Classifying Indonesian Poverty Through 2014 - 2015 using Support Vector Machines and Logistic Regression

Kevin Christian

Statistics Department

*School of Computer Science*   
*Jakarta, 11530*  
[*kevin.christian001@binus.ac.id*](mailto:kevin.christian001@binus.ac.id)

Reyza Farzan Rahmatsyah

Statistics Department  
School of Computer Science  
Jakarta, 11530  
[reyza.rahmatsyad@binus.ac.id](mailto:reyza.rahmatsyad@binus.ac.id)

Zaphenath Paneah Joseph Irawan

Statistics Department  
School of Computer Science  
Jakarta, 11530  
[zaphenath.irawan@binus.ac.id](mailto:zaphenath.irawan@binus.ac.id)

*Abstract*— Poverty is a complex problem that should be tackled in a way that is systematic and structured. Predicting said poverty is a way of helping people correctly identify the underlying connections between aspects that influence poverty. This Research uses the Indonesian Family life survey as a dataset and uses the Logistic Regression and Support Vector Machines. During the Process of Data Preprocessing, this paper uses Kmeans as a way to generate the classes in the dataset. This Research concluded that the SVM model is the better model because of its robustness against the outliers within the dataset.

Keywords—Poverty, Support Vector Machines, Logistic Regression

Abstrak— Kemiskinan merupakan masalah kompleks yang perlu ditangani dengan cara yang sistematis dan terstruktur. Memprediksi kemiskinan merupakan salah satu cara untuk membantu institusi pemerintahan mengidentifikasi dengan benar hubungan mendasar antara aspek-aspek yang memengaruhi kemiskinan. Penelitian ini menggunakan sebagai dataset “Indonesian Family life survey” dan menggunakan Regresi Logistik serta Mesin Vektor Pendukung (Support Vector Machines/SVM) sebagai modelnya. Selama proses Pre-Processing data, penelitian ini menggunakan juga metode K-means untuk menghasilkan kelas-kelas dalam dataset. Penelitian ini menyimpulkan bahwa model SVM adalah model yang lebih baik karena ketangguhannya terhadap outlier dalam dataset tersebut.

Keywords—Kemiskinan, Support Vector Machines, Logistic Regression

# Pendaluhuan

Kemiskinan adalah masalah kompleks yang masih menjadi tantangan besar di berbagai negara di seluruh dunia. Untuk memahami dan mengatasi masalah ini, penting bagi kita untuk menganalisis faktor-faktor yang mempengaruhinya. Penelitian ini bertujuan untuk mengidentifikasi dan menganalisis faktor-faktor utama yang berperan dan mempengaruhi tingkat kemiskinan.

Kemiskinan tidak hanya dipengaruhi oleh satu faktor tunggal, tetapi merupakan hasil dari interaksi berbagai faktor ekonomi, sosial, dan politik. Dalam penelitian ini, kami akan meneliti berbagai faktor yang telah diidentifikasi oleh peneliti sebelumnya, serta faktor-faktor baru yang mungkin muncul sebagai hasil dari perkembangan sosial dan ekonomi. Dengan menganalisis faktor-faktor ini, kita akan dapat memperolah pemahaman yang lebih baik tentang akar permasalahan kemiskinan.

Di era big data, ketersediaan data sosial-ekonomi menjadi peluang untuk bisa memahami dan mengatasi masalah kemiskinan. Model statistik tradisional telah digunakan di bidang ini, tetapi masih kurang efektif ketika menangani data yang kompleks. Karena itu, ada kebutuhan yang semakin meningkat untuk menggunakan pendekatan yang lebih canggih untuk mempelajari kemiskinan. Dalam hal ini, machine learning menunjukkan potensi yang menjanjikan dengan kemampuannya untuk belajar dan membuat prediksi berdasarkan data.

Penelitian ini akan menggunakan metode Support Vector Machine (SVM) dan k-Nearest Neighbors (k-NN). SVM merupakan algoritma klasifikasi yang kuat dan fleksibel, dapat menangani data berdimensi tinggi dan mampu menemukan batas keputusan yang kompleks. Sementara k-NN, dapat mengklasifikasikan titik data baru berdasarkan kedekatannya dengan yang sudah ada.

Kami percaya bahwa dengan menggabungkan teknik machine learning ini dengan data yang luas, kami akan dapat menemukan pengetahuan berharga mengenai pola kemiskinan, sehingga dapat membantu mengembangkan strategi untuk mengurangi kemiskinan.

# Penelitian Serupa

Penelitian mengenai kemiskinan pernah dilakukan sebelumnya, dengan berbagai metode machine learning. Penelitian pertama [5] dilakukan oleh Lutfia Nazila, Alan Prahutama, dan Arief Rachman Hakim mengenai klasifikasi status kemiskinan rumah tangga, mereka menggunakan metode Support Vector Machine dan Clasification and Regression Trees (CART). Penelitian tersebut menggunakan data dari Survei Sosial Ekonomi Nasional (SUSENAS), survei tersebut dilaksakan oleh Badan Pusat Statistik (BPS) provinsi Jawa Tengah pada 2018. Hasil dari penelitian tersebut adalah metode SVM memiliki ketepatan sebesar 89.31%. Metode tersebut bekerja dengan baik dalam memprediksi kelas mayoritas, namun tidak bekerja dengan baik untuk memprediksi kelas minoritas.

Penelitian kedua [1] dilakukan oleh Ni Putu Hendayanti dan Maulinda Nurhidayati, mereka menggunakan regresi logistik biner untuk menentukan ketepatan klasifikasi tingkat kedalaman kemiskinan provinsi-provinsi di Indonesia. Pada penelitian tersebut, faktor yang paling mempengaruhi kemiskinan adalah rata-rata lama sekolah dan harapan lama sekolah.

Penelitian ketiga [8] dilakukan pada tahun 2019, Rengganis Woro Maharsi dan Sulistyo Hadi membadingkan metode Support Vector Machine (SVM) dan regresi logistik biner untuk mengklasifikasi status kemiskinan rumah tangga di provinsi Lampung. Mereka menggunakan data sekunder dari Susenas 2019 dimana data dibagi menjadi data training dan data testing dengan rasio 80:20. Hasil dari penelitian tersebut adalah Support Vector Machine memberikan hasil yang lebih baik daripada regresi logistik biner. Akan tetapi SVM memberikan nilai APER yang cukup tinggi yaitu 15,34%.

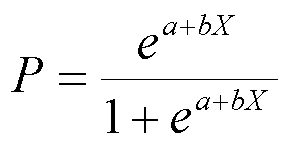
Penelitian keempat [6], dimana Mustafa A. Rahman dari Australia meneliti karakteristik dari kemiskinan rumah tangga dengan regresi logistik. Berdasarkan penelitiannya, faktor-faktor yang sangat mempengaruhi kemiskinan adalah rendahnya tingkat pendidikan, adanya disabilitas pada anggota rumah tangga, dan anggota rumah tangga yang bekerja pada bidang manufaktur.

# Metodologi

Support Vector Machines (SVM) adalah metode machine learning yang digunakan untuk tugas klasifikasi dan regresi. SVM efektif dalam mengklasifikasikan data berdimensi tinggi dan juga bisa sangat efisien ketika datanya memiliki banyak fitur. Prinsip utama dari SVM adalah menemukan hyperplane dalam ruang N-dimensi (N adalah jumlah fitur) yang mengklasifikasikan data dengan cara paling optimal. SVM mencari hyperplane dengan margin terbesar, yaitu jarak terbesar antara batas keputusan (hyperplane) dan titik data terdekat dari setiap kelas. Titik data ini disebut Support Vectors, karena mereka mendukung atau menentukan posisi dan orientasi dari hyperplane.

SVM mencoba untuk memaksimalkan margin ini untuk mendapatkan model yang paling umum dan memiliki bias-variance trade-off yang baik. Jika data tidak dapat dipisahkan secara linier, SVM menggunakan teknik yang disebut kernel trick. Kernel trick memungkinkan SVM untuk bekerja dalam ruang dimensi yang lebih tinggi, sehingga memungkinkan pemisahan non-linier antara kelas. Kernel dapat dipandang sebagai fungsi yang mengubah data masukan menjadi bentuk yang lebih mudah dipisahkan secara linier. Meski SVM sangat kuat dan fleksibel, model ini juga membutuhkan penyetelan parameter yang cermat (seperti pilihan kernel dan parameter kernel) dan penanganan hati-hati terhadap overfitting, terutama dalam kasus di mana dimensi fitur sangat tinggi.

Support Vector Machine (SVM) dan Logistic Regression adalah dua metode yang digunakan dalam machine learning untuk melakukan klasifikasi data. Meskipun keduanya digunakan untuk tujuan yang serupa, mereka memiliki perbedaan dalam hal pendekatan dan bagaimana mereka bekerja. Perbedaan utama antara dua teknik ini adalah cara mereka mencari batas keputusan: Logistic Regression mencoba meminimalkan kesalahan dan memaksimalkan kemungkinan data, sementara SVM mencoba mencari margin maksimum yang memisahkan kelas. SVM juga lebih cenderung untuk lebih baik dalam menggeneralisasi model dalam kasus di mana terdapat banyak dimensi (feature) dan dataset berukuran kecil, sementara Logistic Regression lebih cocok untuk kasus di mana data lebih linear dan proses interpretasi model lebih penting.



K-means clustering adalah salah satu algoritma pengelompokan (clustering) yang paling populer dalam machine learning dan data mining. Algoritma ini berusaha mempartisi dataset menjadi sejumlah 'k' kelompok atau cluster. Proses ini melibatkan pemetaan setiap titik data ke cluster terdekat. Algoritma K-means biasanya berkinerja baik untuk dataset yang memiliki cluster yang jelas dan simetris. Namun, K-means mungkin tidak efektif jika ada cluster yang berukuran sangat berbeda atau berbentuk tidak biasa. K-means sangat dipengaruhi oleh pemilihan centroid awal. Jika centroid awal dipilih secara tidak tepat, K-means mungkin tidak dapat menemukan pengelompokan yang optimal.

## Dataset

Dataset yang berjudul “ *INDONESIAN FAMILY LIFE SURVEY 5*“ data ini dibuat oleh John Strauss yang lalu diunggah di rand.org. (<https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/ifls5.html>). Dataset ini merupakan survey kehidupan berkeluarga di Indonesia dari tahun 2014 sampai 2015. Survey ini mencakup daerah barat dan tengah Indonesia.

## Variables

1. Penjelasan Variable dataset

|  |  |  |
| --- | --- | --- |
| Nama | Definisi | Keterangan |
| ID | Identifikasi dari responden survey | - |
| School Cost | Biaya sekolah dari anak responden | - |
| School Supplies | Biaya alat dan bahan sekolah anak responden | - |
| Transportation | Biaya Transprotasi Responden | - |
| Program Type | Tipe program bantuan yang didapatkan responden | 1. : Program Kompensasi Penuruan Subsidi B 2. : Program Bantuan Tunai Bersyarat (Kelua)   : Bantuan Langsung Tunai Bersyarat (UCT) |
| Amount Received | Nominal Uang yang didapatkan dari *program Type* | - |
| Income Source | Sumber dasri penghasilan responden | 0 : Retirement  1 : Government Scholarship  2 : Private Scholarship  3 : Insurance Money  4 : Winnings / Lottery |
| Amount Received |  |  |
| Received Income | Apakah responden mendapatkan penghasilan dari *income source* | 0 : Did Not Receive  1 : Given  2 : Don’t Know |
| Annual Income | Pendapatan Pertahun |  |
| Poor | Tingkat Pendapatan | 0 : Low Income  1 : Mid Income  2. High Income |

## Model Evaluation

1. Confusion matrix [16]

True Positive : Digunakan untuk mendefinisikan jumlah data yang positif dan terklasifikasi dengan benarTrue Negative : Digunakan untuk mendefinisikan jumlah data yang negatif dan terklasifikasi dengan benarFalse Positive : Digunakan untuk mendefinisikan jumlah data yang positif dan terklasifikasi dengan salahFalse Negative : Digunakan untuk mendefinisikan jumlah data yang negative dan terklasifikasi dengan salah

1. Classification Report [12]

|  |  |
| --- | --- |
| True Positive ( TP ) | False Positive ( FP ) |
| False Negative ( FN ) | True Negative (TN ) |

**Accuracy** : rasio benar dari prediksi yang sudah dilakukan (positif dan negative)

TP + TN / (TP + TN +FP + FN)

**Precision** : rasio prediksi yang benar dibandingkan dengan keseluruhan yang diprediksi benar

TP / (TP + FP)

**Recall** : rasio yang diprediksi dibandingkan dengan keseluruhan data yang sesuai prediksi

TP / ( TP + FN )

**F1-Score** : perbandingan antara *precision* dan *recall*

2\*Precision \* Recall / Precision + Recall

# hasil

## Data Pre-Processing

### Data Processing

Proses Data Cleaning pada paper ini melibatkan proses pemilihan kolom yang dapat digunakan dalam memprediksi kemiskinan. Kolom ini juga diganti namanya agar dapat lebih mudah dimengerti. Karena dataset kita merupakan dataset yang tersebar dalam banyak file, kita melakukan proses data cleaning sebanyak tiga kali dengan proses terakhir adalah menggabungkan 3 dataset yang sudeh tersebar ini. Gabungan 3 dataset ini kami namakan *final.*  NA yang terdapat pada dataset *final* ini dihilangkan.

### Outlier Handling

Model yang digunakan cukup kuat terhadap Outlier sehingga proses ini dapat kita abaikan.

### Clustering

Karena Dataset yang dihasilkan proses Data Processing tidak memiliki variabel target, maka diperlukannya proses clustering agar dapat menghasilkan variabel tersebut. Metode yang digunakan untuk *Clustering* adalah *KMeans.*

A picture containing line, screenshot

Description automatically generated

*Gambar 4.1 Hasil KMeans Clustering*

Dari gambar 4.1, kita dapat mengambil kesimpulan bahwa number of cluster yang ideal untuk digunakan dalam mengelompokkan dataset *final* adalah lima. Hal ini disebakan perbedaan inertiaantara lima dan enam relatif dekat.

A picture containing diagram, circle, screenshot, graphics

Description automatically generated

*Gambar 4.2 Hasil Clustering*

Gambar 4.2 menggambarkan hasil dari clustering *Kmeans* yang menggunakan jumlah tetangga sebanyak lima. Intrepretasi yang bisa kita gunakan adalah :

1. Intrepretasi Clustering

|  |  |
| --- | --- |
| Cluster | Intrepretasi |
| 0 | Low Income |
| 1 | Lower-Middle Income |
| 2 | Middle Income |
| 3 | Upper-Middle Income |
| 4 | High Income |

A screenshot of a game

Description automatically generated with medium confidence

*Gambar 4.3 Heatmap dari data*

## Support Vector Machines

### With Outliers

1. Train Test Comparison for SVM

|  |  |  |  |
| --- | --- | --- | --- |
| Train/Test | **precision** | **recaII** | **fl- s core** |
| **0** | 0 . 21 | 0.62 | 0.32 |
| 0.19 | 0.58 | 0.28 |
| **1** | 1 | 0.8 | 0.89 |
| 1 | 0.79 | 0.88 |
| **2** | 0.34 | 1 | 0.51 |
| 0.41 | 1 | 0.58 |
| **3** | 0.06 | 1 | 0.11 |
| 0.05 | 1 | 0.09 |
| **4** | 1 | 1 | 1 |
| 1 | 1 | 1 |

Accuracy Train : 0.79

Accuracy Test : 0.78

Dari tabel III, model yang tidak menggunakan outlier konsisten ketika menggunakan *Training* ataupun *Testing* Dataset.

### Witthout Outliers

Untuk menghasilkan SVM yang tidak menggunakan Outlier, kita harus menghilangkan outliernya. Akan tetapi, proses penghilangan dari outlier ini menghasilkan data yang variabel targetnya hanya mempunyai 1 nilai saja. Dengan Demikian, SVC yang dapat digunakan adalah SVM yang menggunakan outlier.

### With HyperParameter Tuning

Hyperparameter Tuning menggunakan GridSearch dengan nilai c : 1 dan nilai : 0.001 yang ideal. Berikut merupakan hasil dari GridSearch

1. Train Test Comparison for SVM After Gridearch

|  |  |  |  |
| --- | --- | --- | --- |
| Train/Test | **precision** | **recaII** | **fl- s core** |
| **0** | 1 | 1 | 1 |
| 0.85 | 1 | 0.92 |
| **1** | 1 | 1 | 1 |
| 1 | 0. 95 | 0.97 |
| **2** | 1 | 1 | 1 |
| 1 | 1 | 1 |
| **3** | 1 | 1 | 1 |
| 0.91 | 1 | 0.95 |
| **4** | 1 | 1 | 1 |
| 1 | 1 | 1 |

Accuracy Train : 1.00

Accuracy Test : 0.96

## Logistic Regression

1. Train Test Comparison Logistic Regression

|  |  |  |  |
| --- | --- | --- | --- |
| Train/Test | **precision** | **recaII** | **fl- s core** |
| **0** | 0.59 | 0. 70 | 0.64 |
| 0.61 | 0.73 | 0.67 |
| **1** | 0.93 | 0.85 | 0.89 |
| 0.94 | 0.84 | 0.89 |
| **2** | 0. 62 | 0.82 | 0.7 |
| 0.64 | 0. 93 | 0.76 |
| **3** | 0. 24 | 0.4 | 0.3 |
| 0.31 | 0. 55 | 0.39 |
| **4** | 0.78 | 1 | 0.88 |
| 0.83 | 1 | 0.91 |

Accuracy Train : 0.81

Accuracy Test : 0.81

# Analisis dan diskusi

Dari hasil perbandingan data training dan data testing, bab ini akan membahas mengenai hasil dan artinya untuk setiap model.

## Support Vector Machines Before and After GridSearch

Support Vector Machine yang ideal untuk prediksi model adalah yang sudah melalui proses GridSearch. Precision dan Recall SVM setelah melalui GridSearch dapat memprediksi kategori 0 (*Low Income*) dengan relatif benar. Akurasi dari data testing juga lebih superior yang setelah melakukan Gridsearch.

## Support Vector Machines and Logistic Regression

Jika dibandingkan SVM dan Logistic Regression akurasi data testing SVM lebih bagus. SVM juga lebih ideal dalam precision dan recall untuk prediksi kelas 0.

# Kesimpulan

Poverty merupakan aspek kehidupan masyarakat yang harus ditangani dengan proses yang sistematis dan terskturkur. Prediksi Kemiskinan akan sangat membantu proses pengurangan kemiskinan. Hasil analisis menunjukkan bahwa model SVM dengan Hyperparameter c = 1 dan = 1e-03 merupakan model yang ideal.SVM sendiri merupakan model yang kuat terhadap outlier sehingga prediksi kemiskinan dapat melibatkan lebih banyak variable. Banyak variable sendiri juga ideal karena kemiskinan merupakan masalah yang dapat dihadapi dari berbagai sisi.

# References

1. Hendayanti, N. P. N., & Nurhidayati, M. (2020). Regresi Logistik Biner dalam Penentuan Ketepatan Klasifikasi Tingkat Kedalaman Kemiskinan Provinsi-Provinsi di Indonesia. *Sainstek : Jurnal Sains Dan Teknologi*, *12*(2), 63. <https://doi.org/10.31958/js.v12i2.2483>
2. Hilbe, J. M. (2009). Logistic Regression Models. United Kingdom: CRC Press.
3. J. A. Hartigan, M. A. W. (2012). Algorithm AS 136 A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society Series B Methodological*, *28*(1), 100–108.
4. M. Sokolova dan G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Inf. Process. Manag., vol. 45, no. 4, hal. 427–437, 2009.
5. Nuzula, L., Prahutama, A., & Hakim, A. R. (2020). KLASIFIKASI STATUS KEMISKINAN RUMAH TANGGA DENGAN METODE SUPPORT VECTOR MACHINES (SVM) DAN CLASSIFICATION AND REGRESSION TREES (CART) MENGGUNAKAN GUI R (Studi Kasus di Kabupaten Wonosobo Tahun 2018). Jurnal Gaussian, 9(4), 525-534. <https://doi.org/10.14710/j.gauss.v9i4.29449>
6. Rahman, M. A. (2013). Household Characteristics And Poverty: A Logistic Regression Analysis. *The Journal of Developing Areas*, *47*(1), 303–317. <https://doi.org/10.1353/jda.2013.0001>
7. Steinwart, I., Christmann, A. (2008). Support Vector Machines. Germany: Springer New York.
8. Woro Maharsi, R., Sulistyo Hadi, dan, & kunci, K. (2022). Pemilihan Metode Terbaik Support Vector Machine (SVM) Dan Regresi Logistik Biner Untuk Klasifikasi Status Kemiskinan Rumah Tangga Di Provinsi Lampung Tahun 2019. *Jurnal Siger Matematika*, *03*(02).