

Forecasting Category in Public School Staff Students and Spending in 2018 with Comparison Method: Support Vector Machine and Autoregressive Integrated Moving Average with Python

Cholifaturozzika¹, Amrina Rosyada², Wanda Alifiyah Pramesti³, Dwi Rolliawati⁴

Sistem Informasi, Universitas Islam Negeri Sunan Ampel Surabaya
Gunung Anyar, Surabaya, 60294

e-mail: ¹09020620026@student.uinsby.ac.id, ²09020620020@student.uinsby.ac.id,
³09040620068@student.uinsby.ac.id, ⁴dwi-roll@uinsby.ac.id.

Abstract

Education is one of the efforts to develop capabilities and shape the character and civilization of a dignified nation. Increasingly expensive education costs require parents to prepare more maturely. Therefore, the purpose of this study is to show forecasting of rill costs per student in 2018 using the Timeseries Public School Staff, Students, and Spending 1970-2017 data. In this study, the ARIMA and the SVM model are used to forecast rill costs per student in 2018. ARIMA is one of the simplest models and the most widely used method. This model will be compared using machine learning, namely SVM. From the experimental results that compare the performance of SVM and ARIMA. In the calculation above, the MAPE results from the Arima and SVM models are 0.142 and 0.0029. Whereas for RMSE the results of the ARIMA calculations obtained results of 36.625, while for SVM the RMSE results obtained were 7.470. The results of the study concluded that the SVM model has a better performance in predicting crude oil prices compared to the ARIMA method.

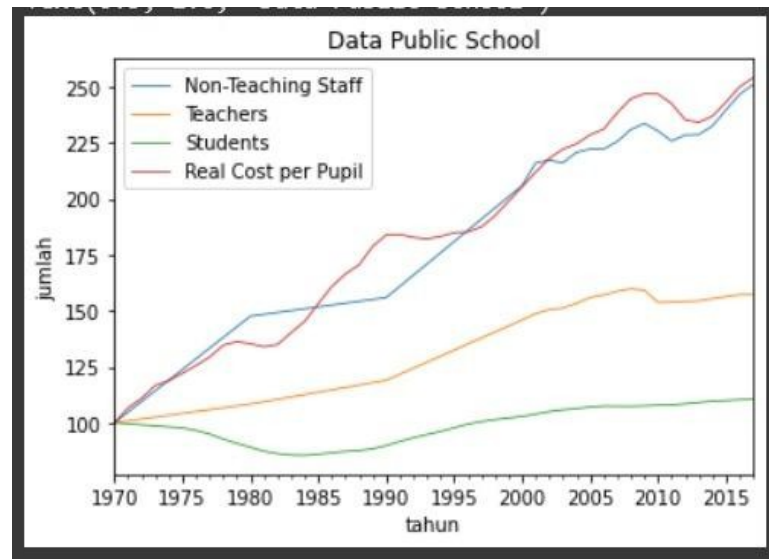
Keyword: ARIMA, Forecasting, Machine Learning, SVM, Timeseries

PENDAHULUAN

Pendidikan merupakan salah satu upaya untuk mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermartabat. Pendidikan juga pada dasarnya merupakan konsep yang mendefinisikan proses menciptakan perubahan dalam perilaku dan struktur pemikiran serta keterampilan (Karacor, dkk. 2018). Pendidikan juga dapat menentukan nilai negara di pasar global dengan keterampilan tenaga kerja dan harga keterampilan tersebut. Struktur Human Capital sebagian besar dibentuk oleh pendidikan.

Berdasarkan Human Capital maka menjadi dasar pentingnya investasi manusia dalam bidang pendidikan. Seiring berkembangnya zaman, individu ingin belajar lebih banyak karena dengan lebih banyak pelatihan berarti mendapat pekerjaan yang lebih baik dan pendapatan yang lebih tinggi. Namun, diketahui pula untuk mendapatkan pendidikan terdapat pula biaya yang harus dikeluarkan setiap individunya. Dalam penelitian ini, peneliti mencoba untuk menunjukkan peramalan biaya pendidikan rill untuk tahun 2018 dengan menggunakan data selang waktu pengeluaran sekolah umum untuk siswa dan staff sepanjang 1970-2017.

Gambar 1 menunjukkan pergerakan jumlah Non-Teaching Staff, Teachers, Students, dan Real cost per pupil selama rentan tahun 1970-2015. Namun demikian, Real cost per pupil dapat diramalkan kedepannya, dan memungkinkan untuk membuat keputusan di masa depan. Pergerakan real cost per pupil merupakan bagian yang akan dianalisis menggunakan variabel dari data Non-Teaching Staff, Teachers, dan Students,



Gambar 1. Menampilkan Data *Public School* Dari Tahun 1970-2015

Dalam artikel ini, model ARIMA (*Autogressive Integrated Moving Average*) dan model SVM (*Support Vector Machine*) digunakan untuk melakukan *forecasting Real cost per Pupil* pada tahun 2018. ARIMA merupakan salah satu model yang paling sederhana dan metode yang banyak digunakan. Model ini akan dikomparasi menggunakan machine learning yaitu SVM. Pemilihan kedua metode ini didasarkan pada penelitian yang dilakukan Makala & Li (2021) dan Pratiwi & Hadijati (2021) yang melakukan prediksi harga emas menggunakan ARIMA dan SVM. Pada penelitian tersebut diketahui bahwa SVM memiliki kinerja yang jauh lebih baik daripada ARIMA. Kemudian penelitian Lestari et al., (2022) melakukan *forecasting* terhadap harga minyak mentah dengan pengaruhnya pada peristiwa ekstrim. Penelitian tersebut menyimpulkan bahwa model SVM memiliki kinerja lebih baik dalam memprediksi harga minyak mentah daripada metode ARIMA.

METODE PENELITIAN

Pada penelitian yang dilakukan tahapan dimulai dengan menentukan dataset yang akan digunakan. Pada penelitian ini menggunakan data *Time Series Public School Staff, Students and Spending 1970-2017*. Data yang digunakan memiliki ciri seri waktu dan memiliki kategori. Data yang digunakan telah disosialisasikan oleh Flourish (flourish.studio). Flourish merupakan salah satu tool online yang digunakan untuk memvisualisasikan data (Mapes & Topham, 2020). Sebelum menerapkan metode peramalan tahapan sebelumnya adalah studi literatur. Ulasan literatur adalah proses menganalisis data penelitian terhadap suatu topik khusus sehingga memahami topik penelitian tersebut. Paper yang telah di review mengenai model peramalan seperti time series, simple moving average, linear regression, *Support Vector Machine* (SVM), *Autoregressive Integrated Moving Average* (ARIMA), dan lainnya. Dari berbagai metode yang telah di review, penelitian ini merupakan komparasi dari dua metode yaitu SVM (*Support Vector Machine*) dan ARIMA (*Autoregressive Integrated Moving Average*). Penggunaan dua metode ini dirasa peneliti sudah seduai dengan kebutuhan penelitian.

1. ARIMA

ARIMA (*Autoregressive Integrated Moving Average*) adalah salah satu teknik statistik yang menangani tren, musim, siklus, dan aspek non stasioner lainnya dari data set. ARIMA sering disebut *timeseries box-jenkins* (Kurniawan et al., 2021). ARIMA adalah model peramalan dalam bidang statistika dan ekonomi yang disajikan dalam deret waktu sehingga memudahkan dalam fluktuasi data. ARIMA biasanya digunakan dalam meramalkan data jangka pendek karena dinilai memiliki nilai keakuratan yang tinggi (Statistika et al., 2022). Sedangkan untuk

jangka waktu, ketepatan peramalannya cenderung konstan atau mendatar sehingga dinilai kurang akurat. ARIMA menggunakan nilai sebelumnya dan saat ini dari variabel dependen untuk menghasilkan peramalan jangka pendek yang akurat. Penggunaan ARIMA akan sesuai dengan Model time series yang terhubung antara satu dengan yang lain. ARIMA hanya menggunakan dataset yang univariat. Box-jenkins terbagi menjadi tiga yaitu AR, MA, dan ARIMA. Dari ketiganya AR dan MA hanya berlaku pada AR karena lebih stasioner Hal ini berarti, stasioner tidak terdapat pertumbuhan atau penurunan pada data.

Dari penjelasan sebelumnya dapat diketahui bahwa time series dataset yang tidak stasioner dapat dilakukan *differencing* untuk menghitung perubahan. ARIMA memiliki 3 model perhitungan, yaitu:

Model Autoregressive (AR)

$$Z_t = +1 Z_{t-1} + a_t \quad (1)$$

Model Moving Average (MA)

$$Z_t = -1 a_t - 1 - 2 a_{t-2} - \dots - q a_{t-q} + a_t \quad (2)$$

Kemudian jika ditemukan data yang dimiliki tidak stasioner, maka perlu tindakan differencing dengan rumus ARIMA (1,1,1) sebagai berikut:

$$Z_t = + (1+1) Z_{t-1} = (1) Z_{t-2} - 1 a_{t-1} + a_t \quad (3)$$

2. SVM

SVM (*Support Vector Machine*) merupakan salah satu algoritma machine learning yang bertujuan menemukan hyperplane terbaik serta memaksimalkan hasil peramalan lebih jauh (Puspita, 2017). *Hyperplane* adalah fungsi yang digunakan untuk pemisah antar kelas atau kategori. *Hyperplane* dapat berupa *line* pada *two dimensional* dan dapat berupa *flat plane* pada *multiple dimensions* (Witanti et al., 2022).

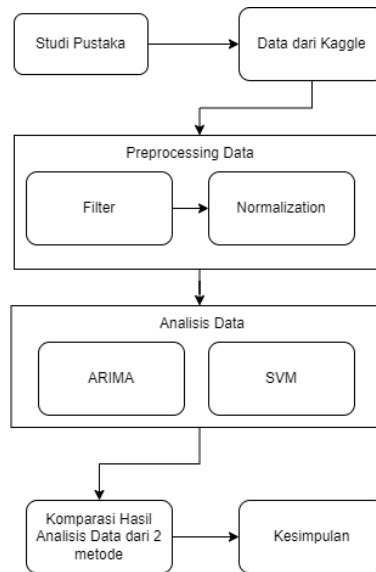
Menurut pendapat Santoso SVM adalah teknik pembelajaran yang hanya menganalisis data untuk klasifikasi dan regresi, hal ini dikarenakan SVM lebih tepat dalam hal kategorisasi. SVM didasarkan pada *linear regression* tetapi dapat mengatasi masalah kategorisasi dan regresi dengan linier maupun non linear. SVM dapat diartikan sebagai teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. SVM memiliki prinsip dasar linear berkategori dimana untuk kasus kategori secara linear akan dipisahkan. SVM telah berkembang dengan dapat bekerja pada kasus non-linear dengan menggunakan konsep kernel berdimensi tinggi. Dimensi tinggi akan dicarikan hyperlane untuk memaksimalkan margin antara kelas data. SVM *hyperplane* dapat dinotasikan sebagai berikut :

$$f(x) = w^T x + b \quad (1)$$

Dari notasi hyperlane tersebut, diperoleh persamaan :

$$w + b \geq 1 \text{ untuk } y_i = +1 \quad (2)$$

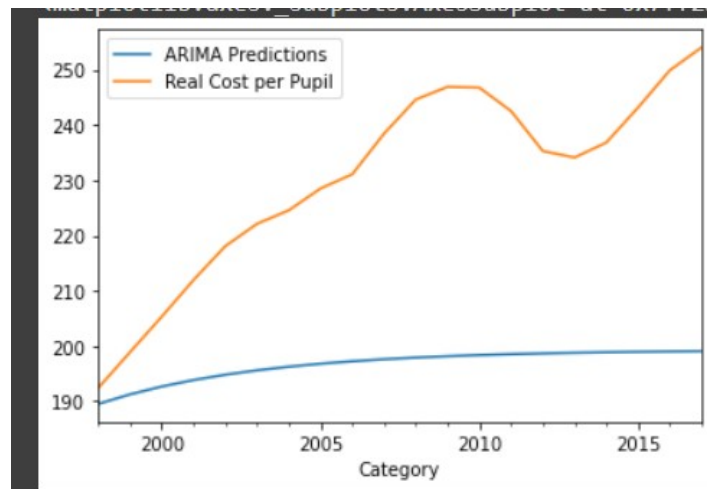
$$w + b \leq -1 \text{ untuk } y_i = -1 \quad (3)$$



Gambar 2. Alur Penelitian

HASIL DAN PEMBAHASAN

1. Hasil Eksperimen ARIMA



Gambar 3. Hasil Dari Visualisasi Pada *Real Cost* Per Pupil Yang Diperoleh Dari Perhitungan Arima Dengan Menggunakan Python

Dari hasil visualisasi data dapat diketahui MAPE lebih kecil dari MAE. Metode peramalan mengandung bias sehingga dibandingkan dengan nilai deret asli akan jauh. Berdasarkan deklarasi data di awal terdapat salah satu library yang ditambahkan yaitu seaborn. Seaborn merupakan library untuk membuat grafik dan statistik dengan menggunakan Python (Waskom, 2021). Seaborn dibangun berdasarkan library matplotlib yang terintegrasi datanya secara terstruktur pada panda. Seaborn juga bertujuan agar visualisasi data yang ditampilkan meningkatkan dalam pemahaman data. Berdasarkan hal tersebut library Seaborn dapat diketahui berorientasi pada fungsi plotting berdasarkan dataset yang diproses pada dataframe dan array yang melakukan semantik dan agregasi statistik untuk menghasilkan grafik yang informatif.

Tabel 1. Hasil SARIMAX

	coef	std error	z	P> z	[0.025	0.975]
ar.L1	0.8237	0.120	6.858	0.000	0.588	1.059
ma.L1	0.1072	0.230	0.467	0.641	-0.343	0.558
sigma2	5.3043	1.785	2.971	0.003	1.805	8.804
Prob(Q) = 0.63			Prob (JB) = 0.76			
Heteroskedascity (H) = 1.18			Skew = 0.06			

Dari tabel 1 terdapat hasil SARIMAX yang merupakan rangkaian perhitungan peramalan deret waktu dengan ARIMA untuk analisis sehingga dapat menghasilkan hasil penelitian yang lebih komprehensif. Pada bagian ini digunakan untuk memprediksi jumlah biaya nyata untuk peramalan sampai akhir tahun 2018.

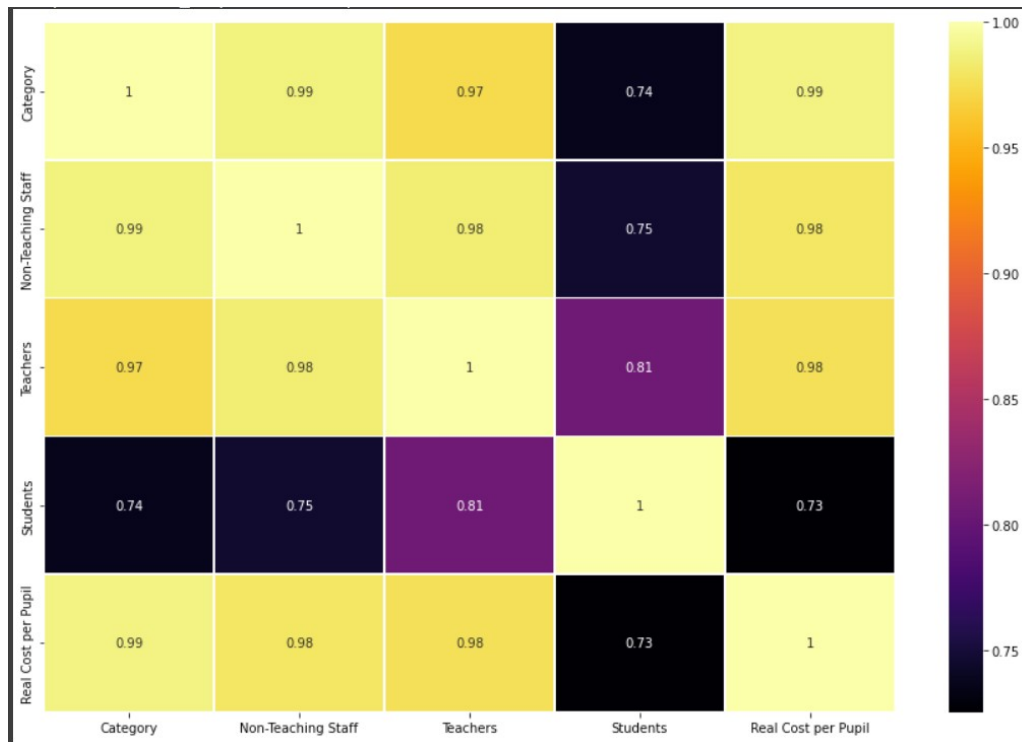
Tabel 2. Hasil Perhitungan ARIMA

MAPE	0.142
MAE	33.664
MPE	-0.142
MSE	1341.386
RMSE	36.625

Sebelum menentukan MAPE (*Mean Percentage Absolute Error*), MAE (*Mean Absolute Error*), MPE (*Mean Percentage Error*), MSE (*Mean Square Error*), dan RMSE (*Root Mean Square Error*) hal pertama yang dilakukan adalah mendeklarasikan matriks dengan mengimpor nilai tengah sebagai acuan menentukan nilai-nilai tersebut. Berdasarkan percobaan yang dilakukan mendapatkan 0, 142 untuk MAPE. Kemudian untuk MAE mendapatkan nilai 33, 664. Untuk MPE menghasilkan -0,142. Sedangkan untuk MSEnya mendapatkan 1341, 386. Dan pada RMSE menghasilkan 36.625 sesuai dengan tabel 2.

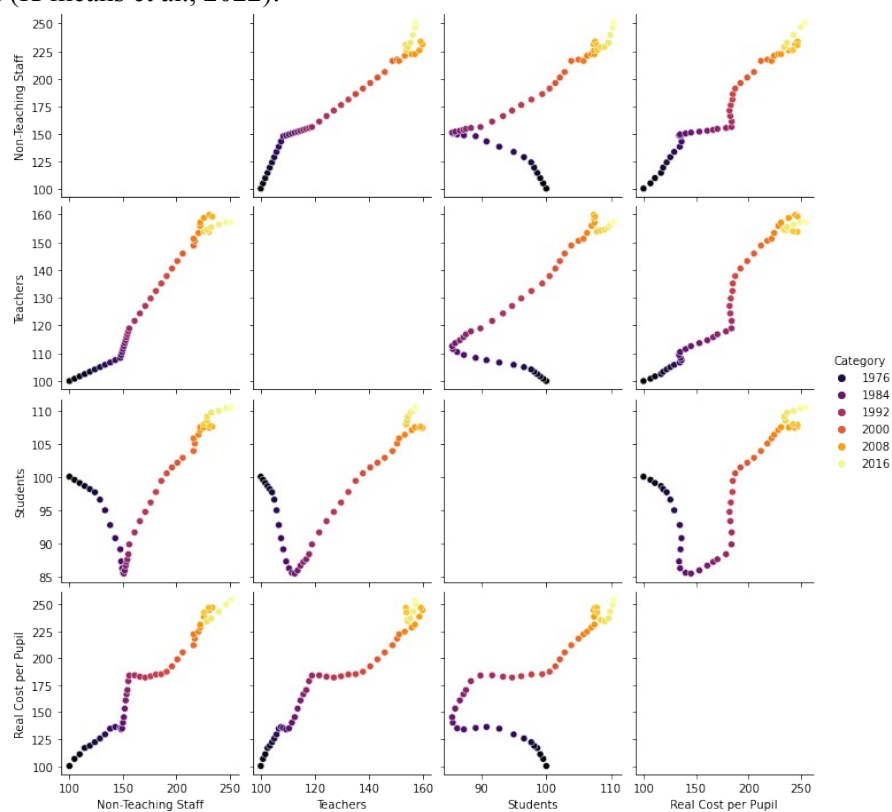
2. Hasil Eksperimen SVM

Vector Machine dikatakan sebagai salah satu model terbaik dari bahasa mesin. Di sini, dalam model ini digunakan tiga jenis kernel dan menentukan mana yang lebih baik dengan membandingkan satu sama lain dan model Arima.



Gambar 4. Pair Plot

Pemetaan semantik dilakukan untuk memudahkan dalam visualisasi data ke tahap berikutnya (K-means et al., 2022).



Gambar 5. Plot Marjinal

Menambahkan pemetaan semantik dan mengubah plot marjinal default menjadi perkiraan kepadatan kernel berlapis (KDE). Untuk memvisualisasikan data tersebut, maka perlu memulai seaborn terlebih dahulu. Dari visualisasi data tersebut dapat merepresentasikan berbagai data dalam satu gambar. Fungsi *pairplot* akan menggabungkan gambar gabungan dan marginal. Fungsi ini tidak hanya terfokus pada *relation* antar data tetapi juga memvisualisasikan setiap kombinasi secara bersama.

Tabel 3. Perhitungan SVM

MAPE	0.003
MAE	5.6000
MPE	0.0019
MSE	55.8
RMSE	7.467

SVM hasilnya dapat dilihat pada Tabel 3 dimana dengan SVM menunjukkan kinerja yang jauh lebih baik dari ketiga model dengan memiliki 0,003 MAPE dan RMSE 7,467. Dengan perolehan hasil pada MAPE pada perhitungan SVM yang lebih kecil, persentase kesalahan yang terjadi juga kecil.

KESIMPULAN

Peramalan deret waktu sangat penting untuk peramalan masa depan karena terlibat dalam proses pengambilan keputusan. Sebelumnya kebutuhan data dan informasi untuk meramalkan masa depan. Di ringkasan, langkah utama untuk peramalan adalah untuk menganalisis data deret waktu historis atau lampau untuk mengidentifikasi pola yang bisa digunakan.

Hasil percobaan yang membandingkan kinerja SVM dan ARIMA adalah perhitungan SVM yang menghasilkan kinerja lebih efektif karena menyediakan teknik alternatif. Pada perhitungan diatas, diperoleh hasil MAPE dari model Arima dan SVM yakni 0.1421 dan 0.0028. Sedangkan hasil yang diperoleh untuk RMSE dari perhitungan ARIMA adalah 36.6249, sedangkan pada SVM diperoleh hasil RMSE 7.4699. Hal ini menunjukkan bahwa SVM lebih unggul dari model ARIMA.

DAFTAR PUSTAKA

- Karaçor, Zeynep., Güvenek, Burcu., Ekinci, Esra and Konya , Sevilay. 2018. PANEL ESTIMATION FOR THE RELATIONSHIP BETWEEN EDUCATION EXPENDITURE AND ECONOMIC GROWTH FOR OECD COUNTRIES. *Forum Scientiae Oeconomia*. 6(2), http://dx.doi.org/10.23762/FSO_vOL6NO2_18_1
- K-means, K. M. A., Kurniadewi, H., Hakim, R. A., Jajuli, M., & Jaman, J. H. (2022). *Pemetaan UMKM dalam Upaya Pengentasan Kemiskinan dan Penyerapan Tenaga*. 6(2), 113–119.
- Kurniawan, A. F., Pane, S. F., & Awangga, R. M. (2021). *Prediksi Jumlah Penjualan Rumah di Bojongsong ditengah Pandemi Covid-19 dengan Metode ARIMA*. 5, 1479–1487. <https://doi.org/10.30865/mib.v5i4.3121>
- Lestari, I., Sumargo, B., & Ladayya, F. (2022). *Pemodelan ARIMA Intervensi untuk Meramalkan Harga Minyak Mentah Dunia*. 22(2), 133–146.
- Mapes, K., & Topham, K. (2020). *Exploring Data Visualization with Flourish*.
- Pratiwi, D., & Hadijati, M. (2021). *Prediction of gold price with ARIMA and SVM*. <https://doi.org/10.1088/1742-6596/1767/1/012022>

- Puspita, V. (2017). *Time Series untuk Peramalan Konsumsi Listrik Menggunakan Metode Support Vector Machine (SVM)*. 3(1), 261–264.
- Statistika, P. S., Matematika, F., Ilmu, D. A. N., Alam, P., & Indonesia, U. I. (2022). *ANALISIS SENTIMEN UU CIPTA KERJA MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM)*.
- Waskom, M. L. (2021). *seaborn: statistical data visualization Statement of need*. 6, 1–4. <https://doi.org/10.21105/joss.03021>
- Witanti, A., D. (2022). ANALISIS SENTIMEN MASYARAKAT TERHADAP VAKSINASI COVID-19 PADA MEDIA SOSIAL TWITTER MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM). *Jurnal Sistem Informasi Dan Informatika (Simika)* P-ISSN, 5, 2622–6901.