang di peroleh dari Badan Pusat Statistik di Indonesia. Berikut ini merupakan contoh datanya:

	2017	2018	2019
Provinsi Jawa Barat	22084	20841	19569
bogor	9 368	8 619	8 120
Sukabumi	7 896	7 158	6 760
Cianjur	7 138	6 310	6 028
Bandung	8 043	5 486	6 993
Garut	8 627	5 833	7 417
Tasikmalaya	6 502	3 745	5 447
Ciamis	5 263	6 125	4 342
Kuningan	6 415	7 392	5 524
Cirebon	7 069	5 121	6 125

Gambar 1 data pns

1.2. Regresi Linear

Regresi Linear adalah metode untuk menyelidiki hubungan antara satu variabel terikat dengan satu variabel bebas. Regresi digunakan untuk mengukur hubungan dua variabel atau lebih yang dinyatakan dengan bentuk hubungan dan fungsi. Regresi Linear ialah bentuk hubungan dimana variabel bebas X maupun Variabel tergantung Y sebagai faktor yang

berpangkat satu[6]. Regresi linier digunakan untuk melihat hubungan antar dua varibale atau lebih[7]. Regresi Linear sering digunakan adalah regresi linear sederhana dengan bentuk fungsi pada rumus berikut:

$Y = a + bX$ (1)

Keterangan :

Y = Variabel dependen

X = Variabel independen

a = Konstanta / Intercept

b = Koefisien regresi / Slope

$$a = \frac{\sum y(\sum x^2) - \sum x.\sum xy}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2}$$

3. Hasil dan Pembahasan

3.1 Data Selection

Data selection merupakan bagian penting dari proses analisis data dan machine learning karena dapat mempengaruhi kualitas analisis dan hasil prediksi[8]. Data yang digunakan yaitu data jumlah pns se Indonesia selama 3 Tahun mulai Tahun 2017 sampai Tahun 2019 berjumlah 1039 dengan 4 atribut yang di peroleh dari Badan Pusat Statistik Indonesia. Berikut ini adalah datanya:

```
In [6]: data.show()
+-------------+----+----+----+
|      Wilayah|2017|2018|2019|
+-------------+----+----+----+
|        bogor|8120|8619|8120|
|     Sukabumi|6760|7158|6760|
|      Cianjur|6028|6310|6028|
|      Bandung|6993|5486|6993|
|        Garut|7417|5833|7417|
|  Tasikmalaya|5447|3745|5447|
|       Ciamis|4342|6125|4342|
|     Kuningan|5524|7392|5524|
|      Cirebon|6125|5121|6125|
|   Majalengka|5256|7918|5256|
|     Sumedang|4689|5848|4689|
|    Indramayu|6236|4687|6236|
|       Subang|5813|6489|5813|
|   Purwakarta|3610|5794|3610|
|     Karawang|5443|6609|5443|
|       bekasi|5219|5611|5219|
|Bandung Barat|3644|3835|3644|
|  Pangandaran|1777|1910|1777|
|    KotaBogor|3406|6754|3406|
|Kota Sukabumi|1767|3578|1767|
```

Gambar 2 data selection

3.2 Data Transformation

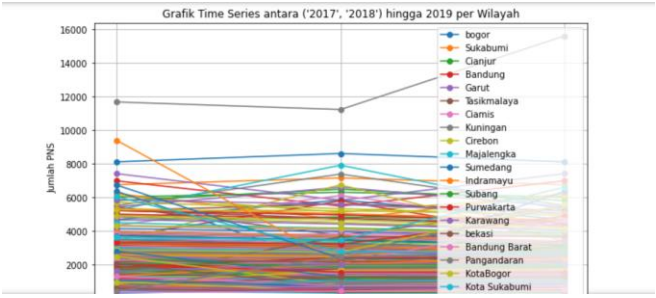
Pada tahap ini tidak diperlukan transformasi karena *value* pada atribut yang digunakan adalah total produksi semuanya telah berupa data *numeric*/ *integer*. Tujuan Transformasi Dataset bertujuan untuk mengubah dataset menjadi dataframe dengan menggunakan library Pandas, dengan begitu dataset dapat dibaca oleh Jupyter Notebook sebagai platform penelitian[9].

```
In [5]: data.printSchema()
root
 |-- Wilayah: string (nullable = true)
 |-- 2017: integer (nullable = true)
 |-- 2018: integer (nullable = true)
 |-- 2019: integer (nullable = true)
```

Gambar 3 data transformation

3.3 Data Mining

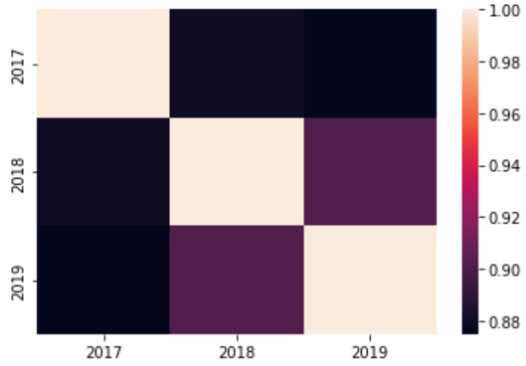
Data mining merupakan sebuah analisa dari observasi data dalam jumlah besar untuk menemukan hubungan yang tidak diketahui sebelumnya dan dua metode baru untuk meringkas data agar mudah dipahami serta kegunaannya untuk pemilih data[10].Berikut ini adalah tampilan grafik time series jumlah pns per wilayah.



Gambar 4 grafik time series

3.4 Data Menghitung Korelasi Antar Data

Pada tahap ini digunakan untuk membuat sebuah heatmap dari matriks korelasi antar variabel-variabel numerik dalam DataFrame. Berikut ini adalah tampilan gambarnya:



Gambar 5 korelasi antar data

3.5 Proses Data Testing dan Data Training

Pada tahap ini dilakukan proses data testing dan data training yang dibagi menjadi data testing 20% dan data training 80%. Dapat dilihat pada gambar berikut:

```
#PEMISAHAN DATA TRAINING DAN DATA TESTING
# Memisahkan data menjadi data latih dan data uji dengan perbandingan 80:20
train_data, test_data = selected_data.randomSplit([0.8, 0.2], seed=42)
train_data.show()
```

Gambar 6 data testing dan training

3.6 Menampilkan Hasil Prediksi

Pada tahap ini menampilkan kolom wilayah,tahun dan prediction dari DataFrame predictions. Berikut ini adalah tampilan hasilnya:

Wilayah	2017	2018	2019	prediction
ACEH BESAR	2183	2073	4647	4647.0
ACEH TAMIANG	1779	1708	3065	3065.0
ACEH TENGGARA	2307	2172	2778	2778.0
BANDA ACEH	11682	11234	15613	15613.0
Bangkalan	490	583	585	585.0
Batu Bara	1290	1288	1240	1240.0
Bojonegoro	314	567	564	564.0
Cirebon	6125	5121	6125	6125.0
Gunungsitoli	1168	1129	1046	1046.0
Humbang Hasundutan	1406	1340	1241	1241.0
Indragiri Hilir	3138	2854	2737	2737.0
Indramayu	6236	4687	6236	6236.0
Jakarta Barat	4305	4054	3682	3682.0
Jakarta Utara	3078	2983	2803	2803.0
Kab Serang	29	29	29	29.000000000000036
Kab. Pasaman	1822	1674	1616	1616.0
Kabupaten Banyumas	5710	5339	4925	4925.0
Kabupaten Sukoharjo	3370	3150	2898	2898.0
Kepulauan Anambas	926	884	923	923.0
Kota Bandung	6367	1874	6367	6367.0

only showing top 20 rows

Gambar 7 hasil prediksi

3.7 Evaluasi Model Regresi Linear

Pada tahap ini melakukan Evaluasi kinerja model regresi menggunakan Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan R-squared (R2). Berikut adalah proses menampilkan evaluasi terebut:

```
In [25]: from pyspark.ml.evaluation import RegressionEvaluator

         # Misalnya, gunakan RMSE sebagai metrik evaluasi
         evaluator_rmse = RegressionEvaluator(labelCol='2019', predictionCol='prediction', metricName='rmse')
         evaluator_mae = RegressionEvaluator(labelCol='2019', predictionCol='prediction', metricName='mae')
         evaluator_r2 = RegressionEvaluator(labelCol='2019', predictionCol='prediction', metricName='r2')

         # Hitung nilai metrik untuk model
         rmse = evaluator_rmse.evaluate(predictions)
         mae = evaluator_mae.evaluate(predictions)
         r2 = evaluator_r2.evaluate(predictions)

         # Tampilkan hasil evaluasi
         print(f'Root Mean Squared Error (RMSE): {rmse}')
         print(f'Mean Absolute Error (MAE): {mae}')
         print(f'R-squared (R2): {r2}')

         Root Mean Squared Error (RMSE): 7.556785117495326e-14
         Mean Absolute Error (MAE): 1.4786970446899383e-14
         R-squared (R2): 1.0
```

Gambar 8 model regresi Linear

4. Kesimpulan

Berdasarkan hasil analisis data jumlah Pegawai Negeri Sipil (PNS) di Indonesia menggunakan Pyspark dengan metode Regresi Linear, dapat kita simpulkan bahwa Nilai RMSE sangat mendekati nol (7.556785117495326e-14), menunjukkan bahwa model regresi linear dapat dengan sangat baik memprediksi jumlah PNS di Indonesia. Semakin kecil nilai RMSE, semakin baik model dalam melakukan prediksi.MAE juga mendekati nol (1.4786970446899383e-14), yang menandakan bahwa perbedaan antara nilai prediksi dan nilai sebenarnya sangat kecil. Semakin kecil nilai MAE, semakin akurat model dalam melakukan prediksi. Nilai R-squared yang mendekati 1.0 (1.0) menunjukkan bahwa model regresi linear mampu menjelaskan variasi data jumlah PNS dengan sangat baik. Nilai R2 mencerminkan proporsi variasi dalam data yang dapat dijelaskan oleh model. Nilai 1.0 menunjukkan tingkat penjelasan yang optimal. Secara keseluruhan, hasil ini menunjukkan bahwa model regresi linear yang diterapkan pada data jumlah PNS di Indonesia menggunakan Pyspark memiliki kinerja yang sangat baik. Model ini dapat digunakan untuk melakukan prediksi jumlah PNS dengan tingkat akurasi yang tinggi, mengingat nilai-nilai evaluasi model (RMSE, MAE, dan R2) mendekati nol atau sempurna.

Referensi

[1] H. S. Nugraha, D. Simarmata, And U. J. Abstrak, “Politik Hukum Pengaturan Netralitas Aparatur Sipil Negara Dalam Pemilihan Kepala Daerah Tahun 2018.”

[2] Baiq Wilda Al Aluf, S. Kom. , M. S. Ari Hernawan, And S. Kom. , M. E. Gibran Satya Nugraha, “Teknik Distributed Naive Bayes Untuk Analisis Sentimen Ulasan Pelanggan Amazon,” 2023.

[3] R. Septiana Et Al., “Perspektif Wisatawan Mancanegara (Wisman) Terhadap Pariwisata Indonesia Menggunakan Latent Dirichlet Allocation (Lda),” Seminar Nasional Sains Data, Vol. 2023.

[4] R. Dwi Shaputra And S. Hidayat, “Implementasi Regresi Linier Untuk Prediksi Penjualan Dan Cash Flow Pada Aplikasi Point Of Sales Restoran.”

[5] R. Septiana Et Al., “Perspektif Wisatawan Mancanegara (Wisman) Terhadap Pariwisata Indonesia Menggunakan Latent Dirichlet Allocation (Lda),” Seminar Nasional Sains Data, Vol. 2023.

[6] M. Galih And P. Dina Atika, “Prediksi Penjualan Menggunakan Algoritma Regresi Linear Pada Koperasi Karyawan Usaha

Bersama," Journal Of Information And Information Security (Jiforty), Vol. 3, No. 2, Pp. 193–202, 2022.

[7] J. Homepage Et Al., "Malcom: Indonesian Journal Of Machine Learning And Computer Science Prediction System For Determine The Number Of Drug Orders Using Linear Regression," Vol. 2, Pp. 62–70, 2022.

[8] A. K. Hermawan And A. Nugroho, "Analisa Data Mining Untuk Prediksi Penyakit Ginjal Kronik Dengan Algoritma Regresi Linier," Bulletin Of Information Technology (Bit), Vol. 4, No. 1, Pp. 37–48, 2023.

[9] R. Stevi, O. #1, And S. Budi, "Analisis Dataset Google Playstore Menggunakan Metode Exploratory Data Analysis."

[10] I. L. L. Gaol, S. Sinurat, And E. R. Siagian, "Implementasi Data Mining Dengan Metode Regresi Linear Berganda Untuk Memprediksi Data Persediaan Buku Pada Pt. Yudhistira Ghalia Indonesia Area Sumatera Utara," Komik (Konferensi Nasional Teknologi Informasi Dan Komputer), Vol. 3, No. 1, Nov. 2019.