Correlation Analysis on Classification of Government Employee Performance Results with Employment Agreements using Algorithms K-Nearest Neighbour (Case Study: Kebumen Regency Government)

Siti Rokhanah
Master of Information Technology, Postgraduate Program, Universitas Teknologi Yogyakarta
Sleman, Indonesia

Arief Hermawan
Master of Information Technology,Postgraduate Program, Universitas Teknologi Yogyakarta
Sleman, Indonesia

ABSTRACT

This research focuses on the performance appraisal of employees in government agencies and how correlations are in data processing to improve model accuracy. By focusing on data preparation, handling missing data, imbalanced data and feature selection. The purpose of the study is to provide an understanding of the interaction between these methods in the context of performance result analysis.

This study includes four experimental scenarios that consider a combination of data preprocessing methods. Each scenario is designed to evaluate the performance of the k-nearest neighbour algorithm on the dataset of performance results of Government Employees with Work Agreements in the Kebumen District Government. The method steps include data preparation, handling missing data and feature selection based on the Correlation Matrix to overcome High Dimensional Data and the K-Nearest Neighbour method to display and produce the final results of processing data.

The test results show that using a combination of data pre-processing methods can significantly increase the accuracy of the K-Nearest Neighbor model on the performance results dataset for Government Employees with Performance Agreements. The highest accuracy was obtained in the pre-processing scenario when applying correlation with the Correlation Matrix technique and K-Nearest Neighbor classification by eliminating attributes that had a high correlation value, namely 100 %

Keywords

Employee Performance; Classification; K-Nearest Neighbour; Correlation Matrix.

1. INTRODUCTION

Employee Performance Appraisal requires agencies to create competitive advantages by conducting Human Resource management. Employees are human resources who play an active role in achieving the vision and mission of an agency. At work, employees are in a social environment that involves emotional intelligence (EQ) and intellectual intelligence (IQ) when carrying out their duties. Nonetheless, the impact of effective performance appraisals on actual employees and the performance of improvement organizations is well established in the literature. Assessment of employee work implementation is an activity process carried out to evaluate the level of work implementation or performance of an employee [2]

Performance appraisal of Government Employees with work agreements (PPPK) is an important factor in the Management of the State Civil Apparatus, especially in considering leaders to conduct performance evaluations. This process is needed as one of the considerations to extend or terminate the PPPK performance contract within the Kebumen Regency Government. In processing PPPK performance appraisal, data mining can help obtain accurate and complete information about PPPK's profile, so that it can be used as a basis for decision making.

Challenges in data processing are high data complexity due to data diversity or heterogeneity [3], inconsistent data, noise, missing data, High Dimensional Data, and data imbalanced [22]. Challenges in data processing can be overcome by preparing data before processing and commonly referred to as data preprocessing [21]. Data preprocessing consists of a series of techniques to clean, transform, and prepare raw data before it is analyzed using classification methods [6]. The purpose of data preprocessing is to improve data quality, reduce noise, eliminate irrelevant data, and normalize variables so that they can be used optimally in the classification process so that accuracy increases. [17]

Pre-processing of data is very important to do, if not done it will have an impact on decreasing classification performance. Incorrect selection of pre-processing techniques may result in incorrect predictions. In today's era of big data, when large volumes of data are generated every second, and the use of such data is the main concern of policymakers, efficient handling of lost data becomes increasingly important [19].

Missing data can be overcome by several methods, of which the most commonly used methods are Removing Features, Mean Imputation, and K-Nearest Neighbor. Removing Features is done by [30] by removing features or attributes when the number of Missing data is more than 55%. KNN replaces the missing value with the corresponding variable average of k its nearest neighbor. KNN does not consider data structures that may exist in the in the dataset. This can lead to inaccurate imputations if there is a complex pattern or relationship between the missing variable and the non-missing variable [4]

Research by [18] states that the Correlation Matrix method can increase the accuracy value, where the results of his research for wine dataset classification after applying the Correlation Matrix to normalize Z-Score accuracy rose from 73. 75% to 75.62% and at Min-Max normalization accuracy rose from 68.12% to 71.25%. Research conducted by [10] also states that feature selection uses Correlation Matrix can increase the accuracy of classification of poisonous mushrooms from 97.97 % to 99.02 %. The study found 2 of the 22 input variables that had a very low contribution to the classification results and tended to be irrelevant confounding variables in the model, so

the 2 variables were omitted.

Research on data pre-processing methods has been carried out, but those that discuss the performance appraisal data of Government Employees with Work Agreements are still very few and are usually carried out for performance appraisals in Civil Servants such as in research by [7] which discusses the classification of Civil Servant Performance Results with 98.8% accuracy results and has not discussed how the correlation affects accuracy on classification method on data on the performance results of Government Employees with Work Agreements. This research will apply the pre-processing technique of feature selection using the Correlation Matrix which will be used to overcome High Dimensional Data, and the method used by K-Nearest Neighbor as an algorithm used to replace the missing value with the average of the corresponding variables from k nearest neighbors because of several previous studies that discuss performance results has never discussed data pre-processing before classification. Classification is a role in data mining that uses a predictive approach method, has the following definition: If there is a set of records (training set) where each record consists of a set of attributes and one attribute is a class. Define a model for class tributes as a function of values from other attributes. The goal is records that were not seen before Previously determined a class as accurately as possible. A test data set is used to determine. accuracy of a model. Generally, the data set provided is divided into a set of training data and test data, where the training data is used to form the model and the test data is used to test it [13]. KNN algorithm has the advantage of being easy to apply so that the accuracy is high.

Based on the background in the description above, the problem can be formulated in this study as follows:

a. What are the pre-processing steps for feature selection data that can be done on the dataset Performance Results of Government Employees with Work Agreements in Kebumen Regency Government?

b. What are the results of the analysis on the classification method after the application of pre-processing of feature selection data on the dataset Performance Results of Government Employees with Work Agreements in the Kebumen Regency Government?

2. RESEARCH METHODS

2.1 Research Authenticity

Before this study was conducted. There are several studies that have the same fields and themes as the research to be carried out by researchers, including the following:

Research conducted by [6]. This study used the classification technique Nearest Centroid Neighbor Classifier Based on K Local Means Using Harmonic Mean Distance (LMKHNCN). This method is a modified method of the K-Nearest Neighbor (KNN) method and is proven to have better performance compared to the original KNN method. F1-Score and accuracy testing is carried out using K-Fold Cross Validation to determine the distribution of accuracy and also Testing of the effect of normalization because there is no normalization information in previous studies. The method in this case produces good classification performance, it is proven that the accuracy and F1-Score results by this method respectively reach 98.8% and 98.1%.

Another study examines performance appraisals with predefined forms. Every year an employee performance evaluation will be carried out to make a decision and consideration in knowing the performance obtained by the employee. The criteria for evaluating employee performance at the Education and Culture Office of Central Bengkulu Regency are Service Orientation, Integrity, Commitment, Discipline, Cooperation, and Leadership. Application of employee performance at the Education and Culture Office Central Bengkulu Regency was created using Visual Basic .Net programming language and SQL Server 2008 Database. Grouping employee data based on the results of performance appraisals that have been carried out, namely assessment orientation, integrity, commitment, discipline, cooperation, and leadership. The results of grouping employee performance appraisal data found that the number of employees in the high cluster (Cluster I) was 22% and the number of employees in the low cluster (Cluster II) was 82% [1]

Another study used the K-Nearest Neighbor (K-NN) algorithm method which is an algorithm that can be used to predict student achievement. Data normalization is required for parsing attribute values so that they are in a smaller range than the actual data. Feature selection is used to eliminate irrelevant features. Data cleansing of irrelevant data in the dataset aims to remove data that can interfere with the classification process. The classification process is carried out by cross-validation by dividing the data into training data by 80% and test data by 20% alternately as much as 5 fold. Euclidean, Manhattan, and Minkowski methods were used to measure the distance between two data. The formed classification model was tested using data separate from the training data and evaluated using a confusion matrix. As a result of the evaluation, an average accuracy of 95.85%, an average precision of 95.97%, was obtained and an average recall of 95.84% [16]

Another study is research that uses three methods, namely C4.5 algorithms, PCA combinations, and C4.5, as well as PCA, discretization, and C4.5 combinations to mine data. The discretization used is entropy-based discretization. In the pre-processing stage, SMOTE over-sampling technique was used to handle 4 training data that experienced class imbalance. In the application of a combination of PCA, discretization, and C4.5 algorithms, dimension reduction is carried out by using the PCA algorithm. The reduced data is discretized and then classified with the C4.5 algorithm [24]

Other research that can be mentioned is research that uses sociometric methods based on artificial intelligence. In this study, what was carried out was to analyze the work of a group in agencies in the form of software applications, making it easier to assess employee performance in realizing competitive advantages. The indicators used are responsibility, cooperation, work ability, appearance, relationship with customers, discipline of working time, and concern for the environment [1]

2.2 Theoritical Basic

Based on the Regulation of the Minister of State Civil Apparatus Empowerment Bureaucratic Reform Number 6 of 2022 concerning Performance Management of the State Civil Apparatus in Chapter 1 of the General Provisions states the meaning of the State Civil Apparatus, namely:

1. Employees of the State Civil Apparatus hereinafter referred to as employees are civil servants and government employees with work agreements appointed by civil service supervisory officials and assigned duties in a government position or assigned other state duties and paid based on applicable regulations;
2. Government Employees with Work Agreements, hereinafter abbreviated as PPPK, are Indonesian citizens who meet certain requirements based on work agreements for a certain period of time in order to carry out

government duties;

3. Employee Annual Performance Evaluation is a process by which the Performance Appraisal Officer reviews the overall work results and work behavior of Employees during one year of performance and determines the annual performance predicate of Employees based on the results of Employee performance;

Government Employees with Work Agreements (PPPK) as elements of the government that provide service to the state and services to the community in an integrated manner in accordance with a good, authoritative, efficient, clean, high-quality service mentality and aware of their responsibilities in carrying out their main duties and functions in government for development activities. The performance of Government Employees with Employment Agreements plays an important role in ensuring the improvement of Employee performance Government with Work Agreement, so by making predictions through classification of performance results will be able to provide results whether they are eligible to continue the work contract with the Kebumen Regency Government or not. Preprocessing is the first step after collecting data that will be used to train a classification model [23].

Pre-processing is carried out to improve data quality so that it will improve the accuracy value on the classification increases. If there is data that is not processed properly, it will lead to analysis results that lead to a wrong understanding of the data. This can have a negative effect on policy making based on the results of the analysis. Pre-processing of data that can be done in order to produce increased accuracy results is to conduct experiments to minimize missing data, High Dimensional Data, Imbalanced data and so on.

The problem that arises in the classification approach is the existence of missing data, namely the loss of value of an attribute due to errors in data collection, errors when entering data, and the inability of respondents to provide accurate answers. Because data classification is supervised, learning requires a complete dataset. [25]. Missing data can be overcome by deleting or imputing methods. Removing missing data can be done on attributes or per empty row of data, this method is done in certain cases, while imputation is a method to fill missing data in a dataset with estimated values that can be generated based on existing information.

High dimensional data is a critical research problem with serious implications in real-world problems [19]. The dataset will be said to be a High Dimensional Data condition when the dataset has a large number of features. One of the main characteristics of High Dimensional Data is its complexity. The large number of variables in High Dimensional Data can affect the efficiency and accuracy of statistical analysis and interpretation of the results. High Dimensional Data can be overcome with Feature Selection and Feature Extraction [28]

Correlation Matrix is a matrix that shows the correlation coefficient between attribute sets [25]. Each random attribute (Xi) in the table correlates with each of the other values in the table (Xj). Large values in this matrix indicate a serious relationship between the attributes involved. This makes it possible to know which pair has the highest correlation. The correlation value ranges from -1 to 1, where the value 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation, and a value of 0 indicates the absence of linear correlation between the two variables [11] The correlation matrix makes all attributes on the same basis, in which case every analysis can be described as an interdependent analysis. This matrix is square which can be described as n x n, where n is the number of variables in the dataset. The main diagonal of the correlation matrix is 1, because the correlation between the attribute and itself is 1.

The K-Nearest Neighbor (KNN) method is one of the basic methods derived from the instance-based learning group [12] The KNN algorithm classifies new data based on the learning distance closest to the object or commonly called the nearest neighbor. The KNN algorithm uses supervised learning where the results of the tested data are classified based on the closest membership of the most test data [31]. The number of the closest neighbors is called k. The value of k can be determined by the condition of n + 1 where n is the number of labels, or using the odd and even methods, if the value of n is odd then the value of k is even, and vice versa. The steps taken to determine the closest distance are, first the data is divided into training data and test data, after obtaining training data and test data then calculating the distance of each test data against training data [27].

2.3 Reseach Steps

The research was carried out in several stages, the stages of research can be seen in Figure 1.

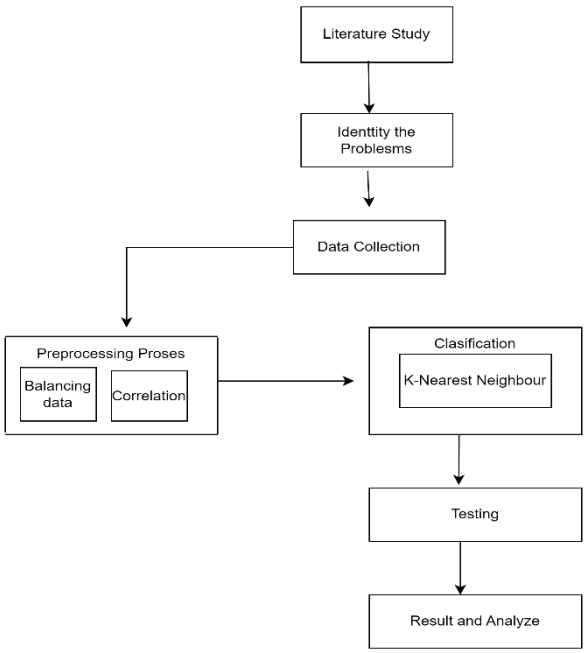

Fig 1: Research Stages Diagram

2.3.1 Literature Study

Research conducted by studying literature for example is journals and materials related to research topics. By conducting literature studies, journals were obtained regarding the classification of Civil Servant performance results, analysis, pre-processing of data on a dataset and machine learning classification algorithms**.**

2.3.2 Identify the Problem

The problem identification process is the process of looking for problems about what is happening that has a relationship with appropriate data pre-processing techniques to increase the accuracy value in the correlation analysis of the results of Government Employee performance with Work Agreements at the Kebumen Regency Government.

2.3.3 Data Collection

This study uses primary data from simpeg.kebumenkab.go.id from the Performance Results of Government Employees with Work Agreements at the Kebumen Regency Government in 2022. This dataset contains

data on the Performance Results of Government Employees with Work Agreements reported at the end of 2022 to early 2023. The dataset consists of 2,089 rows of data and 20 attributes including target variables These datasets have different types of features such as numerical and categorical. Some of the attributes contained in the Dataset of Performance Results of Government Employees with Work Agreements at the Kebumen Regency Government are presented in the Table 1

Table 1 Dataset Attributes of Performance Results of Government Employees with Work Agreements in Kebumen District Government

Num	Features	Num	Features
1.	Nama	11	Kerjasama
2.	NIP_baru	12.	Kepemimpinan
3.	Tanggal_lahir	13.	Lampiran_SKP
4.	Usia	14.	Perilaku
5.	Jenis_kelamin	15.	Capaian_Kinerja
6.	SKP	16.	Umpan_balik
7	Orientasi_Pelayanan	17.	Nama_OPD
8.	Integritas	18.	TMT_SK
9.	Komitmen	19.	Rekomendasi
10.	Disiplin	20.	Perpanjangan_Kontrak

Table 1 is a list of the attributes of the dataset used in the research. This attribute is listed in the data collection on Hasil Kinerja Pegawai Pemerintah dengan Perjanjian Kerja which will be carried out in 2022.

2.3.4 Data Preprocessing

The data preprocessing stage is carried out after the data collection process. Some of the preprocessing stage scenarios carried out are first balancing data, namely reducing the amount of data received to 232 and the second step is handling High Dimensional Data which can be see figure 2

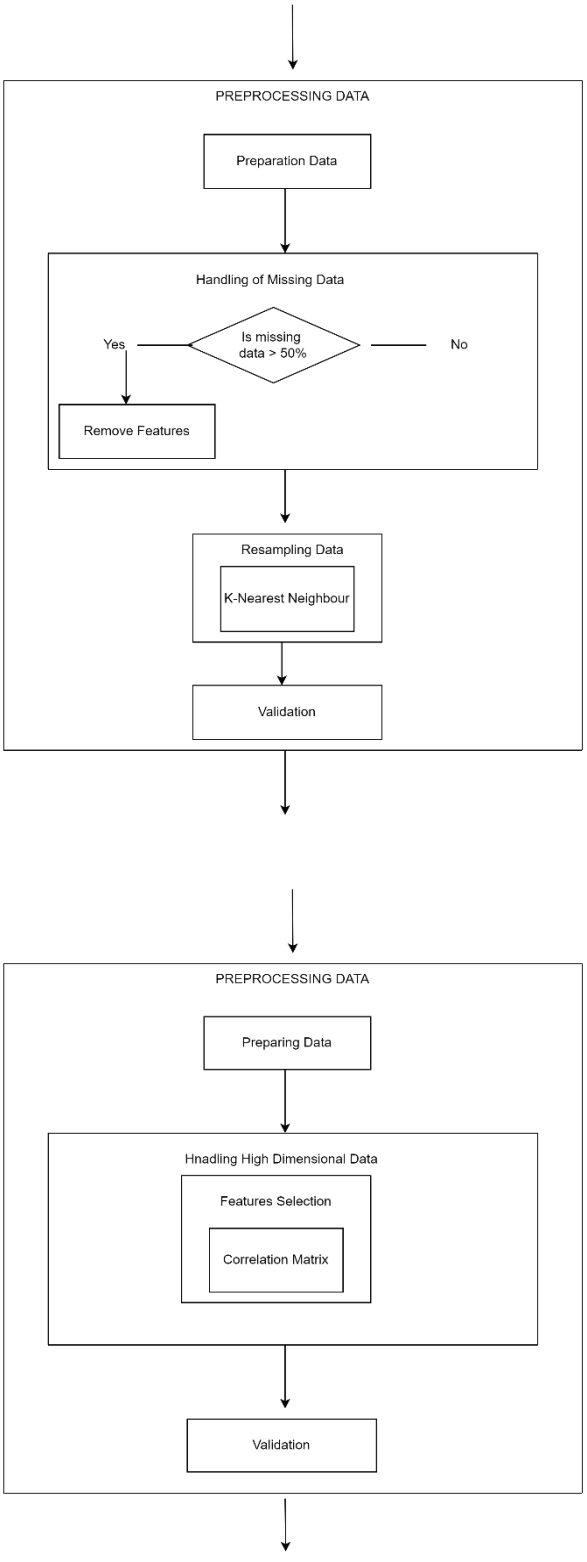

Fig 2: Preprocessing – Handling High Dimension Data

Figure 2 shows the preprocessing steps for handling High Dimension data. The steps taken are to prepare data, then handle High Dimension Data. Features Selection with Correlation Matrix and then perform the Validation process.

2.3.4.1 Data Preparation

Data preparation is carried out by labeling the data so that it can be processed using the K-Nearest Neighbor algorithm. Another

data preparation is to select attributes that will be removed because they are unique and irrelevant for the correlation analysis process and do not have predictive value for the next classification process.

2.3.4.2 Feature Selection
This study uses a feature selection method, namely the Correlation Matrix to analyze and identify the most relevant features and have a significant impact on the process of classifying datasets of performance results of Government Employees with Work Agreements at the Kebumen Regency Government. Features that have a high correlation will be removed, while features below number 1 will be analyzed.

2.3.5 Classification Process
The classification method used in this study is the K-Nearest Neighbor method. The classification process is carried out using Rapidminer. This research uses the set of operators you want to use and input into the Rapidminer tool to process data so that the data can be used optimally. Here are the operators that will be used.

2.3.5.1 Retrive (data)
This operator loads objects (data) into the process. This object is often an ExampleSet but can also be a stored model. Inside this operator is a dataset.

2.3.5.2 Set Roles
This operator classifies an attribute as a special attribute or a standard attribute.
2.3.5.3 Split data
Data split is a technique used to partition datasets is one of several aspects that affect how well a classification model performs on a machine learning Split Data algorithm [15].

2.3.6 Testing
After the classification process is carried out to obtain the results of the correlation analysis, testing is carried out from the classification model obtained. Testing is carried out with the Correlation Matrix to see the value of accuracy, precision and recall. The test scheme carried out is handling after Missing data, after handling High Dimensional Data, after handling Imbalanced data and the last test is testing after handling Missing data, High Dimensional Data and Imbalanced Data.

3. RESULT AND DISCUSSIONS

3.1 Result

In the chapter Research Results and Discussion, the results and analysis of the process above will be discussed. By using the Rapidminer tool which will involve using the K-Nearest Neighbor classification method in the dataset of Government Employee Performance Results with Work Agreements**.**

3.1.1 Data preparation
The data preparation carried out was labeling the Government Employee Performance Results Dataset with Work Agreements at the Kebumen Regency Government which was taken from Simpeg.kebumenkab.go.id which had 2,096 rows of data. In the data preparation step in the Government Employee Work Results Dataset with a Work Agreement, the labeling carried out is to change the data that is still in the form of strings and change the role as a label so that the K-Nearest Neighbor classification can be carried out. Labeling results that convert to label data are shown same as table 1.
The number of attributes in the dataset Performance Results of Government Employees with Work Agreements in Kebumen Regency Government is 19 attributes, and has one specific attribute for the next process. Another data preparation that is done is to eliminate categorical attributes to be processed in correlation analysis. The attributes omitted are "Name", "NIP_baru", "Tanggal_lahir", "Jenis_kelamin", "Leadership", "Performance Results Achievement", "OPD Name", "TMT Contract", and "Leadership Recommendation". The attributes of the dataset Performance Results of Government Employees with Work Agreements in the Kebumen Regency Government are 20 attributes, then as many as 9 attributes are removed because the attributes mentioned above have attributes that do not match the classification chosen in this study, so that the remaining 11 attributes to be tested in the next process.
Figure 3 presents the correlation results with the Correlation Matrix technique after removing 9 irrelevant attributes using the Rapidminer tool

Attribut...	Usia	skp	ORIENT...	INTEGR...	KOMIT...	DISIPLIN	KERJA...	Lampir...	Perilak...	Umpan ...
Usia	1	-0.031	-0.089	-0.088	-0.059	-0.083	-0.082	-0.254	-0.271	-0.271
skp	-0.031	1	0.791	0.794	0.809	0.796	0.795	0.610	0.587	0.568
ORIENT...	-0.089	0.791	1	0.987	0.951	0.985	0.984	0.445	0.472	0.451
INTEGRI...	-0.088	0.794	0.987	1	0.948	0.987	0.987	0.443	0.469	0.448
KOMITM...	-0.059	0.809	0.951	0.948	1	0.954	0.949	0.442	0.472	0.448
DISIPLIN	-0.083	0.796	0.985	0.987	0.954	1	0.987	0.444	0.468	0.446
KERJAS...	-0.082	0.795	0.984	0.987	0.949	0.987	1	0.441	0.466	0.444
Lampira...	-0.254	0.610	0.445	0.443	0.442	0.444	0.441	1	0.937	0.922
Perilaku ...	-0.271	0.587	0.472	0.469	0.472	0.468	0.466	0.937	1	0.984
Umpan ...	-0.271	0.568	0.451	0.448	0.448	0.446	0.444	0.922	0.984	1

Fig 3 : Results of Correlation Matrix

The intensity of the navy blue color in the correlation above shows that the level of correlation is high, in this case showing a strong relationship between the attributes in the dataset Government Employee Performance Results with Work Agreements in the Kebumen Regency Government.
From the matrix in Figure 3 after the attribute selection process based on the Correlation Matrix, from 20 attribute features reduced to 11 attribute features. Then the results are grouped into table 4.2 attributes – attributes that have a high correlation value of 0.9 which will be carried out correlation analysis with the combination of classifications selected in this study.

Table 2. Highes Correlation Value

Num	Atribute	Correlation Value
1	Orientasi pelayanan, Integritas	0,987
2.	Integritas, Disiplin	0,987
3	Integritas, Kerjasama	0,987
4	Kerjasama, Disiplin	0,987
5	Orientasi pelayanan, Disiplin	0,985
6	Orientasi Pelayanan, Kerjasama	0,984
7	Umpan balik, Perilaku	0,984
8	Komitmen, Disiplin	0,954
9	Orientasi Pelayanan, Komitmen	0,951
10	Komitmen, Kerjasama	0.949
11	Komitmen, Integritas	0,948
12	Lampiran SKP, Perilaku	0,937
13	Lampiran SKP, Umpan balik	0,922

3.1.2 K-nearest neighbor classification with correlation variation

This research conducts handling of High Dimensional Data can be handled by preprocessing steps using feature selection techniques based on the Correlation Matrix. This stage involves removing highly correlated features according to table 2 with the aim of reducing data dimensions and maintaining features that contribute significantly to dataset variability which accurate 100 %.

3.1.2.1 K-nearest neighbor classification with feature selection removes integrity attributes

Figure 4 presents the process of combining the results of the Correlation Matrix variation using the K-Nearest Neighbor classification using the Rapidminer tool with a split data ratio of 0.9 and 0.1.

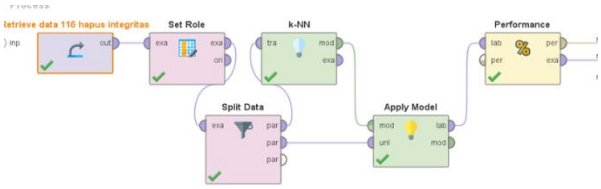

Fig 4 : The K-nearest neighbor classification process with feature selection removes integrity attributes

Figure 4 shows the K-Nearest Neighbor classification process for the dataset Performance Results of Government Employees with Work Agreements in the Kebumen Regency Government with an altered k value then selected, namely 5 and the dataset processed is a dataset resulting from feature selection processing from 11 attribute features reduced to 10 attribute features with a Correlation Matrix. By removing the attribute with a correlation value above 0.9, namely the "Integrity" attribute.

Table 3. K-nearest neighbor classification results with feature selection correlation variations eliminate integrity attributes

Accuracy : 100 %

	True ya	True tidak	Class precision
Pred. Ya	0	0	0.00 %
Pred. tidak	0	23	100 %
Class recall	0.00 %	100 %	

Table 3 shows the results of the K-Nearest Neighbor classification with correlation variations using the Correlation Matrix by removing the "Integrity" attribute with an accuracy value of 100%.

3.1.2.2 K-nearest neighbor classification by feature selection removes disciplinary attributes

In Figure 5 presents correlation analysis using the Correlation Matrix technique with K-Nearest Neighbor classification by removing the "Discipline" attribute in the dataset Performance Results of Government Employees with Work Agreements in the Kebumen Regency Government. The number of attribute features from 11 attributes to 10 attributes applies further to this correlation analysis research with the value k = 5

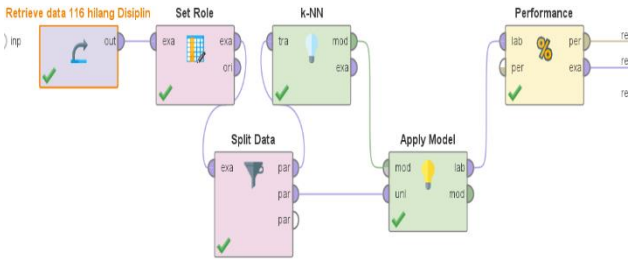

Fig 5 : The K-nearest neighbor classification process with variation in feature selection correlation removes disciplinary attributes

In Table 4 shows the results of the K-Nearest Neighbor classification with correlation variations using the Correlation Matrix by removing the "Discipline" attribute with an accuracy value of 100%.

Table 4. K-nearest neighbor classification results with feature selection correlation variations eliminate disciplinary attributes

Accuracy : 100 %

	True ya	True tidak	Class precision
Pred. Ya	0	0	0.00 %
Pred. tidak	0	23	100 %
Class recall	0.00 %	100 %	

3.1.2.3 K-nearest neighbor classification with feature selection removes the cooperation attribute

In Figure 6 shows the correlation analysis process using the Correlation Matrix technique with K-Nearest Neighbor classification by removing the "Cooperation" attribute in the dataset Performance Results of Government Employees with Work Agreements in the Kebumen Regency Government. The number of attribute features from 11 attributes to 10 attributes applies further to this correlation analysis research with the value k = 5

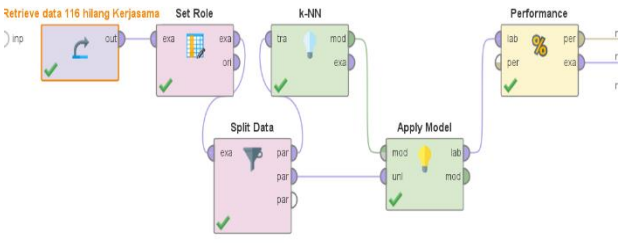

Fig 6 : K-nearest neighbor classification process with variations Correlation Selection Features Eliminate Cooperation Attributes

The results of accuracy, precision and reacall of the K-Nearest Neighbor classification process with variations in feature selection removing the Cooperation attribute are shown in Table 5.

Table 5. K-nearest neighbor classification results with feature selection correlation variations eliminate cooperation attributes

Accuracy : 100 %

	True ya	True tidak	Class precision
Pred. Ya	0	0	0.00 %
Pred. tidak	0	23	100 %
Class recall	0.00 %	100 %	

Table 5 shows the results of the classification of K-Nearest Neighbor with correlation variations using the Correlation Matrix by removing the "Cooperation" attribute with an accuracy value of 100%.

3.1.2.4 K-nearest neighbor classification by feature selection removes service orientation attribute

Figure 7 shows the correlation analysis process using the Correlation Matrix technique with K-Nearest Neighbor classification by removing the "Service Orientation" attribute in the dataset of Government Employee Performance Results with Work Agreements in Kebumen District Government. The number of attribute features from 11 attributes to 10 attributes applies further to this correlation analysis research with the value k = 5

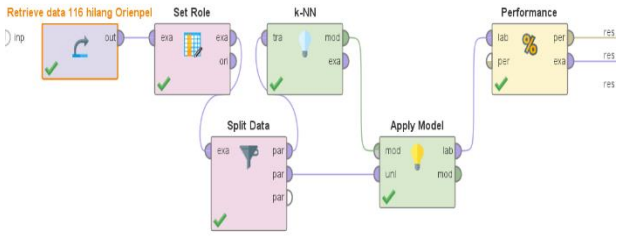

Fig 7 : K-nearest neighbor classification process with variations Feature selection correlation removes service orientation attributes

Table 6 The results of the K-nearest neighbor classification with variations in feature selection correlation eliminate service orientation attributes

Accuracy : 100 %

	True ya	True tidak	Class precision
Pred. Ya	0	0	0.00 %
Pred. tidak	0	23	100 %
Class recall	0.00 %	100 %	

3.1.3 K-nearest neighbour classification with correlation combination

3.1.3.1 The K-nearest neighbor classification with a combination of correlations removes the attributes of service orientation and integrity.

Figure 8 shows the correlation analysis process using the Correlation Matrix technique with K-Nearest Neighbor classification with a combination of removing the "Service Orientation" and "Integrity" attributes in the dataset of Government Employee Performance Results with Work Agreements in Kebumen District Government. The number of attribute features from 11 attributes to 1 special attribute and 8 regular attributes with value k = 5

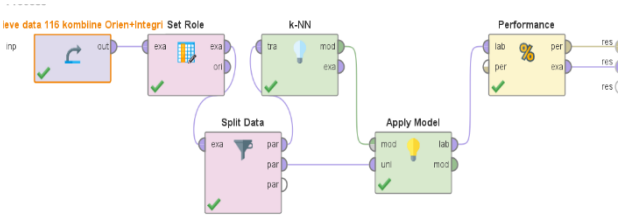

Fig 8: K-nearest neighbor classification process with combination Correlation Removes Service Orientation and Integrity Attributes

Table 7 . K-nearest neighbor classification results with combination Correlation Removes Service Orientation and Integrity Attributes

Accuracy : 100 %

	True ya	True tidak	Class precision
Pred. Ya	0	0	0.00 %
Pred. tidak	0	23	100 %
Class recall	0.00 %	100 %	

Table 7 shows the result of the accuracy of the K-Nearest Neighbor classification process with a combination of correlations removing the attributes "Service Orientation" and "Integrity" is 100%.

For results with 100% accuracy can be seen in the next table in the final result,

3.1.4 K-Nearest Neighbor Classification After Overcoming High Dimensional Data

This research uses preprocessing methods, including using Correlation Matrix as a correlation analysis technique, and K-Nearest Neighbor for classification methods. Preprocessing results on the dataset Performance Results of Government Employees with Work Agreements at the Kebumen Regency Government using 20 features and 2,096 rows of data. The next process uses the K-Nearest Neighbor classification with variations in the correlation results with a high value of 0.9 and with a k value in the classification algorithm = 5. The results of the accuracy of the correlation analysis experiment scenario with classification are shown in Table 8.

Table 8 Classification Results After Conducting Correlation Analysis

Number	Omitted Attribute Name	Accuracy Results
1	Integritas	100 %
2	Komitmen	95.65 %
3	Disiplin	100 %
4	Kerjasama	100 %
5	Perilaku	95.65 %
6	Umpan Balik	95.65 %
7	Lampiran SKP	95.65 %
8	Orientasi Pelayanan	100 %
9	Orientasi Pelayanan dan Integritas	100 %
10	Integrasi dan Disiplin	100 %
11	Integrasi dan Kerjasama	100 %
12	Kerjasama dan Disiplin	100 %
13	Orientasi Pelayanan dan Disiplin	100 %

14	Orientasi Pelayanan dan Kerjasama	100 %
15	Umpan Balik dan Perilaku	95.65 %
16	Komitmen dan Disiplin	100 %
17	Orientasi Pelayanan dan Komitmen	100 %
18	Komitmen dan Kerjasama	100 %
19	Komitmen dan Integritas	100 %
20	Lampiran SKP dan Perilaku	95..65 %
21	Lampiran SKP dan Umpan Balik	95.65 %

The classification results presented in the K-Nearest Neighbor classification results tables with variations in feature selection correlation still show a high degree of accuracy, although the dataset becomes more complex after the application of several correlation analyses. However, there is a decrease in accuracy during the classification process with feature selection techniques by eliminating attributes with high correlation values. This can be due to the complexity of the reduction in the dataset after correlation. However, the accuracy value has decreased. This stage of correlation analysis has the effect of providing benefits in handling High Dimensional Data problems using correlation techniques

3.2 Discussions

This study conducted an experimental scenario process using the Preprocessing data method and correlation analysis using the K-Nearest Neighbor classification in the dataset of Government Employee Performance Results with Work Agreements at the Kebumen Regency Government in 2022. Overall from the scenarios of the analysis process, the highest accuracy obtained is around 100% while the lowest accuracy obtained is 95.65%.

The correlation analysis scenario step involves feature selection using the Correlation Matrix. Dimensionality reduction of data can be done by eliminating features that have high correlation. Although accuracy is still high, there is a decrease of about 4-5% after conducting correlation analysis by eliminating attributes that have high correlation values and applying the K-Nearest Neighbor classification. The results vary but still get a high accuracy value. In this case it shows a trade off between dimensionality reduction and maintenance of important information, and also the technique of removing features based on correlation alone does not necessarily improve model performance.

The overall results showed that correlation analysis with the Correlation Matrix technique and applying the K-Nearest Neighbor classification showed superiority in accuracy, and showed the effectiveness of attributes in handling which indicators were relevant in the dataset Performance Results of Government Employees with Work Agreements at the Kebumen District Government. The highest accuracy is obtained in scenarios with Correlation Analysis with the Correlation Matrix technique and applying the K-Nearest Neighbor classification, which is 100%.

4. CONCLUSION

Based on the results of tests that have been carried out using several scenarios to determine the effect of the correlation carried out, the results can be concluded as follows:

1. This study focuses on data pre-processing techniques to support the correlation analysis of Government Employee Performance Results dataset with Work Agreement on the K-Nearest Neighbor classification model.

2. The aspects carried out in this study involve the process of data preparation, handling Missing Data, Feature Selection and Correlation.

3. Correlation analysis with K-Nearest Neighbor classification is carried out through several different data pre-processing scenarios, namely:

a. Preprocessing with Removing Features

b. Preprocessing with Feature Selection using Correlation Matrix.

c. Combination of Preprocessing Removing Features, Feature Selection and K-Nearest Neighbor.

4. The results of this study show that the use of a combination of Pre-Processing data methods in the Correlation analysis on the dataset of Government Employee Performance Results with Work Agreements in the Kebumen Regency Government has a significant influence in producing high accuracy scores and which features are relevant to use.

5. The highest accuracy is achieved in the Preprocessing step with Removing Features, which is with an accuracy of 100 %.

6. There are several trade-offs between correlation and accuracy, so these findings provide useful insights to improve the effectiveness of relevant correlation analysis in assessing the performance results of Government Employees with Work Agreements at the Kebumen District Government

5. ACKNOWLEDGMENTS

Our thanks to Badan Kepegawaian dan Pengembangan Sumber Daya Manusia on Kebumen Regency

6. REFERENCES

[1] Andrika P.D., RD Kusumanto, R. K., Dimas, MD., & Anisah, M. 2022. Aplikasi Penilaian Kinerja Pegawai dengan Metode Sosiometri Berbasis Artificial Intelegence. *Journal Locus Penelitian Dan Pengabdian*, *1*(5), 348–360. https://doi.org/10.58344/locus.v1i5.90

[2] BPK, B.P. 2020. Pp_No46 tahun 2011(pasal.1). https://peraturan.bpk.go.id/Home/Details/135059/pp-no-21-tahun-2020

[3] Benhar, H. I.-A. 2020. Benhar, H., Idri, A., & Fernández-A J. L. 2020. Data preprocessing for heart disease classification: A systematic literature review. Computer Methods and Programs in Biomedicine, 195, 105635.

[4] Dahj, J. N. M., & Ogudo, K. A. 2023. Machine Learning-Based Imputation Approach with Dynamic Feature Extraction for Wireless RAN Performance Data Preprocessing. *Symmetry*, *15*(6). https://doi.org/10.3390/sym15061161

[5] Endang ED. S. 2020. IMPLEMENTASI DATA MINING MENGGUNAKAN ALGORITME NAIVE BAYES CLASSIFIER DAN C4.5 UNTUK MEMPREDIKSI KELULUSAN MAHASISWA. *TELEMATIKA*, 56-67.

[6] Hidayatullah, A. S., Bachtiar, F. A., & Cholissodin, I. 2021 Penerapan Algoritme Nearest Centroid Neighbor

Classifier Based on K Local Means Using Harmonic Mean Distance (LMKHNCN) Untuk Klasifikasi Hasil Kinerja Pegawai Negeri Sipil. Jurnal Teknologi Informasi Dan Ilmu Komputer, 8(6), 1287. https://doi.org/10.25126/jtiik.202183443

[7] 1Ilmaniati, A., & Putro, B. E. 2019. Analisis komponen utama faktor-faktor pendahulu (antecendents) berbagi pengetahuan pada usaha mikro, kecil, dan menengah (UMKM) di Indonesia. *Jurnal Teknologi*, *11*(1), 67–78. https://jurnal.umj.ac.id/index.php/jurtek/article/view/2652

[8] Hapsari R.K dan Indriyani. 2022. implementasi algoritma smote sebagai Penyelesaian Imbalance Hight Dimensional Datasets.

[9] Hermawan, A., & Permana W.A. (2022). Implementasi Korelasi untuk Seleksi Fitur pada Klasifikasi Jamur Beracun Menggunakan Jaringan Syaraf Tiruan. 5. https://www.kaggle.com/uciml/mushroom-

[10] Joan D & VIncent .R. 2020. *The Design of Rijndael: AES - The Advanced Encryption Standard*. Springer Berlin Heidelberg: https://doi.org/10.1007/978-3-662-60769-5_6.

[11] Kripsiandita, Y., Arifianto, D., & A'yun, Q. 2021. Deteksi Gangguan Autis Pada Anak Menggunakan Metode Modified K-Nearst Neighbor. *JUSTINDO (Jurnal Sistem Dan Teknologi Informasi Indonesia)*, *6*(1), 31–37. https://doi.org/10.32528/justindo.v6i1.4357

[12] Lestari, P. I., Ratnawati, D. E., & Muflikhah, L. 2019. Implementasi Algoritme K-Means Clustering Dan Naive Bayes Classifier Untuk Klasifikasi Diagnosa Penyakit Pada Kucing. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *3*(1), 968– 973

[13] Oktara, P., Yulianti, L., & Fredricka, J. 2021. Analisis Kinerja Pegawai Menggunakan Algoritma K-Means Pada Dinas Pendidikan Dan Kebudayaan Kabupaten Bengkulu Tengah. *Jurnal Media Infotama*,

[14] Prasetyawan, D., & Gatra, R. 2022. Algoritma K-Nearest Neighbor untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi JISKA (Jurnal Informatika Sunan Kalijaga), 7(1), 56–67. https://doi.org/10.14421/jiska.2022.7.1.56-67

[15] Oktafiani, R., Hermawan, A., & Avianto, D. 2023. Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning. *Jurnal Sains Dan Informatika*, 19–28. https://doi.org/10.34128/jsi.v9i1.622

[16] Joshi, A. P., & Patel, B. V.2021. Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Oriental Journal of Computer Science and Technology*, *13*(0203), 78–81. https://doi.org/10.13005/ojcst13.0203.03

[17] Khakim, E. N. R., Hermawan, A., & Avianto, D. 2023. IMPLEMENTASI CORRELATION MATRIX PADA KLASIFIKASI DATASET WINE. *JIKO (Jurnal Informatika Dan Komputer)*, *7*(1), 158. https://doi.org/10.26798/jiko.v7i1.771

[18] Khan, S. I., & Hoque, A. S. M. L.2020. SICE: an improved missing data imputation technique. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00313-w

[19] Muneer, A., Mohd Taib, S., Mohamed Fati, S., O. Balogun, A., & Abdul Aziz, I.2022 A Hybrid Deep Learning-Based Unsupervised Anomaly Detection in High Dimensional Data. *Computers, Materials & Continua*, *70*(3), 5363–5381. https://doi.org/10.32604/cmc.2022.021113

[20] Nugroho, B., & Denih, A. 2020. Perbandingan kinerja metode pra-pemrosesan dalam pengklasifikasian otomatis dokumen paten. *Komputasi: Jurnal Ilmiah Ilmu Komputer dan Matematika*, *17*(2), 381-387.

[21] Nazabal, A., Williams, C. K. I., Colavizza, G., Smith, C. R., & Williams, A. 2020. *Data Engineering for Data Analytics: A Classification of the Issues, and Case Studies*. http://arxiv.org/abs/2004.12929

[22] Oğuz, M. B. 2020 Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *https://doi.org/10.1016/j.specom.2019.12.001*.

[23] Rahmawan, H. 2020. Penentuan Rekomendasi Pelatihan Pengembangan Diri Bagi Pegawai Negeri Sipil Menggunakan Algoritma C4.5 Dengan Principal Component Analysis Dan Diskritisasi. Jurnal Tekno Kompak, 14(1), 5. https://doi.org/10.33365/jtk.v14i1.531

[24] R Kurniawan, P Pizaini, F Insani. 2021. Penerapan Algoritma K-Means Clustering dan Correlation Matrix untuk Menganalisis Risiko Penyebaran Demam Berdarah di Kota Pekanbaru. *ejurnal.unmerpas.ac.id*, 1-6.

[25] Seto A,A., Dewi S,K., Agustianto, K., Wiryawan, G., & Jember, P. N. (n.d.). *PENGARUH PREDIKSI MISSING VALUE PADA KLASIFIKASI DECISION TREE C4.5*. https://doi.org/10.25126/jtiik.202294778

[26] Setianto, Y. A., Kusrini, K., & Henderi, H. 2019 Penerapan Algoritma K- Nearest Neighbour Dalam Menentukan Pembinaan Koperasi Kabupaten Kotawaringin Timur. Creative Information Technology Journal, 5(3), 232. https://doi.org/10.24076/citec.2018v5i3.179

[27] Shen Y, Elke H & Qiong B & Tom B & Geert. 2020. Towards better road safety management: Lessons learned from inter-national benchmarking. *Accident Analysis & Prevention*

[28] Tuhina Banerjee, P. R. 2021. Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*.

[29] Vincenzo Moscato Giancarlo Sperlí Antonio Picariello. 2021 A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*.

[30] Wahyudi, M. 2022. Diagnosa Gejala Kecanduan Game Online Dengan Metode K-Nearest Neighbor. Seminar Nasional Informatika (Senatika), 6(3), 106– 117.*17*(2), 341139.

[31] Wets, Y. S. 2020. Towards better road safety management: Lessons learned from inter-national benchmarking. *Accident Analysis & Prevention*.