**PENELITIAN DATA ANALISIS MENGENAI PENJUALAN ALPUKAT**

Jonathan Sutanto / 00000053777

Major of Information System, Faculty of Engineering and Informatics, Multimedia Nusantara University, Tangerang, Indonesia

[jonathan.sutanto@student.umn.ac.id](mailto:jonathan.sutanto@student.umn.ac.id)

**Abstrak — penelitian ini dilakukan untuk mengkomparasikan algoritma antara peneliti dengan algoritma kelompok yang digunakan pada week**

**14. Data yang digunakan adalah data penjualan data mengenai harga alpukat dan jumlah penjualan alpukat di berbagai pasar di Amerika Serikat. Diharapkan dengan mengkomparasikan kedua algoritma tersebut bisa didapatkan hasil yang maksimal dalam melakukan penelitian terdapat suatu data.**

1. **METODE ANALISIS**

Dalam penelitian yang dilakukan, terdapat beberapa penggunaan rumus yang digunakan oleh peneliti untuk membantu dalam penelitian. Berikut merupakan rumus yang digunakan oleh peneliti:

1. Klasifikasi *K-Nearest Neighbor*

KNN adalah sebuah metode klasifikasi yang digunakan untuk penelitian suatu objek dengan pembelajaran mengetahui data yang paling dekat dengan objek tersebut[1].

Tiap sampel data yang ada akan menjadi sebuah titik, dan pada sebuah objek yang dipilih akan diperiksa nilai terdekat dari nilai tersebut sehingga bisa memprediksikan klasifikasi berdasarkan data pembelajaran terdekat.

Berikut merupakan rumus *K-Nearest Neighbor* yaitu[2] :

1. Regresi dengan Scatterplot

Analisis Regresi digunakan untuk meneliti hubungan antar dua atau lebih variabel[3]. Dengan menggunakan regresi pada data, peneliti dapat mengetahui hubungan antara hubungan variabel dependen dengan variabel independen lainnya. Scatterplot adalah sebuah grafik yang biasa digunakan untuk mengetahui hubungan antara 2 variabel. Peneliti menggunakan Scatterplot dikarenakan dapat Scatterplot dapat menampilkan hasil dari metode regresi dikarenakan kesamaan fungsinya. Berikut merupakan rumus yang digunakan pada *Scatterplot*[4] :

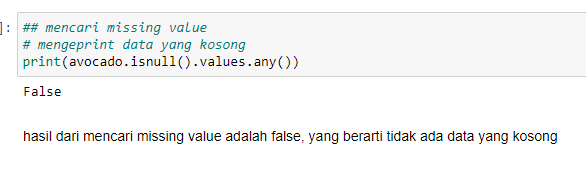
1. Evaluasi. Evaluasi adalah sebuah metode yang digunakan untuk menganalisa apakah hasil dari prediksi penelitian dapat dipercaya atau seberapa akurat nilai prediksi tersebut dengan realitasnya. Pada metode evaluasi, peneliti menggunakan rumus R Squared untuk mengetahui hasil evaluasi pada regresi. Berikut merupakan rumus R Squared.

II. **HASIL PENELITIAN**

Berikut merupakan hasil komparasi antara algoritma 1 yaitu peneliti dengan algoritma 2 yang berasal dari penelitian yang telah dilakukan secara kelompok.

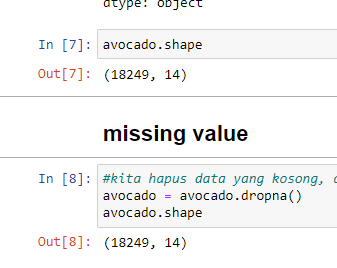
1. *Missing Value*

Pada uji coba pertama, peneliti mencoba untuk mencari *Missing Value*, dan berikut hasil yang didapatkan oleh kedua algoritma. *Missing Value*  digunakan untuk mencari data yang kosong pada sebuah dataset.



Gambar 1 Ujicoba *Missing Value* dengan algoritma 1.

Pada gambar tersebut, dicoba untuk melakukan print pada kolom yang kosong, hasil yang dimunculkan adalah *False*, maka dari itu tidak ada data yang kosong.



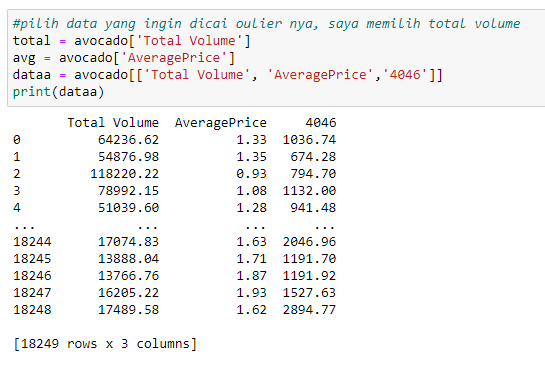
Gambar 2 Ujicoba *Missing Value* dengan algoritma 2.

Pada algoritma 2, kita menggunakan shape untuk melihat data awal sebelum dilakukan pengecekan *Missing Value*, lalu setelah itu dilakukan *dropna* di dataset, setelah itu diperiksa kembali dan masih mendapatkan hasil yang sama. Yang berarti tidak ada yang kosong pada dataset.

1. *Data Preprocessing and Preparation*

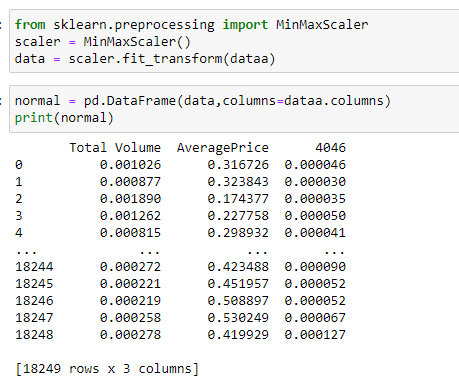
Selanjutnya terdapat ujicoba *Data Preprocessing and Preparation*. *Data Preprocessing and Preparation* adalah teknik awal yang digunakan pada data mining untuk mengubah data mentah menjadi informasi yang bersih, berkualitas dan dapat digunakan dengan baik pada pengolahan data selanjutnya[5]. Terdapat berbagai macam metoda yang digunakan untuk melakukan *Preprocessing and Preparation* data.

Pada algoritma 1, peneliti menggunakan metode Normalisasi. Normalisasi adalah pengelompokan data sehingga data variabel memiliki nilai yang sama. Sehingga data tersebut lebih mudah dibaca dan digunakan dalam analisis[6]



Gambar 3 Data sebelum di normalisasi pada algoritma 1.

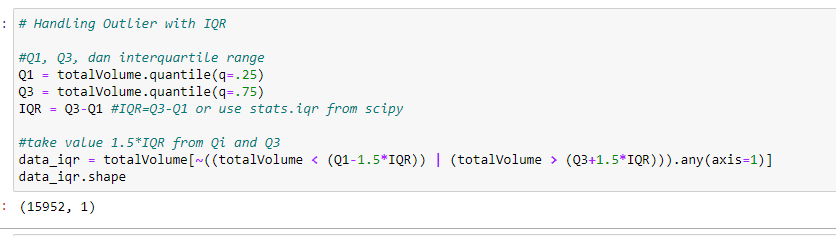
Disini bisa dilihat data masih berjumlah tinggi dan memiliki nilai yang sembarang.



Gambar 4 Data setelah di normalisasi pada algoritma 1.

Setelah di normalisasi, bisa dilihat bahwa data menjadi lebih kecil dan setiap nilai telah dikelompokkan menjadi satu satuan yang sama.

Sedangkan pada algoritma 2, peneliti menggunakan metode *Outlier. Outlier* adalah karakteristik unik yang memiliki nilai jauh dibanding data-data lainnya dan muncul dalam bentuk nilai ekstrim baik untuk sebuah variabel tunggu atau kombinasi[7]. Karena hal tersebut, nilai yang bersifat *outlier* harus dihapus agar tidak mengganggu nilai dari data lainnya pada metode kedepannya.

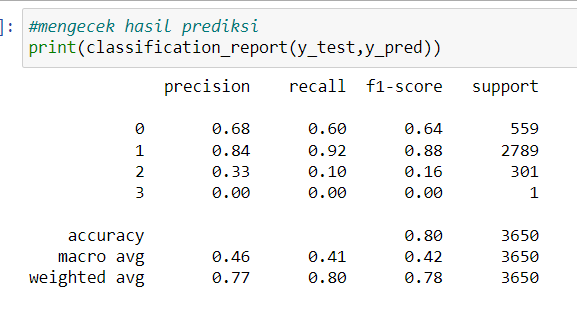


Gambar 5 Penggunaan *IQR* pada *Outlier.*

Seperti yang diketahui pada gambar 3, jumlah row data pada dataset berjumlah 18249 row, tetapi setelah dilakukan uji coba outlier dengan menggunakan rumus *IQR(InterQuartile Range)* didapatkan tersisa 15952 row saja. Ini berarti terdapat 2.297 nilai *Outlier* yang dibuang.

1. Klasifikasi

Selanjutnya peneliti melakukan uji coba klasifikasi. Disini dilakukan peneliti menggunakan algoritma KNN (*K-Nearest Neighbor)*. Dengan menggunakan algoritma tersebut, kita bisa mengklasifikasikan dataset yang ada secara akurat.

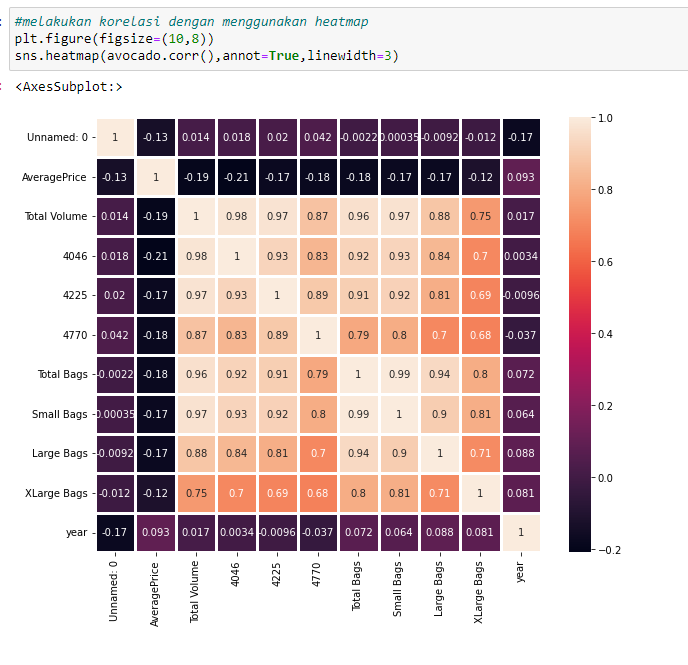
****

Gambar 6 Hasil dari algoritma KNN

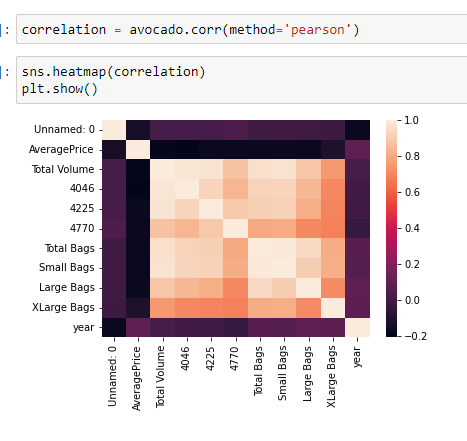
Berikut merupakan gambar hasil dari algoritma KNN pada penelitian 1. Dari hasil tersebut bisa didapatkan informasi bahwa nilai akurasi hasil *micro average* prediksi sebesar 46%, dan *weighted average* sebesar 77%.

1. Korelasi

Pada ujicoba keempat terdapat ujicoba korelasi. Metode korelasi adalah metode yang digunakan untuk mengetahui apakah antar variabel terdapat hubungan atau tidak, dan jika ada seberapa besar/kuat hubungan tersebut. Pada uji korelasi ini kedua algoritma menggunakan rumus *corr* dan dibantu dengan grafik *heatmap.*



Gambar 7 Hasil uji coba korelasi pada algoritma 1.

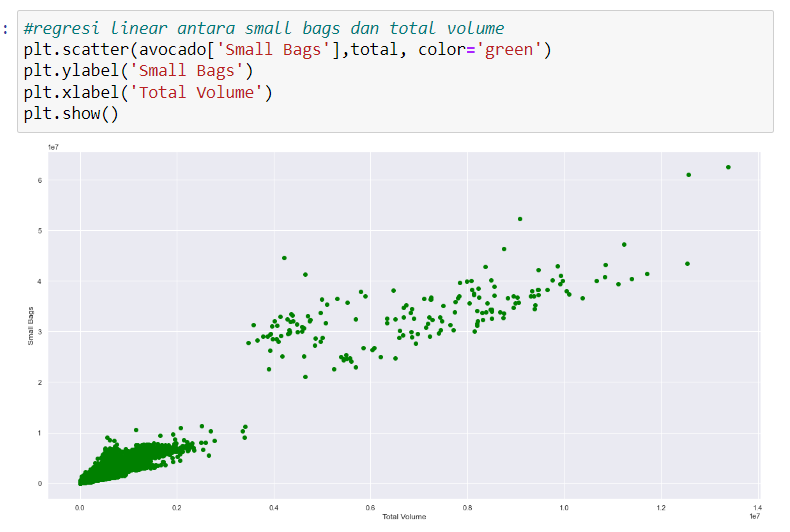


Gambar 8 Ujicoba korelasi pada algoritma 2.

Bisa dilihat pada gambar diatas memiliki grafik yang sama. Terdapat korelasi yang tinggi pada bagian dalam grafik, sedangkan pada bagian luar grafik terdapat korelasi yang rendah.

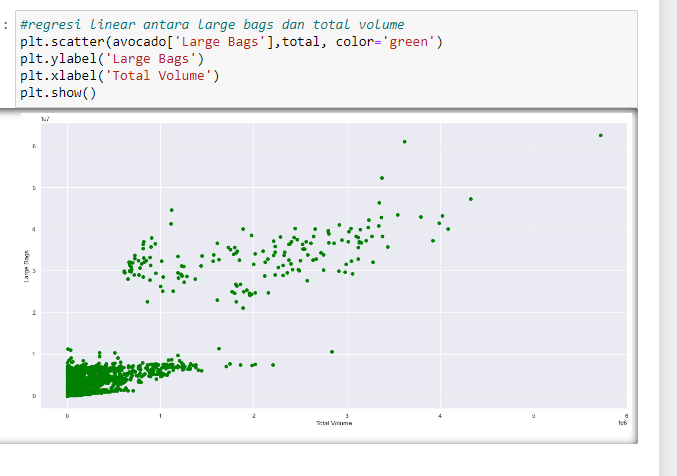
1. Regresi

Ujicoba kelima terdapat uji coba Regresi. Ujicoba regresi berfungsi untuk mengetahui seberapa jauh pengaruh pada 2 variabel yang dianggap saling berhubungan[9] dan berbeda dengan uji coba korelasi yang mencari apakah terdapat hubungan antar variabel yang ada. Pada kedua algoritma, peneliti menggunakan dasar data yang sama, yaitu menggunakan antara *Small Bags & Total Volume* dan *Large Bags & Total Volume*. Dikarenakan hanya menggunakan 2 variabel saja, maka digunakan regresi linier untuk uji coba regresi.



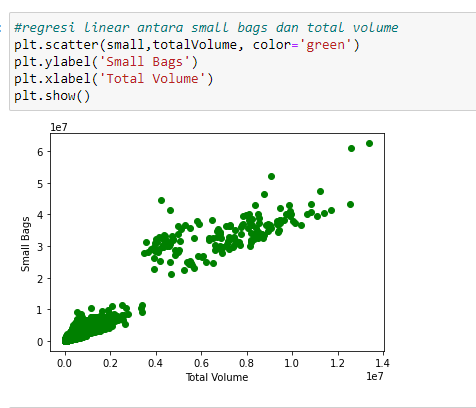
Gambar 9 Uji Regresi antara *Small Bags & Total Volume* pada algoritma 1.

Pada gambar diatas, bisa dilihat bahwa kedua variabel tersebut memiliki hubungan yang positif. Pada bagian awal memiliki hubungan yang sangat kuat tetapi di bagian akhir hubungan menurun menjadi kuat tetapi tetap positif.



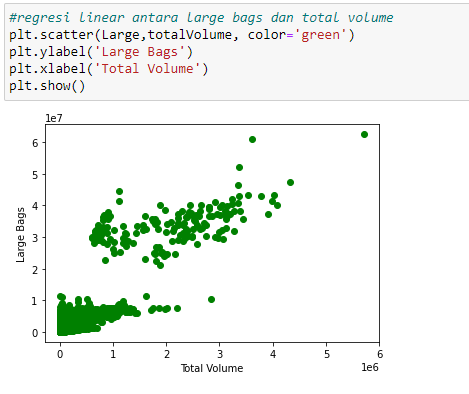
Gambar 10 Uji regresi antara *Large Bags & Total Volume* pada algoritma 1.

Pada gambar diatas, bisa dilihat terdapat hubungan yang sangat kuat pada awal grafik, tetapi setelah itu data mulai sedikit tidak kuat tetapi tetap positif.



Gambar 11 Uji regresi antara *Small Bags & Total Volume* pada algoritma 2.

Pada gambar diatas, bisa dilihat sama dengan hasil pada algoritma 1 yaitu data diawal memiliki hubungan yang sangat kuat, tetapi perlahan menjadi kuat tetapi tetap bersifat positif karena data tetap mengarah ke kanan atas.

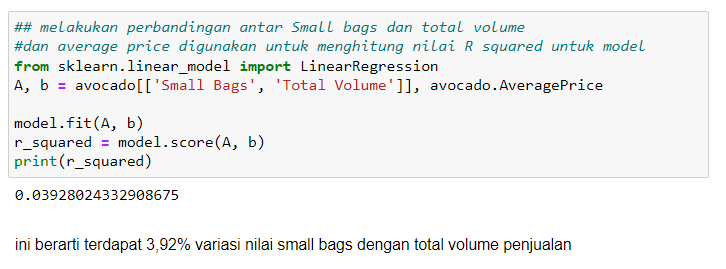


Gambar 12 Uji regresi antara *Large Bags & Total Volume* pada algoritma 2.

Pada regresi di algoritma 2 juga sama dengan algoritma 1. Sehingga hasil dari penelitian regresi dari *Small Bags & Total Volume* dan *Large Bags & Total Volume*  algoritma 1 dan algoritma 2 memiliki hasil yang

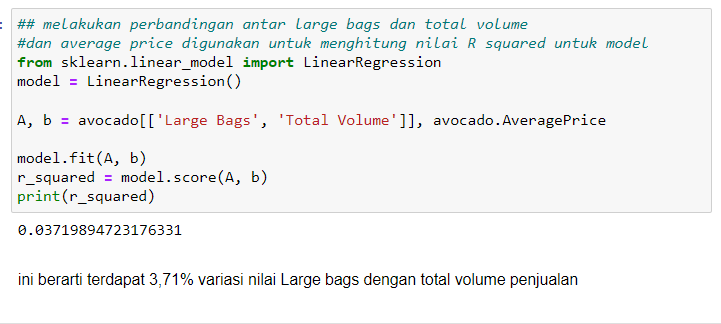
1. Evaluasi

Uji coba terakhir adalah evaluasi. Pada tahap ini, model akan dilakukan evaluasi untuk dicek antara nilai prediksi yang sudah dibuat dengan nilai yang sesungguhnya. Dengan begitu model dapat dinilai seberapa presisi dan ketepatan dan tingkat kesalahan yang ada pada model tersebut[10].



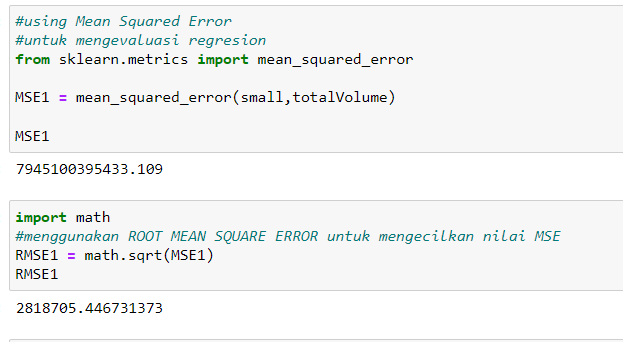
Gambar 13 Uji evaluasi pada regresi pertama dengan algoritma 1.

Pada uji coba diatas, peneliti menggunakan algoritma *R Squared* untuk membantu dalam mencari nilai evaluasi dan didapatkan nilai 3.92% antara *Small Bags & Total Volume.*



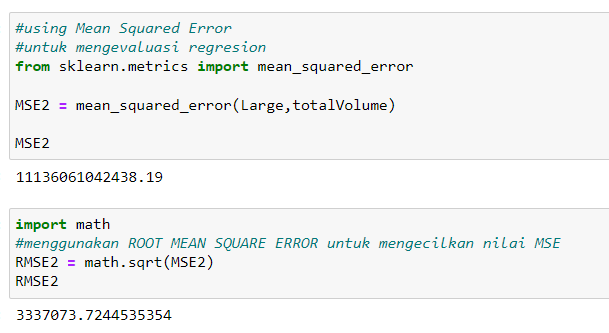
Gambar 14 Uji evaluasi pada regresi kedua dengan algoritma 1.

Lalu selanjutnya, saya melakukan ujicoba pada regresi kedua dengan menggunakan algoritma yang sama dengan uji coba regresi pertama. Dan didapatkan hasil evaluasi 3.71% antara *Large Bags & Total Volume.*



Gambar 15 Uji evaluasi regresi pertama dengan algoritma 2.

Pada algoritma 2, peneliti menggunakan algoritma MSE dan didapatkan nilai yang lumayan tinggi yaitu 7945100395433.109. Lalu untuk mengecilkan nilai tersebut, peneliti menggunakan algoritma RMSE sehingga didapatkan hasil yang lebih kecil dan lebih akurat yaitu 2818705.446731373.



Gambar 16 Uji evaluasi regresi kedua dengan algoritma 2.

Pada hasil evaluasi regresi ini terdapat nilai yang lumayan tinggi yaitu 11136061042438.19. Peneliti juga menggunakan kembali algoritma RMSE untuk mendapatkan hasil yang lebih kecil dan akurat dan didapatkan hasil 3337073.7244535354.

**III. KESIMPULAN DAN SARAN**

1. Kesimpulan

Kesimpulan yang didapatkan dari penelitian ini adalah dengan menggunakan rumus yang berbeda pada metode penelitian akan memberikan hasil yang berbeda. Tetapi perbedaan tersebut hanya berpengaruh pada perhitungan di metode tersebut, sisanya tidak akan berpengaruh secara signifikan. Penelitian ini juga membantu peneliti dalam mengolah data dan mempelajari lebih baik mengenai metode pengolahan data yang lebih baik. Dan juga menjadi diharapkan penelitian ini bisa menjadi ilmu bagi peneliti dan juga pembaca.

1. Saran

Penelitian ini juga tidak luput dari saran yang diberikan oleh rekan peneliti, berikut merupakan saran yang diterima oleh peneliti :

1. Terdapat saran dari Pak Rudi Sutomo selaku dosen pembimbing mata kuliah Data Analysis untuk memberikan masukan mengenai isi dari jurnal penelitian ini.
2. Terdapat saran dari Bill Kiki yaitu membuat penelitian jangan sampai salah dari hipotesis yang ada.
3. Terdapat saran Stanley Jeven yang memberikan masukan mengenai rumus yang digunakan untuk metode metode yang ada.

Peneliti berterima kasih dengan adanya saran tersebut, peneliti dapat membuat jurnal penelitian dengan baik dan sesuai dari rubrik penilaian.

**references**

[1]Dinus, Repository. (2019). Classification Algoritma kNN

(k-Nearest Neighbor). Retrieved from <https://repository.dinus.ac.id/docs/ajar/pert_ke_7_sistem_cerdas.pdf>

[2]Afifah, L. (2020, July 9). *3 Metode Normalisasi Data (Feature Scaling) di Python*. IlmudataPy. Retrieved December 13, 2022, from https://ilmudatapy.com/metode-normalisasi-data/

[3]aquila, R. (2020, 6 30). BAB\_II. *Landasan Teori*, 6. https://kc.umn.ac.id/15266/4/BAB\_II.pdf

[4]BAB IV-88. (18, 2 6). *analisis data dan pembahasan*, 36. http://eprints.ums.ac.id/59154/14/BAB%20IV-88.pdf

[5]DQ Lab. (2021, April 12). *Pentingnya Preprocessing dalam Pengolahan Data Statistik*. DQLab. Retrieved December 13, 2022, from https://www.dqlab.id/pentingnya-preprocessing-dalam-pengolahan-data-statistik

[6]Pragota, A. (n.d.). *Scatterplot: Visualisasi untuk Regresi • Nural Learning*. Nural Learning. Retrieved December 13, 2022, from https://learn.nural.id/course/statistics/regresi-linier/scatterplot

[7]Retnawati, H. (2020, 5 13). Pengantar Analisis Regresi dan Korelasi. 5. http://staffnew.uny.ac.id/upload/132255129/pengabdian/4-materi-Pengantar%20Analisis%20Regresi-alhamdulillah.pdf

[8]Salsabila, S. E. (2021, 2 1). MODEL PREDIKSI PENJUALAN MULTI-ITEM TIME SERIES BERBASIS MACHINE LEARNING MENGGUNAKAN METODE AUTOREGRESSIVE INTEGRATED MOVING AVERAGE DAN LONG SHORT-TERM MEMORY PADA PRODUK PERISHABLE (STUDI KASUS: RETAIL SAYUR TOSAGA). *TUGAS AKHIR*. https://dspace.uii.ac.id/bitstream/handle/123456789/29257/16522173%20Shelly%20Elvina.pdf?sequence=1&isAllowed=y

[9]Universitas Gadjah Mada. (n.d.). *Regresi Linear*. Menara Ilmu Metode Statistika. Retrieved December 13, 2022, from <https://metstat.mipa.ugm.ac.id/teori/regresi-linear/>

[10]andriyani, w. (n.d.). *Apa perbedaan Root Mean Square Error (RMSE) dan Standard Deviation? Bagaimana cara penggunaannya?* Quora. Retrieved December 18, 2022, from https://id.quora.com/Apa-perbedaan-Root-Mean-Square-Error-RMSE-dan-Standard-Deviation-Bagaimana-cara-penggunaannya