JSE Journal of Science and Engineering, Vol. 3, No. 2 (page: 80-84)
e-ISSN: 2723-3944 January 2025
DOI: https://doi.org/10.30650/jse.v3i2.3868

ID CARD PRINTING TIME PREDICTION WITH MEAN ABSOLUTE ERROR IN LINEAR REGRESSION

Akmal Mugni Fawwazrin[1], Arbansyah[2], Amin Rezaeipanah[3]
[1-2] Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
[3]Department of Computer Engineering, University of Rahjuyan Danesh Borazjan, Bushehr, Iran
* Corresponding Email: 2011102441166@umkt.ac.id

Abstract – The Population and Civil Regrestration Center, know in Indonesia as the Dinas Kependudukan dan Catatan Sipil (Dukcapil), is a government agency responsible for managing and recording population data in specific regions. Dukcapil is responsible for recording births, deaths, marriages, divorces, and documenting identities through the Electronic Resident Identity Card (E-KTP). Dukcapil faces the need to address delays in document issuance, which will be the focus of this research. The data were collected directly when researchers conducted Praktek Kerja Lapangan (PKL) at Dukcapil. This study utilizes the linear regression method to estimate the time required for printing E-KTP cards that will be produced and calculates the Mean Absolute Error (MAE). Thorough MAE calculations using the linear regression method, researchers obtained a result of 10.2923 from the available data.

Keywords: Prediction; E-KTP; Linear Regression; MAE

Submitted: 17 December 2023 - Revised: 20 December 2023 - Accepted: 31 January 2025

1. Introduction

The Population and Civil Registration Center or as known in Indonesia as Dinas Kependudukan dan Pencatatan Sipil (Dukcapil) plays a crucial role in the administration of a country's population. Dukcapil is a government institution responsible for managing and recording population data in specific region [1]. It ensures the accuracy, reliability, and accessibility of population data. Dukcapil is responsible for recording births, deaths, marriages, divorces, and identity documentation, fostering collaborations with educational institutions and the business sector to support its objectives. As a central government institution, Dukcapil contributes significantly to creating an organized and reliable population administration system. This aligns with the nation's need for a robust population data infrastructure, particularly in the era of globalization and information technology advancements.

Dukcapil also plays a vital role in meeting citizens' needs, such as issuing Electronic Identity Cards (E-KTP). Kartu Tanda Penduduk Elektronik (E-KTP) is an identity card created electronically, meaning both its physical and usage are computerized [2]. To address delays in the issuance of population documents, this project aims to use linear regression methods to estimate the time needed for printing E-KTP. Regression analysis is a statistical technique used to determine the strength of the linear relationship between an independent variable (X) and a dependent variable (Y). Regression analysis is based on the cause-and-effect correlation or the functional correlation between one variable and another [3]. In the context of Field Work Practice (PKL) implementation, collaboration between educational institutions, Dukcapil, and students is crucial. Praktek Kerja Lapangan (PKL) is one form of activities that take place directly in a work environment. Internship can be undertaken by vocational high school students, college students, or new employees. At the college level, an internship is a systematic and synchronized implementation between the educational program at school and the mastery program of skills acquired through direct work activities in the professional world to achieve a certain level of expertise. In the context of college, it is commonly referred to as "internship" or "internship program." [4]. Students not only apply theoretical knowledge from lectures but also gain practical experience, acquire additional knowledge, and develop social skills in a real working environment.

The benefits of implementing PKL extend to Dukcapil and the university directly. Dukcapil receives assistance in easing workload burdens, while the university enhances the quality of students through practical experiences during PKL. The positive relationship forged between the university and Dukcapil creates a mutually beneficial and sustainable partnership. Overall, this project not only provides a solution for efficient population administration through the prediction of E-KTP printing time but also

e-ISSN: 2723-3944

yields long-term benefits through collaboration between the education sector, government, and business world.

2. Prediction

Prediction is an estimation of the demand level for a product or service during a specific period in the future [5]. According to Ica Admirani (2018:10), prediction is not merely speculation but also a technique that utilizes historical data to estimate projections or trends that may occur in the future. In this context, the use of historical data is crucial in the prediction process. This data serves as the foundation to identify specific patterns or trends that can be used as a reference to make estimations about events or outcomes in the future. Therefore, prediction is not solely based on assumptions or speculation but also on a careful analysis of the information that has occurred.

In a more detailed definition, prediction can be seen as an effort to make forecasts about events or outcomes in the future by referencing the knowledge and information gathered at the present time. By utilizing existing data, prediction provides a representation or estimation of the potential developments or changes in the future.

It is important to note that prediction is not absolute and always involves a degree of uncertainty. Various factors can influence the accuracy of predictions, and the results can vary depending on the analytical methods and the quality of the data used. Nevertheless, in various fields such as economics, meteorology, and other sciences, prediction remains a crucial tool to aid in future planning and decision-making.

3. Python

Python is the most relevant programming language used by data scientists for various data science applications. It also has excellent functionality for handling mathematics, statistics, and scientific functions [9]. Notably, Python stands out for its clean and straightforward syntax, making it accessible for both beginners and experienced developers. The language prioritizes readability, reducing the cost of program maintenance and development.

One of Python's distinctive features is its extensive library ecosystem. This rich set of libraries empowers programmers to build advanced applications using concise and seemingly uncomplicated source code. Python's philosophy of "batteries included" reflects the idea that the language comes bundled with a diverse range of modules and libraries, covering everything from web development and data analysis to artificial intelligence.

4. Data

According to Triska Apriyani (2017:2), data is a collection of events extracted from reality in the form of numbers, letters, or specific symbols or combinations thereof that have not yet revealed much information. Therefore, further processing is required [6].

5. Linear Regression

Linear regression serves as a method to examine the correlation between independent and dependent variables [7]. In the context of simple linear regression, we consider the relationship between one independent variable (X) and one dependent variable (Y). The goal is to build a mathematical model that can predict the value of Y based on the value of X.

1) Formula for Simple Linear Regression:

$$Y = a + bX \quad (1)$$

Description:
- Y : Respons variable,
- a : Intercept contant,
- b : Regression coefficient
- X : Predictor variabel

2) Calculation of Regression Coefficient (b) and Intercept (a):

$$b = \frac{\sum_{i=1}^{n}(x_{i-}\bar{x})(y_{i-}\bar{y})}{\sum_{i=1}^{n}(x_{i-}\bar{x})^2} \quad (2)$$

$$a = \bar{y} - b\bar{x} \quad (3)$$

Description:
- n : Number of observations
- x_i y_i : Value observation for variables X and Y
- $\bar{x}$ $\bar{y}$: The averages of X and Y, respectively

By using the calculated values of a and b, we can create a simple linear regression model that can be used to predict the value of Y based on the value of X. This model aims to estimate the best-fitting straight line representing the linear relationship between the two variables

6. Mean Absolute Error (MAE)

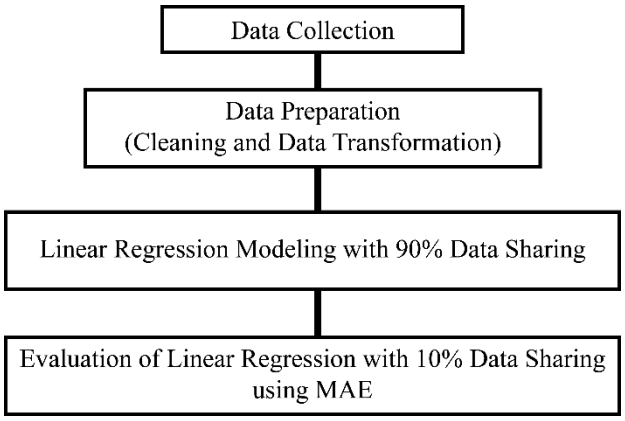
Mean Absolute Error (MAE) is a method used to measure the accuracy of a forecasting model [8]. The MAE value indicates the average absolute error. The formula for MAE is expressed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (4)$$

Description:
- f_i : Predicted value
- y_i : Actual value
- n : Number of data points

MAE calculates the average error by assigning equal weight to all data points (i = 1...n) intuitively. For forecasting model evaluation, MAE provides an intuitive measure of the average error across all data points. In this research, selecting MAE is appropriate because all data points are given equal weight.

e-ISSN: 2723-3944



7. Result and Discussion

To solve the problems, the experiment in this research has four stages, as shown in Figure. As in Figure, this research starts with collecting data from District office of Lempake. Furthermore, data preparation must be conducted to ensure Linear Regression can process the data. The Linear Regression acts as a predict model to predict the duration for printing Identity Card (E-KTP). Then, for evaluation, this research employs a Mean Absolut Error as a tool to calculate the average difference between the actual value and the predicted value.

7.1. Data Collection

The dataset used in this study was directly obtained from the district office where the author actively participated during the fieldwork internship. Collected during practical work experience, this data is particularly relevant as it is associated with the specific operational context of the kecamatan (district) office. The total number of data points used in our analysis is 368, focusing on ID card printing.

There are 11 attributes in the dataset, namely KK (Family Card), NIK (National Identification Number), Nama, Umur, Jenis_Kelamin, RT (Residential Code), Kelurahan (Sub-district), Tanggal_Pengajuan, Waktu_Pengajuan, and Waktu_Selesai.

7.2. Data Preparation

The data obtained from the Lempake District will undergo a preparation phase to clean the data from its previously unstructured form into a more organized structure, facilitating the prediction process. The data preparation process involves steps for data Prepocessing and data Transformation. Here are the steps:

1) Data Preprocessing

The initial phase of research begins with the preprocessing stage. In this step, specific emphasis is placed on data that includes empty or absent values. The presence of data with missing values is a primary concern because such data does not include values or contains empty information. Therefore, the first crucial step is to clean and manage data with missing values to establish a robust and consistent foundation for subsequent research processes. The initial dataset consisted of 368 rows, and after preprocessing, it was reduced to 351 rows.

2) Data Transformation

In this step, data transformation is carried out with the aim of converting categorical data types into numeric forms using the label encode technique. This process is crucial as linear regression modeling can only handle numerical data. For instance, the data undergoes transformation by replacing each value in a column with consecutive numbers such as 1, 2, 3, and so on.

```
data['Jenis_Kelamin'].replace(['L','P'], [1,0], inplace=True)
data['Keterangan'].replace(['PRR','Hilang','Ubah Data',
                            'Rusak','Pindahan','Pemekaran'], [1,2,3,4,5,6], inplace=True)
data['Kelurahan'].replace(['SPJB', 'SPJU', 'SPJS', 'SPJT',
                           'TM', 'LPK', 'SS', 'BDP'], [1,2,3,4,5,6,7,8], inplace=True)

data['Pengajuan_Minutes'] = data['Waktu_Pengajuan'].apply(
                            lambda x: int(x.split(':')[0]) * 60 + int(x.split(':')[1]))
data['Selesai_Minutes'] = data['Waktu_Selesai'].apply(
                            lambda x: int(x.split(':')[0]) * 60 + int(x.split(':')[1]))
```

Figure 1. Transformation Code

The provided Python code operates on a DataFrame named 'data.' It transforms categorical data by replacing specific values in the 'Jenis_Kelamin,' 'Keterangan,' and 'Kelurahan' columns with numerical equivalents. For instance, 'L' and 'P' in 'Jenis_Kelamin' are replaced with 1 and 0, respectively. Similarly, values in 'Keterangan' and 'Kelurahan' are replaced with corresponding numerical codes.

Additionally, the code creates two new columns, 'Pengajuan_Minutes' and 'Selesai_Minutes,' by converting time values from 'Waktu_Pengajuan' and 'Waktu_Selesai' columns into minutes. This is achieved by splitting the time strings into hours and minutes, converting hours to minutes, and summing the results.

7.3. Modeling and Evaluation

During the Field Work Practice (PKL) period at the Civil Registration Office (Disdukcapil), the author successfully developed a program for predicting the printing of ID cards using the Linear Regression method. Here are the results of the internship work:

1) Correlation

This stage serves to explore the relationships among attributes, which can be done through correlation analysis. The objective is to identify which attributes play a more significant role in this prediction. Correlation is determined by examining the largest positive values of attributes concerning the label data.

	KK	NIK	Umur	Jenis_Kelamin	RT	Kelurahan	Keterangan	Pengajuan_Minutes	Selesai_Minutes
KK	1.000000	0.465218	-0.258294	0.093662	0.075520	-0.084155	-0.199002	0.086739	0.095894
NIK	0.465218	1.000000	-0.074910	0.033347	-0.034309	-0.039114	-0.168293	0.041442	0.043302
Umur	-0.258294	-0.074910	1.000000	0.007535	0.039239	-0.001773	0.599792	-0.060347	-0.118225
Jenis_Kelamin	0.093662	0.033347	0.007535	1.000000	0.002069	-0.070832	-0.051478	-0.031465	-0.031989
RT	0.075520	-0.034309	0.039239	0.002069	1.000000	0.072160	0.065781	-0.060082	-0.064589
Kelurahan	-0.084155	-0.039114	-0.001773	-0.070832	0.072160	1.000000	0.064166	-0.014703	-0.009037
Keterangan	-0.199002	-0.168293	0.599792	-0.051478	0.065781	0.064166	1.000000	-0.115603	-0.183161
Pengajuan_Minutes	0.086739	0.041442	-0.060347	-0.031465	-0.060082	-0.014703	-0.115603	1.000000	0.989692
Selesai_Minutes	0.095894	0.043302	-0.118225	-0.031989	-0.064589	-0.009037	-0.183161	0.989692	1.000000

Figure 2. Attributes Relationships

e-ISSN: 2723-3944

Based on Figure 3, three attributes with the highest correlation values will be selected. In the above picture, it is evident that the top three correlation values are 'Umur', Keterangan' and 'Pengajuan_Minutes'.

2) Split Data

Before proceeding with the evaluation, the next step is to split the dataset into 90% training data with a total of 316 and 10% testing data with a total of 35.

3) Modeling

```
from sklearn.linear_model import LinearRegression
reg = LinearRegression()
reg.fit(xtrain, ytrain)
```

Figure 3. Modeling

Figure 3 show the creation of linear regression model in python begins by importing the Linear Regression model from the scikit-learn library. Subsequently, it initializes an instance of the Linear Regression class, which serves as the linear regression model. The `fit` method is then employed to train the linear regression model using the provided training data. Here, `xtrain` denotes the input features, and `ytrain` corresponds to the target values (labels). Through this training process, the model acquires knowledge of the association between the input features and the target values.

4) Result

The subsequent step involves determining the predicted values corresponding to the input data. These values are then converted into hours and minutes as part of the prediction value conversion process, resulting in the prediction outcome in the form of hours and minutes.

Table 1
Converted into Hours and Minutes

No	Predicted Value	Format Hours
1	625	10:25
2	559	09:19
3	966	16:06
4	699	11:39
5	640	10:40
….	….	….
….	….	….
34	643	10:43
35	649	10:49

It can be observed here that the initial predicted value of 625 has been converted into hours and minutes, resulting in 10:25. This process involves treating the prediction value as minutes, which will later be converted into hours and minutes. Thus, 625 minutes are equivalent to 10 hours and 25 minutes.

5) Comparison of Prediction Values

In this stage, a comparison is made between the actual values and the predicted values. This analysis allows for evaluating how well the model or method used can accurately predict values. By comparing the differences between the actual and predicted values, it is possible to identify the extent of the prediction quality. Table 3 show the results of this comparison can provide valuable insights for assessing the accuracy and performance of the model used in the context of the ongoing analysis or prediction.

Table 2
Comparison Between Predicted and Actual Value

No	Predicted Value	Actual Value
1	625	622
2	559	559
3	966	945
4	699	697
5	640	647
….	….	….
….	….	….
34	643	641
35	649	643

The results of the process in Table 2 above can be presented in the form of a graph in Figure 4. The graph provides an illustration of the comparison between the original data (blue) and the predicted data (red), visualizing how well the linear regression model approximates the actual values. This graphical representation aids in understanding the extent to which the model captures the patterns and trends present in the dataset.

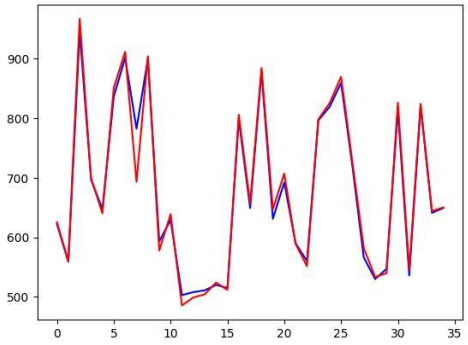


Figure 4. Plot Between Predicted and Actual Value

6) Evaluation

After processing the data using the Linear Regression algorithm, the next step is to test the error values of the Linear Regression algorithm using Mean Absolute Error (MAE). We can calculate the Mean Absolute Error (MAE) based on the data in Table 2.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

$$= \frac{(625-622)+(559-559)+(966-945)+\cdots+(649-643)}{35}$$

$$= \frac{3+0+21+\cdots+6}{35}$$

e-ISSN: 2723-3944

$= 10.29236$

Based on the Mean Absolute Error (MAE) calculation that has been performed, the obtained Mean Absolute Error (MAE) value is 10.2923. A decreasing MAE value indicates that the error is getting smaller, signifying the model's ability to make accurate predictions with minimal deviation from the actual values.

8. Conclusion

In this project, the author successfully developed a programming solution enabling the Department of Population and Civil Registration (DUKCAPIL) to predict the duration of E-KTP (Electronic Identity Card) issuance. Drawing insights from fieldwork at the DUKCAPIL Samarinda and research utilizing the Linear Regression algorithm for predicting E-KTP processing times, several conclusions can be drawn:

1) Significance of Technological Implementation:

- The utilization of technology, particularly through Python programming and the Linear Regression algorithm, has significantly contributed to optimizing the estimation of E-KTP issuance duration at DUKCAPIL.
- The implementation of technology, as demonstrated through a dedicated dashboard, played a crucial role in enhancing visualization and understanding of the prediction results.

2) Influence of Specific Factors:

- The use of technology, specifically Python programming and the Linear Regression algorithm, proved to be a substantial contributor to refining the estimates of E-KTP processing duration.
- The project highlighted the capability of the algorithm to identify and analyze specific factors influencing the timeline for E-KTP issuance.

These conclusions underscore the pivotal role of technology and scientific approaches in enhancing the efficiency and effectiveness of DUKCAPIL in managing and predicting the processing time of E-KTP. The findings provide valuable insights for the continuous improvement and development of administrative processes related to population and civil registration.

Acknowledgements

The author expresses heartfelt gratitude to those who have provided assistance and support, both in the implementation of the internship and in the writing of this research journal, especially to:

1. Mr. Arbansyah, S.KOM., as the Head of the S1 Computer Engineering Program at the University of Muhammadiyah East Kalimantan and the Internship Advisor.
2. To Parents and siblings who consistently pray for and provide support, both morally and materially.
3. To other individuals who cannot be mentioned one by one but have provided encouragement, motivation, and assistance, both directly and indirectly, for the smooth preparation of this research journal.

The author acknowledges that there are still shortcomings in this research journal. Therefore, the author welcomes contructive criticism and suggestions for improvement in the refinement of this research journal.

References

[1] House of Representatives of the Republic of Indonesia. (2006). Law of the Republic of Indonesia Number 23 of 2006. https://www.dpr.go.id/dokjdih/document/uu/UU_2006_23.pdf

[2] Wikipedia (2023). Electronic Identity Card. https://id.wikipedia.org/wiki/Kartu_Tanda_Penduduk_elektronik

[3] U. N. Padang, S. Barat, and G. L. Rizal, "The relationship between cyberloafing and work procrastination in Bukittinggi Fauza City employees," J. Ris. Psychol., vol. 5, no. 4, pp. 167–177, 2022.

[4] STIMIK DUMAI (2023). Field Work Practice. https://stmikdumai.ac.id/praktek-kerja-lapangan/

[5] I. S. D. Andani, H. Oktavianto, and T. T. Warisaji, "Comparison of Forecasting the Number of Coffee Product Sales Using the Linear Regression and Single Moving Average Method," J. Apl. Sist. Inf. and Elektron., vol. 4, no. 2, pp. 70–77, 2022.

[6] I. A. W. M. Arfa Andika Candra, "WEB-BASED ACHIEVEMENT INFORMATION SYSTEM AT SMP NEGERI 7 METRO CITY," J. Mhs. Ilmu Komput., vol. 16, no. 4, pp. 327–332, 2021, doi: 10.22141/2224-0721.16.4.2020.208486.

[7] [1] E. Mardiani, N. R. ...

e-ISSN: 2723-3944