Leveraging Temporal Feature Expansion for Enhanced Prediction of Naive Bayes and Random Forest Classification on SWSR

1st Trisula Darmawan
School of Computing
Telkom University
Bandung, Indonesia
trisuladarmawan@student.telkomuniversity.ac.id

2nd Sri Suryani Prasetiyowati
School of Computing
Telkom University
Bandung, Indonesia
srisuryani@telkomuniversity.ac.id

3rd Yuliant Sibaroni
School of Computing
Telkom University
Bandung, Indonesia
yuliant@telkomuniversity.ac.id

***Abstract*— Based on data from the Central Statistics Agency in the first semester of 2023, Central Java is one of the provinces in Indonesia with a percentage of poor people exceeding the national average rate. From these data, it can be understood that Central Java needs more attention to reduce poverty, including through effective data management of the Social Welfare Service Recipients (SWSR) database so that it can be the basis for developing social welfare service programs. Therefore, this research uses Naïve Bayes and Random Forest algorithms and combines them with a temporal feature expansion method that allows machine learning models to capture time-based patterns in the data so that the model can predict the classification of SWSR distribution in all districts/cities in Central Java for the next few years. The use of the time-based feature expansion method in machine learning classification has the advantage of identifying factors that affect future classification predictions, in contrast to time series or LSTM methods that only produce predictions without revealing these factors. The results show that the performance between the two methods is similar by getting an accuracy score of 85.71% on the best t-k model. Meanwhile, in terms of prediction length, Naïve Bayes Time-Based can predict up to the next 10 years and better than Random Forest Time-Based which is able to predict for the next 9 years. This research is expected to be able to obtain accurate and reliable prediction results to support decision-making in social welfare policies in Central Java Province.**

Keywords*—*social welfare service, naive bayes, random forest, feature expansion

I. Introduction

Social Welfare Service Recipients (SWSR) known as Pemerlu Pelayanan Kesejahteraan Sosial (PPKS) in Indonesia is an individual, family, group, and/or community who, due to an obstacle, difficulty, or disturbance, cannot carry out their social functions, thus requiring social services to fulfill their physical and spiritual and social needs adequately and reasonably [1]. SWSR is one of the scopes of Data Terpadu Kesejahteraan Sosial (DTKS), an integrated social welfare data, which is used as a reference database in organizing social welfare services.

The location of this research is in Central Java Province, which according to data from the Central Statistics Agency in the first semester of 2023 is one of the provinces in Indonesia that has a percentage of poor people exceeding the national average percentage [2]. The percentage of poor people in Central Java is 10.77%, while the national percentage is 9.36%. Based on DTKS, in 2023 the number of SWSR in Central Java was 4,112,263 people [3], covering 10.95% of Central Java's population. These data show the social welfare problems that exist in Central Java, as well as highlighting the importance of effective and targeted social welfare programs to resolve these problems. One important step in designing a good program is to understand the conditions in each region in depth. Therefore, a map predicting the future distribution of SWSR is needed to provide an overview of the severity level in each area. With this map, social welfare policies can be tailored to the specific needs of each region.

Several studies have been conducted to classify social assistance recipients and economic status using Naïve Bayes and Random Forest algorithms. Research on Naïve Bayes has shown its effectiveness, such as in the Philippines with an error rate of 0.0014 [4], and in Indonesia reaching an accuracy rate of 89.04% [5], 95.83% [6], and 84.24% [7] to classify the eligibility of social assistance recipients. Similarly, Random Forest shows its superiority, such as in classifying economic status in Cirebon City with 93% accuracy [8], beating other methods in Indonesia [9] and China [10], and utilizing geospatial and multi-source data to measure poverty with high reliability [11]. These findings confirm the ability of both algorithms to perform classification in social and economic domains.

Research [12] compared the Support Vector Machine (SVM) Time-Based, Long Short Term Memory (LSTM), and K-Nearest Neighbor (KNN) methods on the SWSR dataset in Central Java which is similar to the dataset used in this study. The results showed that the t-k model with the highest accuracy reaching 87.14% was obtained by the SVM Time-Based model which outperformed KNN with its best t-k model with an accuracy of 80.89%. SVM Time-Based also outperforms the accuracy of LSTM prediction results by 54.28%.

Time-based feature expansion methods have proven to be effective in improving the capabilities of Naïve Bayes and Random Forest so that they can be used for future classification predictions. A study [13] obtained high accuracy on the Dengue Hemorrhagic Fever (DHF) dataset and rainfall dataset using the Naïve Bayes Time-Based method. Meanwhile, [14] used a Random Forest Time-Based model for classification prediction on the DHF dataset. The high accuracy of classification prediction results shows the advantages of using the Random Forest Time-Based method.

Considering the results of previous studies that show the superiority of time-based feature expansion which is proven to improve the ability of machine learning to predict future classifications, in this study the time-based feature expansion method is used in the classification of SWSR distribution in the Central Java province of Indonesia. The machine learning algorithms used are Naive Bayes and Random Forest, combined with the temporal feature expansion method to capture time-based patterns within the data. Besides predicting the future distribution of SWSR classifications in each district/city in Central Java, this innovative approach allows the model to identify important factors that influence the prediction. This is an advantage of this method when compared to the conventional time series method or LSTM which only focuses on making forecasts.

This research aims to create an accurate prediction map of the SWSR distribution in Central Java in the next few years, which can provide valuable information for policymakers to design targeted and effective social welfare programs according to the conditions in each region. In addition, this research is also expected to identify key factors that influence the increase in the number of SWSRs, so that policymakers understand the aspects that need to be prioritized to improve social welfare.

II. Materials and Method

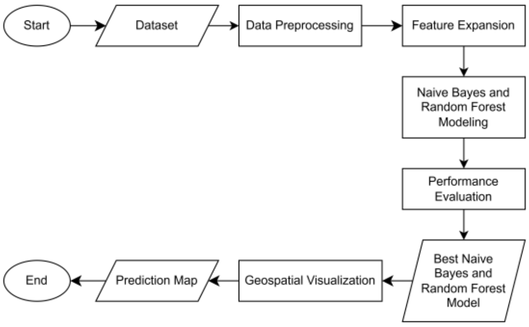

Fig. 1 System Design

A. Dataset

The dataset used in this research is the SWSR dataset in Central Java which has been used in previous research [12] The original dataset of the prior study contains spatio-temporal data related to SWSR in each district/city in Central Java in 2013-2022. The target of this dataset is the number of SWSR for each districts/cities in Central Java per year. This research adds the target in 2023 which has been released on the Central Java Social Service website [3] to the dataset. Meanwhile, the features represented by $X1, X2, \dots, Xn$ in this dataset are factors that are closely related to the number of SWSR as follows:

TABLE I. Data Features

Features	Description
X1	Total Population
X2	Minimum Wage
X3	Average Number of Family Members
X4	Percentage of Poverty
X5	Number of Unemployed People
X6	Number of Working People
X7	Elementary School Graduate
X8	Junior High School Graduate
X9	High School Graduate
X10	College Graduate
X11	Average Length of School
X12	Male Human Development Index
X13	Female Human Development Index
X14	Expenditure per Capita per Person (rupiah)

The targets in the dataset, which are continuous values, need to be converted into categorical data in order to be classified. Given that the government has not made any official provision regarding the classification of the number of SWSRs, this research will use the data class label proposed by previous research [12] using the equal frequency data binning method.

TABLE II. Data Class Labels

Categories	Class	Range
Low	0	0 ≤ SWSR ≤ 78.500
Medium	1	78.500 < SWSR ≤ 146.000
High	2	SWSR > 146.000

B. Data Preprocessing

Data preprocessing is necessary to improve data quality so that machine learning algorithms can more easily recognize the nature of each attribute in the dataset [15]. The data preprocessing technique carried out in this research includes the following steps:

1) *Augmented Dickey-Fuller test:* The Augmented Dickey-Fuller test is a method for testing the stationarity of a time series. This method proposes a null hypothesis that the time series is not stationary because it has a unit root. Subsequently, the null hypothesis is evaluated against the probability of being statistically rejected in favor of the alternative hypothesis that the time series is stationary [16].

2) *Min-Max Normalization:* Min-Max Normalization is a data scaling technique used to change the range of data values from 0 to 1 [17]. This process homogenizes the scale of data values on each feature to reduce potential bias towards certain features.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

3) *Synthetic Minority Oversampling Technique (SMOTE):* Among various oversampling techniques, SMOTE is the most widely utilized method [18]. SMOTE addresses data imbalance by creating new samples in minority data classes until the number of samples in each class becomes uniform. New samples are obtained through a random interpolation process between a sample and its neighboring samples [19]. This stage is carried out after the data is divided with a proportion of 80% for training data and 20% for validation data. SMOTE is applied to the training data to ensure data balance.

C. Feature Expansion

In this research, the feature expansion method is used to enable machine learning models to predict future SWSR classifications. Feature expansion works by adding features of other objects to combine the relationship between the features of both [13]. The application of feature expansion in this study is to combine the features of data in the previous k years to predict the classification in the next k years.

The $t - k$ model is built by applying a combination of features from k years before year t that are evaluated against the target in year t. The following is a combination of features used in the $t - k$ model based on available data from 2013 to 2023: