JOSTEN
Journal of Science, Technology and Neurology
Open Access

Analysis of the Accuracy of *CatBoost*, *Random Forest*, and SVM Models in Poverty Level Classification in Indonesia

Pasha Khatami Hasibuan[1*] Rezkya Nurdiana[2] Atika Pratiwi Harahap[3]

[1]Department of Electronic and Informatics Engineering, Universitas Negeri Yogyakarta, Sleman, Indonesia
[2]Information Systems Study Program, Universitas Malikussaleh, Aceh, Indonesia
Email: pashakhatamihsb.@gmail.com

Article Information:
Received: 27 October 2025
Revised: 29 December 2025
Accepted: 01 January 2025
Published: 02 January 2025

https://doi.org

Abstract

Introduction: Poverty alleviation is a key focus in Indonesia's national development agenda, but the effectiveness of social assistance distribution is often hampered by exclusion errors and slow data updates

Objective: This study aims to conduct a rigorous comparative analysis of the performance of three leading machine learning algorithms, namely CatBoost, Random Forest, and Support Vector Machine (SVM), in classifying poverty levels in districts/cities in Indonesia to improve the accuracy of policy targeting

Methods: Using a comprehensive dataset from the Central Statistics Agency (BPS) covering 11 socio-economic indicators from 514 administrative regions, this study applied standard data pre-processing techniques, data sharing using stratified sampling (80% trained, 20% tested), and model validation through a 5-Fold Cross-Validation scheme to ensure consistency of results.

Results: The results of the experiment showed that the CatBoost model obtained better predictive performance with an Accuracy of 94.17%, F1-Macro of 86.79%, MCC of 73.90%, and an AUC-ROC of 95.97%, beating the Random Forest model in terms of generalization (Accuracy of 94.17%) and SVM (Accuracy of 86.41%). The main scientific findings from the feature importance analysis show that Average School Age, Expenditure per Capita, and Access to Decent Sanitation are the three most significant factors that affect the poverty conditions of an area.

Conclusion: This study shows that algorithms that use gradient boosting (CatBoost) are more efficient and resilient than bagging or kernel-based methods in overcoming the heterogeneity of Indonesian demographic data. The results of this study encourage the government to implement a data-based approach in setting program targets, with an emphasis on intervention on improving the quality of human resources and basic infrastructure.

Keywords: Poverty Rate, CatBoost, Random Forest, SVM, Indonesia, Classification, Analysis

Introduction

Economic underdevelopment continues to be one of the main obstacles in Indonesia's national development plan. Data from the Central Statistics Agency (BPS) as of March 2024 shows that the national poverty rate reached 9.03%, which has an impact on around 25.22 million people [1]. Although this data reflects a downward trend, the Indonesian government still targets the elimination of extreme poverty to 0% by 2024 based on Presidential Instruction Number 4 of 2022 [2]. This ambitious goal is in line with the global commitment of Sustainable Development Goals (SDGs) number 1, namely "No Poverty" [3]. These ambitious goals

emphasize the need for appropriate and immediate policies.

Although budget allocations for social protection programs continue to increase, their effectiveness is often eroded by fundamental problems in targeting, namely inclusion error and exclusion error [4]. This error stems from the limitations of traditional poverty identification methods that are expensive, time-consuming, and often result in data that is no longer relevant. In response, the use of Machine Learning (ML) presents a different paradigm. Machine Learning algorithms excel at recognizing complex patterns in large datasets, exceeding traditional statistical models [5], [6]. A number of previous studies have shown the success of the application of Machine Learning. Jean [6] utilizing satellite data, while other studies examined the application of Neural Networks in predicting poverty but often encountered overfitting issues on limited datasets [7]. In Indonesia, Utomo's research [8] and Setyawan [9] implement KNN and Decision Tree. However, Handayani's research [10] and Saputra [11] noted the importance of special treatment of imbalanced datasets that often arise in cases of poverty, a crucial issue that was also addressed in a comprehensive survey by Haixiang [12].

Based on these gaps, the study provides scientific innovation by directly comparing three advanced algorithms: CatBoost, Random Forest, and Support Vector Machine (SVM). CatBoost was chosen because of its speciality in managing categorical features [13], [14], which is often the weakness of competitors' algorithms like XGBoost in terms of *training speed* [15]. Random Forest was chosen for the stability *of its* ensemble [16], [17], and SVM due to its effectiveness in high-dimensional spaces [18], [19].

The application of Machine Learning in poverty studies is becoming very urgent due to the limitations of conventional statistical methods in capturing the complexity of dynamic socio-economic data. Poverty is a multi-dimensional problem involving non-linear relationships between variables that often escape the usual linear regression modeling. With intelligent computing capabilities, machine learning offers a solution to minimize exclusion errors more precisely through complex data patterns, ensuring social assistance can be delivered with much higher target accuracy.

The selection of the three algorithms in this study is based on the advantages of their respective characteristics in handling poverty data. CatBoost was chosen as the lead algorithm because of its superior ability to handle categorical features automatically and minimize overfitting on varied datasets. Random Forest is included for its stability through an ensemble bagging method that is resistant to data noise, while the Support Vector Machine (SVM) is used as a reliable comparator in search of the optimal hyperplane in high-dimensional spaces. The comparison of the three provides comprehensive insights into the best approach to the characteristics of Indonesia's demographic data.

Method

Types of research

This study uses a comparative quantitative approach to assess the effectiveness of machine learning algorithms in classifying poverty. The object of the study covers 514 districts/cities in Indonesia by utilizing the latest secondary data obtained from the Central Statistics Agency (BPS) [1]. The research flow begins with the pre-processing stage which includes cleaning, normalization using StandardScaler, and data splitting. The data distribution was carried out in an 80:20 ratio using the Stratified Sampling method. This method was chosen to ensure that the proportion of minority classes remained balanced in both subsets, in accordance with Kohavi's suggestion for datasets that have [20] Unequal class distribution.

$$F_t(x) = F_{t-1}(x) + \alpha h_t(x) \quad (1)$$

The final result is determined by a majority vote of all trees created, as stated in the equation:

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_B(x) \quad (2)$$

The final prediction is determined by the method of the most votes of all the resulting trees, as stated in the equation:

$$\omega \cdot x + b = 0 \quad (3)$$

Guaranteeing that the model obtains an optimal hyperplane, the optimization process is carried out by reducing the weight norm $||\omega||$ provided that each sample of data is appropriately classified on the appropriate margin side:

$$y_i(\omega \cdot x_i + b) \geq 1 \quad (4)$$

Model performance was assessed using the Accuracy, Precision, Recall, F1-Score, and AUC-ROC metrics. The use of F1-Score and AUC takes precedence over accuracy alone, as it provides a more accurate representation of unbalanced binary classification situations, as evidenced by Chicco and Jurman [21]. Validation is performed using 5-Fold Cross-Validation to ensure that the model has good generalization capabilities [22].

In addition to standard metrics, the study added an evaluation using the Matthews Correlation Coefficient (MCC) to ensure the validity of the model on a dataset that may be unbalanced. The MCC is considered one of the most honest metrics for binary classification because it effectively accounts for all four components of the confusion matrix (True Positives, True Negatives, False Positives, and False Negatives). MCC values range from -1 to +1, where +1 indicates a perfect prediction, 0 indicates a random prediction, and -1 indicates a completely false prediction.

Results and Discussion

Results

The dataset used in this study consists of 12 variables that describe multidimensional well-being indicators. Details about the predictor and target variables can be seen in Table 1. The dataset includes education, health, economy, and basic infrastructure indicators related to poverty measurement in Indonesia:

Table 1. Descriptive Statistics of Poverty Indicator Variables

Variable	Rerata (Mean)	Std. Deviation	Minimum	Maximum
Average School Length (Years)	8.44	1.63	1.42	12.83
Expenditure per Capita (Thousand Rp)	10,324.79	10,324.79	3,976.00	23,888.00
Human Development Index (HDI)	69.93	6.50	32.84	87.18
Life expectancy (years)	69.66	3.45	55.43	77.73
Access to Decent Sanitation (%)	77.20	18.58	0.00	99.97
Access to Decent Drinking Water (%)	85.14	15.70	0.00	100.00
Open Unemployment Rate (%)	5.06	2.64	0.00	13.37
Labour Force Participation Rate (%)	69.46	6.40	56.39	97.93
GDP on the basis of constant prices (Rupiah) (%)	21,964.08	47,904.92	147.49	460,081.00

Comparative evaluation indicated that all three models could perform the classification well, but there were significant performance differences in sensitivity metrics. A summary of the model's performance on the test data is shown in Table 2

Table 2. Comparison of Working Metrics of Classification Model

Model	Accuracy (Test)	F1-*macro* (Test)	MCC (Test)	AUC-ROC (Test)

CatBoost	94.17%	86.79%	73.90%	95.97%
Random Forest	94.17%	93.17%	71.70%	96.89%
SVM	86.41%	87.60%	54.91%	93.22%

Based on Table 2, the results of the experiment show a tight performance competition between boosting (CatBoost) and bagging (Random Forest) based models. Both models recorded an identical accuracy rate of 94.17% on the test data. However, in the context of poverty classification that has the characteristics of imbalanced data, accuracy alone can be biased.

Therefore, the evaluation was focused on the F1-Macro and Matthews Correlation Coefficient (MCC). In both of these metrics, CatBoost proved to be superior with an F1-Macro score of 86.79% and an MCC of 73.90%. This identifies that CatBoost is more stable in balancing precision and recall between classes than Random Forest (MCC 71.70%) and SVM (MCC 54.91%). Meanwhile, Random Forest recorded a slightly higher AUC-ROC value (96.89%), but CatBoost provides better decision boundary predictive certainty as indicated by the higher MCC value.

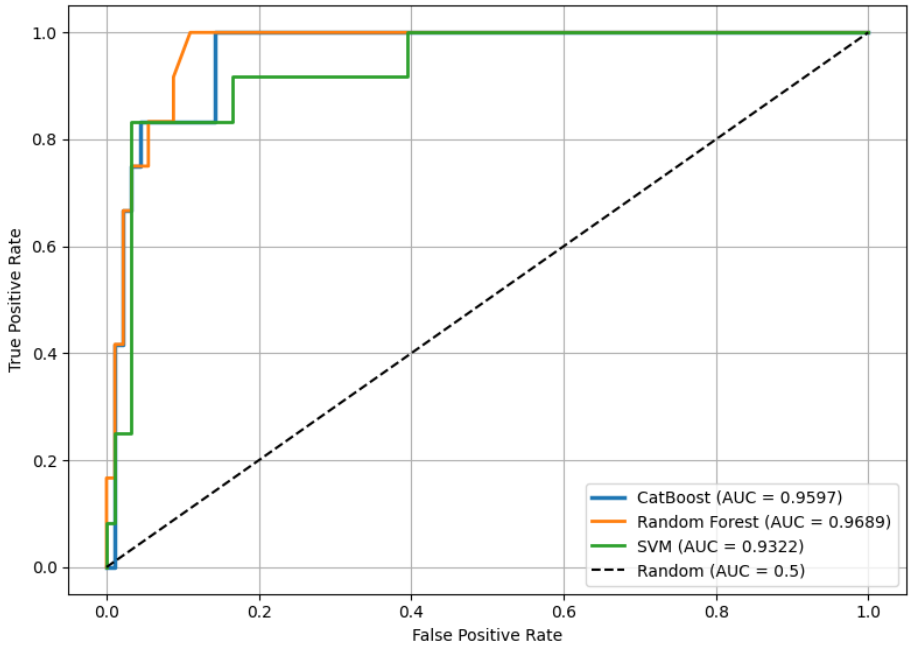
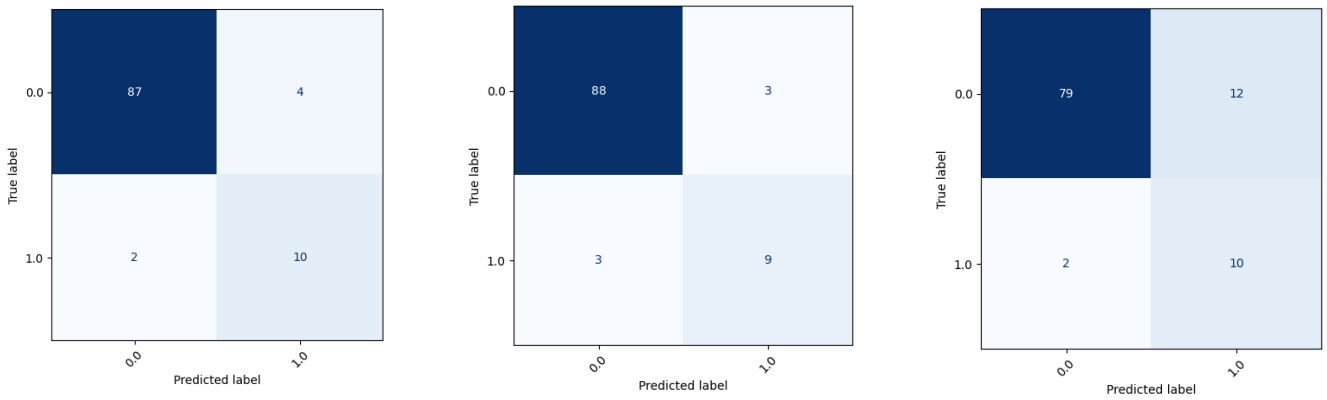


Figure 1. Comparison of the Three Model ROC-AUC Curves

Statistical validation using the Matthews Correlation Coefficient (MCC) provides deeper insight into the quality of model predictions. CatBoost's MCC value of 73.90% is included in the strong correlation category. The +0.022 MCC point advantage over Random Forest means that CatBoost has better generalization capabilities in handling False Postive and False Negative cases effectively. In contrast, a significant decrease in performance was seen in SVM which only reached an MCC of 54.91%, confirming that hyperpline-based models are less effective in handling the complexity of socio-economic features than tree-ensemble-based methods.



(a) (b) (c)

Figure 2. Confusion Matrix untuk Model (a) CatBoost, (b) Random Forest, (c) SVM

Discussion

The analysis of the experimental results showed that the algorithm based on gradient boosting (CatBoost) had a stronger performance than the bagging (Random Forest) and hyperplane-based (SVM) methods in the context of multidimensional poverty data in Indonesia. These results are in line with Setyawan's research [9] and Hancock & Khoshgoftaar [14] and a comparative study by Ben Jabeur [23] which puts gradient boosting algorithms above traditional methods in socio-economic data analysis. The main scientific results of the Feature Importance analysis show the Average School Length as the most powerful indicator. This corroborates the theory of Human Capital and the report from the World Bank [24] which states that each additional year in school has a great effect on the increase in per capita income.

Access to Adequate Sanitation is a key factor, supporting the Multidimensional Poverty Index (MPI) indicator [25]. This is also in line with the Asian Development Bank (ADB) report [26] which confirms that the lack of basic infrastructure is the main cause of structural poverty in rural Indonesia. These results show that government intervention can not only be in the form of cash assistance, but must be accompanied by investment in sanitation infrastructure

The important scientific findings of this study focus on the analysis of the factors that determine poverty. Based on the Feature Importance of the CatBoost model (Figure 3), the variables Average School Length and Expenditure per Capita were detected as the most dominant predictors.

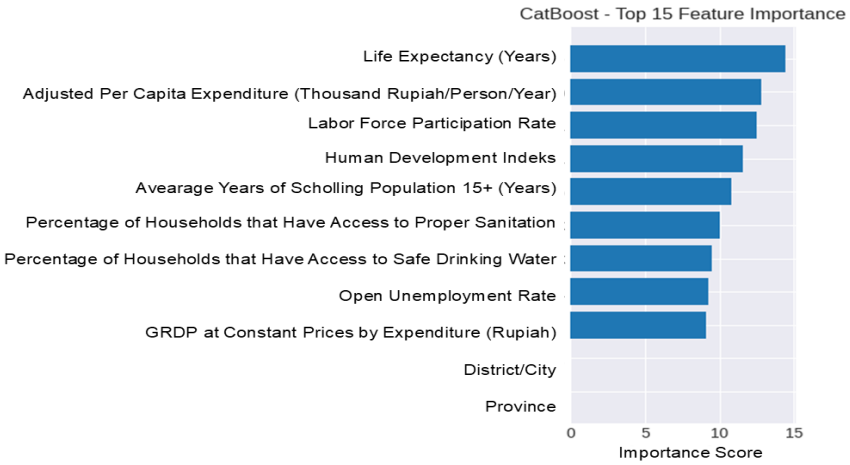


Figure 1. Feature Importance Rating of the Model

The dominance of the education variable (Average School Age) supports the *theory of Human Capital* which states that education investment has a direct impact on productivity and the ability to get out of poverty. This is consistent with Utomo's research in East Nusa Tenggara and Arbianti & Suchaina [8], [27] which found a close relationship between the education index and poverty conditions. In addition, the presence of the Decent Sanitation Access variable as the third main feature shows that poverty in Indonesia is not only related to financial aspects, but also to the lack of basic infrastructure. These findings support Alkire & Santos' argument [25] related to the Multidimensional Poverty Index (MPI), where lack of access to sanitation is a strong indicator of structural poverty.

Conclusion

This study has succeeded in meeting the objectives of the study by proving that CatBoost has the best predictive performance in classifying poverty levels in districts/cities in Indonesia, with an accuracy of 94.17% and F1-Macro of 86.79% and MCC of 73.90%. A significant scientific finding is the introduction of Average School Age, Expenditure per Capita, and Access to Decent Sanitation as the most important factors influencing poverty status. In conclusion, the application of machine learning technology, especially gradient boosting algorithms such as CatBoost, can substantially improve the accuracy and efficiency of poverty alleviation programs. The resulting policy recommendation is that the government focuses on interventions that directly improve human quality and living standards through improving basic infrastructure.

References

[1] BPS, *Profil Kemiskinan di Indonesia Maret 2024*, vol. 50/07/Th., no. 50. 2024. [Online]. Available: https://www.bps.go.id/pressrelease/2023/07/17/2016/profil-kemiskinan-di-indonesia-maret-2023.html#:~:text=Jumlah penduduk miskin pada Maret,yang sebesar 7%2C53 persen.

[2] BPK, "Instruksi Presiden Republik Indonesia Nomor 4 Tahun 2022 Tentang Percepatan Penghapusan Kemiskinan Ekstrem," *Badan Pemeriksaan Keuangan*, no. 146187, pp. 1–15, 2022, [Online]. Available: https://peraturan.bpk.go.id/Details/211477/inpres-no-4-tahun-2022

[3] U. Nations, *The Sustainable Development Goals Report*. 2023. [Online]. Available: https://unstats.un.org/sdgs/report/2023/

[4] J. Y. Kim, "Using Machine Learning to Predict Poverty Status in Costa Rican Households," *SSRN Electronic Journal*, 2021, doi: 10.2139/ssrn.3971979.

[5] C. Zeng, "Poverty Prediction Using Machine Learning Approach," no. 2020, pp. 1–6, 2022.

[6] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science (1979)*, vol. 353, no. 6301, pp. 790–794, 2016, doi: 10.1126/science.aaf7894.

[7] R. Pino-Mejías, A. Pérez-Fargallo, C. Rubio-Bellido, and J. A. Pulido-Arcas, "Artificial neural networks and linear regression prediction models for social housing allocation: Fuel Poverty Potential Risk Index," *Energy*, vol. 164, pp. 627–641, Dec. 2018, doi: 10.1016/J.ENERGY.2018.09.056.

[8] K. S. Utomo, "Perbandingan Algoritma Machine Learning Untuk Penentuan Klasifikasi Kemiskinan Multidimensi Di Provinsi Nusa Tenggara Timur," *Jurnal Statistika Terapan (ISSN 2807-6214)*, vol. 2, no. 01, pp. 36–46, 2022.

[9] A. Setyawan, A. Fitriani, E. Rilvani, U. P. Bangsa, and K. Bekasi, "Klasifikasi Kemiskinan Di Indonesia Dengan Decision Tree Menggunakan Rapidminer," vol. 3, no. 7, 2025.

[10] D. N. Handayani and S. Qutub, "Penerapan Random Forest Untuk Prediksi Dan Analisis Kemiskinan," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 405–412, 2025, doi: 10.31004/riggs.v4i2.512.

[11] T. Terttiaavini, A. Heryati, and T. S. Saputra, "Optimizing Socioeconomic Features for Poverty Prediction in South Sumatera," *TIERS Information Technology Journal*, vol. 6, no. 1, pp. 16–32, 2025, doi: 10.38043/tiers.v6i1.6244.

[12] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst Appl*, vol. 73, pp. 220–239, May 2017, doi: 10.1016/J.ESWA.2016.12.035.

[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Adv Neural Inf Process Syst*, vol. 2018-Decem, no. Section 4, pp. 6638–6648, 2018.

[14] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *Journal of Big Data 2020 7:1*, vol. 7, no. 1, pp. 94-, Nov. 2020, doi: 10.1186/S40537-020-00369-8.

[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.

[16] G. Louppe, "Understanding Random Forests: From Theory to Practice," no. July, 2015, [Online]. Available: http://arxiv.org/abs/1407.7502

[17] M. Belgiu and L. Drăgu, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, Apr. 2016, doi: 10.1016/J.ISPRSJPRS.2016.01.011.

[18] C. Cortes, V. Vapnik, and L. Saitta, "Support-vector networks," *Machine Learning 1995 20:3*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[19] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, Jul. 1999, doi: 10.1016/S0893-6080(99)00032-5.

[20] R. Kohavi and S. Edu, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," pp. 1–7, 2006, [Online]. Available: papers://5e3e5e59-48a2-47c1-b6b1-a778137d3ec1/Paper/p2015

[21] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification

evaluation," *BMC Genomics 2019 21:1*, vol. 21, no. 1, pp. 6-, Jan. 2020, doi: 10.1186/S12864-019-6413-7.

[22] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," *Encyclopedia of Database Systems*, pp. 532–538, 2009, doi: 10.1007/978-0-387-39940-9_565.

[23] S. Ben Jabeur, C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, "CatBoost model and artificial intelligence techniques for corporate failure prediction," *Technol Forecast Soc Change*, vol. 166, p. 120658, May 2021, doi: 10.1016/J.TECHFORE.2021.120658.

[24] W. Bank, "The Promise of Education in Indonesia," *The Promise of Education in Indonesia*, 2020, doi: 10.1596/34807.

[25] S. Alkire and M. E. Santos, "A Multidimensional Approach: Poverty Measurement & Beyond," *Social Indicators Research 2013 112:2*, vol. 112, no. 2, pp. 239–257, Feb. 2013, doi: 10.1007/S11205-013-0257-3.

[26] Asian Development Bank, "Indonesia , 2020 – 2024 — Emerging Stronger," no. September, pp. 2020–2024, 2020.

[27] S. Arbianti and Suchaina, "Peran Pendidikan dan Kesehatan dalam Mengurangi Ketimpangan dan Kemiskinan di Indonesia: Pendekatan Human Capital," *Jurnal Ekonomi-Qu*, vol. 15, no. 1, pp. 54–64, 2025, [Online]. Available: http://dx.doi.org/10.35448/jequ.####