Predicting Rice Production In Sumatra Island Using Linear Regression

Uun Patrio[1], Yuliska[2], Yohana Dewi Lulu Widyasari [3]
{uun22mttk@mahasiswa.pcr.ac.id[1], yuliska@pcr.ac.id[2], yohana@pcr.ac.id[3]}

Master Of Applied Computer Engineering, Politeknik Caltex Riau, Pekanbaru, Riau, Indonesia[1,2,3]

Abstract. This research explores the application of machine learning techniques to predict rice production in Sumatra Island using Linear Regression and compares it with five other algorithms: Random Forest Regression, Gradient Boosting, SVR, K-Nearest Neighbors Regression and Decision Tree Regression. After comparing the various models, the linear regression technique has the greatest R2-score (85.53%), demonstrating that it adequately explains the variation in the data. In comparison to some of the other algorithms examined, it has a Mean Absolute Error (MAE) value of 221938.68 and a Mean Squared Error (MSE) value of 176940698374.90, showing that Linear Regression tends to produce more precise predictions that are nearer to the true value. Therefore, this algorithm is considered the best choice for predicting agricultural production in Sumatra, in accordance with the research objectives. These findings highlight the potential of machine learning in improving rice production prediction models for agricultural planning and decision-making in the region.

Keywords: Prediction, Rice production, Sumatra Island, Linear Regression, Machine Learning

1 Introduction

Sumatra Island has more than 50 percent of agricultural land in each province with the most dominant major food commodity being rice, while other minor commodities are corn, peanuts, and sweet potatoes. Agricultural products in Sumatra Island are very vulnerable to climate change and its negative impacts can affect planting patterns, planting time, production, and quality of crops. Additionally, a rise in earth's temperature brought on by global warming will have an impact on weather patterns, evaporation, water runoff, soil moisture, and drastically changing climates, all of which could endanger agricultural production.

In recent years, advances in machine learning have opened up new opportunities in analyzing agricultural data and forecasting rice production more accurately. Linear regression, one of the most widely used machine learning techniques, enables us to describe the relationship between independent factors and dependent variables. The goal of this work is to forecast rice production on the island of Sumatra using machine learning techniques, particularly linear regression. By utilizing historical data of rice production, climate, and other factors that affect rice production. In order to conduct this analysis, we gathered information on rice production, climate, and other pertinent factors from dependable sources between the years 1993 and 2020. The provinces included in this study are Nanggroe Aceh Darussalam, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, and Lampung. After that, utilize linear regression and other machine learning techniques to analyze and evaluate the data in order to create a precise prediction model. Finally, use evaluation matrices like r2-score, mean squared error, and mean

ABEC 2023, September 21, Bengkalis, Indonesia

DOI 10.4108/eai.21-9-2023.2342997

absolute error to assess the model's performance. The outcomes of this study are anticipated to have a favorable effect on raising the effectiveness and production of Sumatra's agriculture sector.

2 Research Methods

In this research will go through several processes, the following is a flowchart of the research path process.

Fig. 1. Research methodology

2.1 Data Collection

Data was obtained through the kaggle.com website and the BPS website for the main food crop categories from 8 provinces on the island of Sumatra, namely Nanggroe Aceh Darussalam (NAD), North Sumatra, Riau, Jambi, South Sumatra, Bengkulu, and Lampung. The data used is from 1993 to 2020 for the rice dataset. The data contains annual production and harvest area or land area. Then the weather change data is obtained through the BMKG website for daily data on rainfall, humidity, and average temperature or average temperature from 1993 to 2020.

Table 1. Example dataset

Province	Year	Production	Land Area	Rainfall	Humidity	Average Temperature
Aceh	2016	2180754.00	293067.00	1096.00	83.32	27.12
Aceh	2017	2478922.00	294483.00	1905.90	85.57	26.51
Aceh	2018	1751996.94	329515.78	1427.80	83.98	26.48
Aceh	2019	1714437.60	310012.46	1931.40	83.90	26.65
Aceh	2020	1861567.10	317869.41	1619.20	80.82	25.41

The total data is 224 which for each province has 28 annual data. The dataset consists of the following attributes:

1. Province: Name of province
2. Year: Year of rice production
3. Production: Production results or annual harvest (tons)
4. Land Area: Agricultural area (hectares)
5. Rainfall: Average amount of rainfall in a year (millimeters)
6. Humidity: Average humidity level in a year (percentage)
7. Average Temperature: The average degree of temperature in a year (celsius)

Attributes number 1 - 4 were collected from the Indonesian Central Bureau of Statistics Database (BPS), and other attributes were collected from the Indonesian Agency for Meteorology, Climatology and Geophysics Database (BMKG)

2.2 EDA (Exploratory Data Analysis)

The purpose of EDA is to explore, understand, and analyze the data used in the study before conducting a prediction model. The following are some of the EDA stages carried out:

1. Overview Statistics Descriptive

Table 2. Statistics descriptive

	Production	Land Area	Rainfall	Humidity	Average Temperature
count	2.240000e+02	224.000000	224.000000	224.000000	224.000000
mean	1.679701e+06	374349.966920	2452.490759	80.948705	26.801964
std	1.161387e+06	232751.161987	1031.972625	4.878680	1.197041
min	4.293800e+04	63142.040000	222.500000	54.200000	22.190000
25%	5.488570e+05	146919.500000	1703.525000	78.975000	26.177500
50%	1.667773e+06	373551.500000	2315.700000	82.375000	26.730000
75%	2.436851e+06	514570.250000	3039.700000	84.000000	27.200000
max	4.881089e+06	872737.000000	5522.000000	90.600000	29.850000

The average yield in the 8 provinces over 28 years was 1679700,887 tons with the lowest yield of 42938 tons and the highest of 4881089 tons. The average area of agricultural land is 37,433,450 hectares. It is clear from the data description above that there is little difference between the mean and median values for each attribute. Consequently, it can be said that the data are regularly distributed.

2. Province-by-province overview of rice production

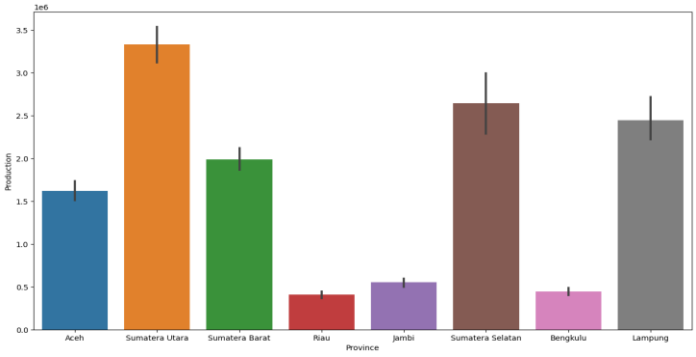

Fig. 2. Province-by-province overview of rice production

According to Figure 2, the dataset's highest crop production is found in North Sumatra, followed by South Sumatra and Lampung.

3. Distribution production of rice by year

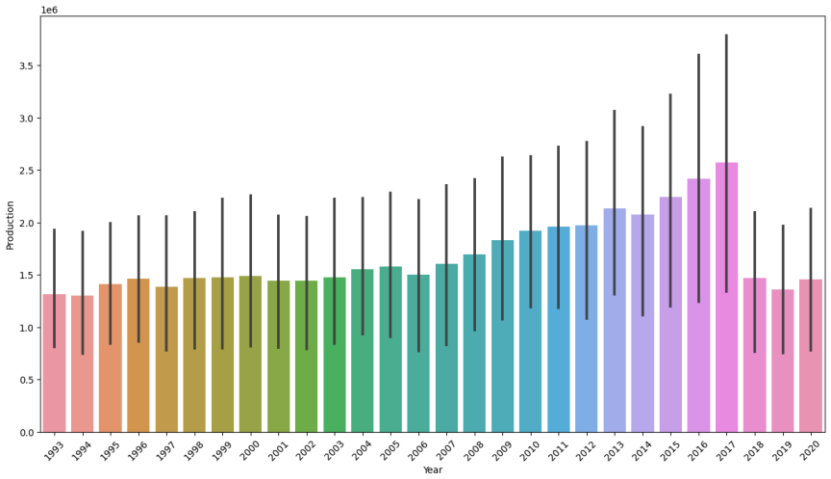

Fig. 3. Distribution production of rice by years

The year with the largest production was in 2017, but in the following years production has decreased significantly.

4. Correlation Matrix between numerical variables

Table 3. Correlation Matrix between numerical variables

	Year	Land Area	Rainfall	Humidity	Average Temperature	Production
Year	1.000000	-0.045951	-0.047645	-0.033474	0.004923	0.182527
Land Area	-0.045951	1.000000	-0.092975	-0.061121	0.115726	0.905622
Rainfall	-0.047645	-0.092975	1.000000	0.056466	-0.228699	-0.042129
Humidity	-0.033474	-0.061121	0.056466	1.000000	-0.407799	-0.052316
Average Temperature	0.004923	0.115726	-0.228699	-0.407799	1.000000	0.041160
Production	0.182527	0.905622	-0.042129	-0.052316	0.041160	1.000000

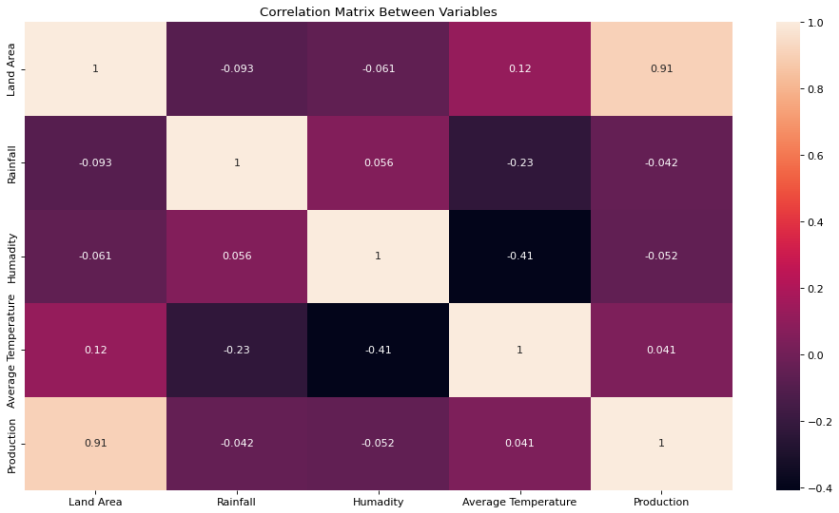

Fig. 4. Correlation Matrix between numerical variables

The degree of correlation between the independent factors and the output variable is depicted in the above image. Farmland area and average temperature are positively associated, which means that when the independent variable's value rises, so will the yield, however this increase may not be statistically significant (temperature effect). While the yield will increase when the movement of the graph is reversed, the rain and humidity variables have a negative correlation value, meaning that when both variables decrease. However, the association is not very strong.

2.3 Data Preprocessing

2.3.1 Feature Selection

We separated the data into training and testing data in this feature selection step and eliminated the year, province, and production columns. Due to the need for as many data points as feasible during sample training, data is frequently split unevenly. For training and assessment, a split ratio of 70/30 or 80/20 is employed. The initial set of data used to teach the machine learning algorithm how to learn and generate precise predictions is known as the training dataset. The training dataset makes up 70% of the dataset. How well the ML algorithm is taught with the training dataset is assessed using the testing dataset. The testing dataset makes up 30% of the whole dataset.

2.3.2 Feature Scaling

Features in this dataset have a wide range of magnitudes, units, and extents. When determining distance, high-intensity objects will be given more weight than low-intensity ones. We must level out all the features in order to get rid of this effect. Scaling can be used to accomplish this.

2.4 Modelling

Modeling is done using 6 algorithms namely:

1) Linear Regression
2) Random Forest Regression
3) Gradient Boosting
4) Support Vector Regression (SVR)
5) Decision Tree Regression
6) K-Nearest Neighbors.

Based on the R2-score, hyperparameter tuning will be applied to each model to improve model performance. Grid Search CV or Randomized Search CV will be used to discover the optimal parameters, depending on how the algorithm operates. Machine learning models are assessed using the cross-validation (CV) resampling technique using a small data sample. Using a sample of the data, the process's sole parameter, k, will be divided into multiples. K-fold cross-validation is the term that is usually used to describe the procedure.

To check how the model predicts, it will represent visualization in the form of graphical images of distplots from the results estimated by the algorithm with the original data. Later an evaluation will be carried out to determine the best algorithm out of the six algorithms to be selected. When finished doing hyperparameter tuning using Randomized Search CV, the code is used as a comment because when the code is run again, the tuning results will be different, even though the model scores are not much different.

2.4.1 Linear Regression

The linear regression technique is used in regression modeling to predict the value of a variable based on the value of another variable. In this study, the linear regression modeling is carried out using the Python library of scikit-learn (sklearn). The first step is to import the "LinearRegression" class from scikit-learn in order to create a linear regression model. Following the model's training using the training set of data "x_train" and "y_train," the model object “LinReg” is produced. After training, the computer uses the model “LinReg” to make predictions on the test data “x_test”. “sc.inverse_transform” is used to restore the prediction results to their original scale since the data was initially standardized. The DataFrame y_pred_LinReg contains the prediction results. This program then creates a distribution plot of the actual data (y_test), which is normalized to the original scale, and a distribution plot of the projected results (y_pred_LinReg), both using the Seaborn (sns) package. We can distinguish between the distributions of the actual and predicted data using the legend, which enables us to assess how well the linear regression model predicts the data.

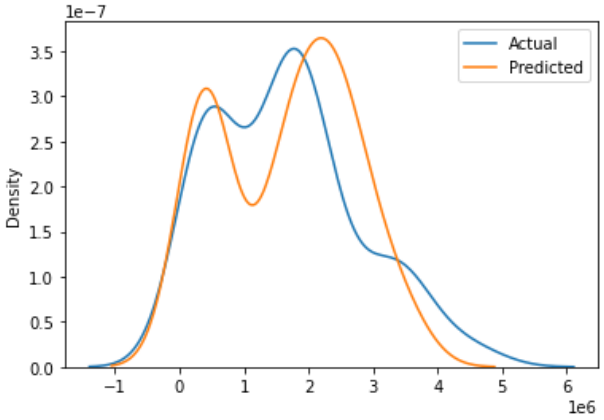

Fig. 5. Visualize the prediction

The performance of the linear regression model was then evaluated using cross-validation, training data, and test data to gauge its accuracy. The model's accuracy score, which measures how well the model matches the training and testing data, is first generated using the training and testing data. To test the model's overall performance, cross-validation with 10 folds (K-Fold) is also carried out. The variable "LinReg_score" contains the cross-validation analysis findings. Finally, the program outputs the average cross-validation score, which is transformed into a percentage, as well as the model's accuracy on the training and testing data. By analyzing these results, one may determine how well the linear regression model fit the data and whether there are any signs of overfitting or underfitting that should be taken into account while creating this model. The evaluation of model performance thus becomes more thorough and educational.

Output:

```
Linear Regression
```

Train : 84.68578496928782
Test : 86.37276105524145
The Average Cross Validation Score is 82.9

2.4.2 Hyperparameter Tuning

The performance and generalization of models in machine learning algorithms are optimized through the use of hyperparameter tuning. The goal of this approach is to optimize the model's hyperparameters so that it can make predictions that are more accurate and handle brand-new data more effectively. By using the Grid Search Cross-Validation (GridSearchCV) method, the Linear Regression model's parameters are tuned. First, the LinReg.get_params() command is used to acquire the default parameters of the Linear Regression model (LinReg). The program then chooses the various parameter combinations to test, including "copy_X," "fit_intercept," "n_jobs," and "positive," as well as the potential values for each parameter. A new Linear Regression model (LinReg_tuning) was developed after testing the parameters. The program then employs GridSearchCV to identify the parameters that produce the best R-squared score, which gauges the strength of the regression model. GridSearchCV tries every combination of parameters while evaluating its performance using training data (x_train and y_train).

The best settings, copy_X: True, fit_intercept: False, n_jobs: 1, and positive: True, are shown in the program output, along with the best R-squared score of around 0.839. The Linear Regression model has been tuned using these parameters to produce the best results depending on the data used in the tuning process. As a result, you can enhance the model's ability to anticipate data based on the preset configuration of the parameters. The new model of Linear Regression has improved just a little bit after hyperparameter tuning.

```
Linear Regression

Train :  84.650818949420
Test  :  86.89566591106708
The Average Cross Validation Score is 83.0
```

2.5 Model Evaluation

The average absolute error, average squared error, and r2 score are used to evaluate the model. The method performs better when its mean absolute error and mean square error are lower, but it's R2 score is higher. The proportion of variance for the items (plants) in the regression model will be represented by the R2-score (coefficient of determination) of the regression score function.

1) The R2-score indicates how well the terms (data points) match the curve or line.
2) The mean absolute error (MAE) is the average absolute difference between the data set's actual and expected values. The MAE computes the average of the residuals in the data set.
3) The mean squared error (MSE) is the average squared difference between the original value and the predicted value in the data set. It determines the variance of the residuals.

Table 4. Model Evaluation

	R2-score	Mean Absolute Error	Mean Squared Error
Linear Regression	85.526260	221938.682671	176940698374.908051
Gradient Boosting	84.864426	267336.839755	185031585168.890594
SVR	82.858047	268481.507628	209559453665.217255
Random Forest	82.823129	253959.746132	215503565358.131348
K-Neighbors	82.716213	316904.942391	211293372515.067657
Decision Tree	81.421643	281254.415963	227119415141.265686

From the results viewed above, a model with a Linear Regression algorithm has the highest R2-score 85,5. So, this model can be the best choice for use in predicting agricultural production in Sumatra, in accordance with the objectives described earlier.

3. Result and Discussion

This research focuses on the application of machine learning techniques to predict rice production in Sumatra Island using 6 algorithms, namely Linear Regression, Random Forest, Gradient Boosting, SVR, Decision Tree, and K-nearest neighbors. The dataset used consists of historical production data, climate data, and soil characteristics relevant to rice production in the region. After tuning the hyperparameters and evaluating the model, then measuring the extent to which the model is able to explain variations in the data using R2-score. The result shows that Linear Regression has the highest R2-score value of 85.53% which indicates this model is very good against variations in test data. Gradient Boosting also produces a fairly good R2-score value of 84.86% while it is in third place followed by SVR which has a slightly lower R2-score value than the two previous models which is 82.86%. Then measuring the average absolute error between the model prediction and the actual value using Mean Absolute Error (MAE), Linear Regression has the lowest MAE value of 221,938.68, indicating a high level of accuracy in prediction. Gradient Boosting has a higher MAE of 267,336.84, while SVR has an almost comparable MAE of 268,481.51. Furthermore, in measuring Mean Squared Error (MSE), which is used to measure the average of the squared error between the model's prediction and the true value, Linear Regression also excels with an MSE of 176,940,698,374.91. Gradient Boosting has a higher MSE of 185,031,585,168.89, and SVR has the highest MSE of 209,559,453,665.22.

Based on the results of this evaluation, Linear Regression is a strong choice as it provides the highest R2-score, lowest MAE, and lowest MSE, which demonstrates its ability to well explain data variations and provide accurate predictions in the context of the dataset used. On the other hand, although other algorithms show good performance, they may be more sensitive to overfitting or less interpretable compared to Linear Regression. For example, ensemble methods such as Random Forest and Gradient Boosting can be prone to overfitting if not properly tuned, and their complexity can obscure the underlying relationships in the data. It is important to recognize that despite the high R2 values achieved by Linear Regression models, there are still challenges in accurately predicting rice production due to the complex and dynamic nature of agricultural systems. Further study is required to increase the model's robustness because external factors such as climate fluctuation, changing environmental conditions, and other factors can have a substantial impact on rice output.

Overall, this research demonstrates the potential of machine learning, specifically Linear Regression, as a valuable tool for predicting agricultural production in Sumatra. This study emphasizes the importance of using data-driven approaches in planning and decision-making in agriculture to optimize production and resource allocation. Future studies can explore incorporating additional data sources, considering more advanced machine learning techniques, and addressing the challenges posed by dynamic farmland to further improve the prediction accuracy of rice production models in the region.

4. Conclusion

In this study, linear regression was used to estimate rice production in Sumatra Island and was contrasted with five different algorithms: Random Forest, Gradient Boosting, SVR, Decision Tree, and K-Nearest Neighbors. After comparing the various models, the linear regression technique has the greatest R2-score (85.53%), demonstrating that it adequately explains the

variation in the data. In comparison to some of the other algorithms examined, it has a Mean Absolute Error (MAE) value of 221938.68 and a Mean Squared Error (MSE) value of 176940698374.90, showing that Linear Regression tends to produce more precise predictions that are nearer to the true value. These findings demonstrate the potential of machine learning in improving rice production prediction models for agricultural planning and decision-making in the region. In addition, this study shows the advantages and disadvantages of each algorithm, which provides valuable insights in choosing the right method for a specific application. Although several algorithms showed good performance, linear regression with its simplicity and high interpretability remains the superior choice in some cases.

However, this study also recognizes that there are several challenges in predicting rice production, such as complex variations in environmental conditions and climate change. Therefore, improving data quality, selecting appropriate features, and more sophisticated data management techniques will be the focus of future research. In the context of agricultural development and resource planning, the application of machine learning in rice paddy production prediction in Sumatra offers a great opportunity to improve efficiency and accuracy in decision-making. It is envisaged that the findings of this study would be useful in managing Sumatra's agricultural resources and will also serve as a roadmap for future research in this area.

Acknowledgement

We want to thank everyone who contributed to the success of our research work. We also wish to express our gratitude to our supervisors for their guidance, inspiration, and supportive counsel during the research process. We also appreciate our family and friends for their help.

References

[1] Bharath, S., L, Y. B., & Javalagi, V. R. (n.d.). *Comparative Analysis of Machine Learning Algorithms in The Study of Crop and Crop yield Prediction*. www.ijert.org

[2] F Mardianto, M. F., Tjahjono, E., Rifada, M., Putra, A. L., & Utama, K. A. (n.d.). *The Prediction of Rice Production in Indonesia Provinces for Developing Sustainable Agriculture*.

[3] Inka Sari, C., & Darma Setiawan, B. (2019). *Prediksi Volume Impor Beras Nasional menggunakan Metode Support Vector Regression (SVR)* (Vol. 3, Issue 5). http://j-ptiik.ub.ac.id

[4] Jiya, E. A., Illiyasu, U., & Akinyemi, M. (2023). Rice Yield Forecasting: A Comparative Analysis of Multiple Machine Learning Algorithms. *Journal of Information Systems and Informatics*, *5*(2), 785–799. https://doi.org/10.51519/journalisi.v5i2.506

[5] Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., & Uwamahoro, A. (2023). Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture (Switzerland)*, *13*(1). https://doi.org/10.3390/agriculture13010225

[6] Mardhika, D. A., Darma Setiawan, B., & Wihandika, R. C. (2019). *Penerapan Algoritma Support Vector Regression Pada Peramalan Hasil Panen Padi Studi Kasus Kabupaten Malang* (Vol. 3, Issue 10). http://j-ptiik.ub.ac.id

[7] Meizenty, S., Sahid, S. S. D., & Sari, N. J. (2021). Rice Quality Detection Based on Digital Image Using Classification Method. *International Applied Business and Engineering Conference 2021*, 1–4.

[8] Putra, H., & Ulfa Walmi, N. (2020). Penerapan Prediksi Produksi Padi Menggunakan Artificial Neural Network Algoritma Backpropagation. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *6*(2), 100–107. https://doi.org/10.25077/teknosi.v6i2.2020.100-107

[9] Rizkyana, st, & Yohana Dewi Lulu Widyasari, nd. (2021). Investigating the Effect of Climate on National Rice Production using Machine Learning. In *International Applied Business and Engineering Conference*.
[10] Rusmilawati, N., & Prasetyaningrum, P. T. (n.d.). *Penerapan Data Mining Dalam Prediksi Hasil Produksi Kelapa Sawit PT Borneo Ketapang Indah Menggunaka Metode Linier Regression*.
[11] Sari, R. N., Saumi, F., Olivia, M., Maidita, R., Salsabilah, A., & Agustina, I. (2022). Application of Linear Regression to The Factors Affecting The Rice Production Level in Langsa City. *Mathline : Jurnal Matematika Dan Pendidikan Matematika*, *7*(2), 315–329. https://doi.org/10.31943/mathline.v7i2.282
[12] Yuliska, & Sakai, T. (2019). A Comparative Study of Deep Learning Approaches for Query-Focused Extractive Multi-Document Summarization. *IEEE 2nd International Conference on Information and Computer Technologies*.