JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 7, No. 2. Maret 2025 DOI: https://doi.org/10.34288/jri.v7i2.321
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

PREDICTION OF PIP RECIPIENTS USING K-NEAREST NEIGHBOR AT MI NURUL QOLBI

Dea Fitra Ningrum[-1], Desti Fitriati[-2]

Informatics Engineering Study Program[1,2]
Pancasila University[1,2]
South Jakarta City, Special Capital Region of Jakarta, Indonesia
4520210023@univpancasila.ac.id[-1], desti.fitriati@univpancasila.ac.id[-2]

Abstract

Education is a key foundation in the development of quality human resources. However, the rising cost of education makes some children unable to attend school due to their parents' financial limitations. To address this problem, the government launched the Indonesia Smart Program (PIP) which provides education funding assistance to eligible students. This research aims to develop an Information System that can predict the eligibility of students to receive PIP assistance using the K-Nearest Neighbors (KNN) algorithm. The data used comes from all students of Madrasah Ibtidaiyah (MI) Nurul Qolbi in the 2022-2023 school year. This research methodology involves testing with a value of k=13 and model evaluation is done using split ratio and cross-validation techniques. The results showed an accuracy of 98.98% from various split ratios (10:90, 20:80, 30:70, 40:60) and an accuracy of 99.24% using the 10-fold cross-validation technique. The accuracy results show excellent performance and provide important significance in the development of prediction systems to help the selection process of aid recipients more efficiently and reduce the administrative burden for schools. However, its application on a wider scale still requires further research, especially to test its consistency and effectiveness in different contexts and with more diverse datasets.

Keywords: Education; Prediction; PIP; K-Nearest Neighbors; Cross Validation

Abstrak

Pendidikan merupakan fondasi utama dalam pengembangan sumber daya manusia yang berkualitas. Namun, meningkatnya biaya pendidikan membuat sebagian anak-anak tidak dapat bersekolah karena keterbatasan finansial orang tua mereka. Untuk mengatasi masalah ini, pemerintah meluncurkan Program Indonesia Pintar (PIP) yang memberikan bantuan dana pendidikan kepada siswa yang memenuhi syarat. Penelitian ini bertujuan untuk mengembangkan sebuah Sistem Informasi yang dapat memprediksi kelayakan siswa menerima bantuan PIP menggunakan algoritma K-Nearest Neighbors (KNN). Data yang digunakan berasal dari seluruh siswa Madrasah Ibtidaiyah (MI) Nurul Qolbi tahun ajaran 2022-2023. Metodologi penelitian ini melibatkan pengujian dengan nilai k=13 dan evaluasi model dilakukan menggunakan teknik split ratio dan validasi silang (cross-validation). Hasil penelitian menunjukkan akurasi sebesar 98,98% dari berbagai split ratio (10:90, 20:80, 30:70, 40:60) dan akurasi sebesar 99,24% menggunakan teknik 10-fold cross-validation. Hasil akurasi menunjukkan kinerja yang sangat baik dan memberikan signifikansi penting dalam pengembangan sistem prediksi untuk membantu proses seleksi penerima bantuan secara lebih efisien dan mengurangi beban administratif untuk pihak sekolah. Namun penerapannya pada skala yang lebih luas masih memerlukan penelitian lebih lanjut, terutama untuk menguji konsistensi dan efektivitasnya dalam konteks yang berbeda dan dengan dataset yang lebih beragam.

Kata kunci: Pendidikan; Prediksi; Program Indonesia Pintar (PIP); K-Nearest Neighbors; Validasi Silang

INTRODUCTION

Education plays a very important role in the development of quality human resources (Penulis, 2023) . One way for the government to fulfill this is by implementing a 12-year compulsory education program. The goal is to ensure that children in Indonesia receive sufficient basic education. The increasingly expensive cost of education keeps some children out of school

because their parents cannot afford to pay tuition fees (Puslapdik, 2022) . To overcome this, the government created one of the education programs, the Smart Indonesia Program or Program Indonesia Pintar (PIP).

According to information from the Education Financing Service Center (PUSLAPDIK), PIP funds for 2023 reached 9.6 trillion Rupiah that has been distributed nationally. The funds are intended to support 18,109,119 students who are at various levels of education, including elementary, junior high, high school, and vocational school (Puslapdik, 2023a) . PUSLAPDIK also noted that a total of 1.8 trillion Rupiah has been allocated to support the education of 3.4 million students at various levels of education in West Java province (Puslapdik, 2023b).

Although PIP is designed to help students from underprivileged families, determining beneficiaries is often a challenge for educational institutions. One of them is Madrasah Ibtidaiyah (MI) Nurul Qolbi, which receives funds from this program. Manual determination of recipients is not only time-consuming, but also risks bias because it involves many factors, such as students' economic circumstances, family conditions, and other aspects that are difficult to measure objectively. Therefore, a system is needed that is able to predict the eligibility of students to receive PIP assistance more accurately and efficiently.

On this issue, research has been conducted by Angga Pebdika (Pebdika, Herdiana, & Solihudin, 2023) , where research was conducted using the naïve bayes method to predict PIP recipients with an accuracy of 88.89%. However, the use of naïve bayes has drawbacks, especially when facing zero conditional probabilities, which can cause errors in prediction (Pebdika et al., 2023) , as well as many gaps that reduce the effectiveness of this method. This will certainly affect the accuracy of the research. Therefore, this study chose to use the K-Nearest Neighbor (KNN) method. Apart from its ability to classify objects based on their proximity, KNN was also chosen because of its simplicity of implementation and flexibility in handling various types of data without requiring certain distribution assumptions. This is important, because student data has the possibility of irregular distribution or does not follow a certain pattern (Rahmadini, LorencisLubis, Priansyah, N, & Meutia, 2023).

RESEARCH METHODS

This research aims to develop a prediction system for the eligibility of Indonesia Smart Program (PIP) beneficiaries using the K-Nearest Neighbors (KNN) method. KNN was chosen because of its ability to classify data based on the closeness of the distance between data points, which can handle non-linear and diverse data (Priyambodo, Nugroho, & Zaman, n.d.) . The main advantages of KNN are the ease of interpretation of results and the ability to provide transparent decisions (Sumiah & Mirantika, 2020) , which is very important in the context of selecting recipients of educational assistance.

Time and Place of Research

The research was conducted for 5 months, starting from March 2024 to July 2024. The research was conducted by analyzing and calculating the data obtained.

Machine Learning

Machine learning belongs to the field of artificial intelligence (AI), which is a field of science that involves the design and development of algorithms that allow computers to develop behavior based on information or facts obtained. In this context, the term “machine” refers more to a system (Ibnu Daqiqil, 2021). Machine Learning is often described as a learning process that involves experience or is done without direct human supervision (Baharuddin, Azis, & Hasanuddin, 2019). The main focus in machine learning is how to automatically identify complex patterns and make intelligent decisions based on data. Machine learning explores building systems based on learning from data, or is the study of how to program computers to learn (Rahmadini et al., 2023).

Classification

Classification is a way of grouping data into one or more predefined classes or the stage of finding a model that describes and distinguishes classes of data (Nata & Royal, 2022) . Some commonly used classification techniques include Neural methods, Rough sets, K-Nearest Neighbor, Bayesian Classifiers, and various other techniques (Yandi Saputra & Primadasa, 2018). Classification is one of the data mining methods that has a function in making predictions. Classification can be defined as a prediction function that predicts and categorizes a data item into a certain class. The classification process involves using a set of training data that already has a predetermined

class, as well as its previously known characteristics (Khoirunnisa, Susanti, Rokhmah, & Stianingsih, 2021) . Evaluation of classifier performance is generally done by measuring the level of accuracy (Anwar Pauji, Aisyah, Surip, Saputra, & Ali, 2022).

Prediction

Prediction is a systematic process of estimating future events based on past and current information to reduce forecast errors (Kushartanto & Aldisa, 2023). Predictions do not have to be exact, but can provide the most accurate answer possible to something that will happen. Prediction shows what will happen to a situation and acts as input in the planning and decision-making process (Rahmadini et al., 2023).

K-Nearest Neighbor

K-Nearest Neighbor (KNN) is one of the most basic supervised learning algorithms in classification (Noviana, Susanti, & Susanto, 2019) . KNN makes predictions by calculating the similarity between objects in the training data and the object being tested, measured using a distance function, the longer the distance means the less similar the tested data is to the training data (Ibnu Daqiqil, 2021). The working principle of KNN is to find the closest distance between the data to be evaluated and its k nearest neighbors in the training data (Anwar Pauji et al., 2022) . The selection of the optimal k value is very important because it affects the accuracy and reliability of the model (Winarno, 2023). In general, higher k values can reduce the impact of noise on the classification process, but can also blur the boundaries between classes. Therefore, the optimal k value is chosen by considering the balance between bias and variance, and is usually an odd number to avoid balanced voting results (Ibnu Daqiqil, 2021) . To ensure that the optimal k-value is selected, techniques such as cross-validation are used to reduce the risk of overfitting and ensure that the model can be generalized to other data.

One of the distance calculation methods commonly used in the calculation of the KNN algorithm is to use the euclidean distance calculation (Cholil, Handayani, Prathivi, & Ardianita, 2021) . Euclidean distance is related to the Phytagoras theorem. This calculation will be done by calculating the straight line distance between two points (Ibnu Daqiqil, 2021). The calculation of euclidean distance is using equation 1 (Yandi Saputra & Primadasa, 2018):

$$euc = \sqrt{(p_i - q_i)^2 + \ldots + (p_i - q_i)^2} = \sqrt{(\sum_{i=1}^{n} (p_i - q_i)^2)} \ldots (1)$$

Keterangan:
p_i = Data Training
q_i = Data Testing
i = Variable Data
n = Dimension Atributtes

The steps to calculate the K-Nearest Neighbor method involve several stages, namely (Yandi Saputra & Primadasa, 2018):

1. Setting the value of parameter K (number of nearest neighbors), parameter K in testing is determined based on the optimum K value during training.
2. Calculating the distance between training data (training) and test data (testing) using the euclidean distance calculation.
3. Sort the objects into groups that have the smallest distance.
4. Map the corresponding class.
5. Identifies the number of nearest neighbor classes and sets that class as the class to be evaluated on the data.

Procedure

The stages of this research begin with a literature study to understand the priorities of PIP recipients. After that, student data was collected based on criteria such as participation in PKH, ownership of KKS, SKTM, and KIP (Ramdhani, 2023) . The data is then processed, separated into training and test data, and used to build a predictive model with the K-Nearest Neighbor method. This stage includes determining the optimal k value and calculating the probability of PIP recipients based on existing attributes. The following is an explanation of each stage of this research:

1. Literature Study
   Literature study or literature review is carried out to find out about the Smart Indonesia Program such as the right and necessary criteria to be used as attributes in making predictions.

2. Data Collection
   The data used in this study comes from the results of distributing questionnaires to parents of Madrasah Ibtidaiyah (MI) Nurul Qolbi students in the 2022-2023 school year. The data is processed by the school and presented in tabular form with Excel format. There are 404 data entries containing student information, including student names, ownership of Indonesia Smart Card (KIP),

Prosperous Family Card (KKS), participation in the Family Hope Program (PKH), ownership of a Certificate of Disadvantage (SKTM), information regarding the amount of parental income, number of dependents, and eligibility status of PIP recipients.

3. Preprocessing
After the data collection stage, the next step is the data processing or preprocessing stage. The purpose of this stage is to ensure that the data is ready to be used for the modeling stage so that it can provide more accurate results. At this stage, checking for missing values and duplicate data is done. The next stage is transformation, where the data is converted into numeric types as needed. Then perform standardization or normalization, which aims to ensure that the data is ready to be used in the modeling stage with a consistent scale. The last stage is to carry out the data division process, where the feature variables and target variables are separated, then the data is divided into test data (training) and training data (testing). This data division is done with a variety of ratios such as 90:10, 80:20, 70:30, or 60:40.

4. Modeling Using K-Nearest Neighbor (KNN)
After data processing, the next step is to apply the K-Nearest Neighbor algorithm to the dataset to be used. In this research, the library from Scikit Learn will be used for the modeling stage. The first step is to determine the value of k. After the k value is determined, a prediction is made on the new data using the model that has been created with the k value. From the prediction results, information can be obtained regarding the eligibility of students, whether they are eligible or not to receive education funding assistance.

5. Model Evaluation
After obtaining the accuracy level, the next step is to conduct an evaluation stage to determine the reliability of the model using cross-validation techniques. In cross-validation, the dataset is divided into subsets, and the model is trained and tested repeatedly using different subsets as test and training data. This process helps ensure that the model is not only effective on one particular dataset, but can also generalize to data not seen before.

6. Putting the model into a pickle
After finding the best model, the next step is to save it in the form of a pickle. The goal is that the model that has been created can be used on the website to be created.

7. System development or creation
This stage is the stage for making a website-based system to predict students receiving PIP. Making this system uses the Streamlit application with the Python Programming Language.

8. System testing
After the web-based system is created, the system will then be tested by entering student data to be predicted. After success, the system will produce prediction results from the data.

Data, Instruments, and Data Collection Techniques

The data used in this study are data on all MI Nurul Qolbi students in 2022-2023 by conducting interviews with the management of the Indonesia Indonesia Pintar Program at MI Nurul Qolbi regarding data. After that the school distributed questionnaires in order to get the data needed for research. The questionnaire data is processed by the school and then presented in table form with excel format. The data has 404 rows containing student data and 6 feature columns containing the variables to be predicted. Table 1 is an explanation of the features that will be used.

Table 1. Data Description

No.	Variable Name	Description
1.	KIP	An indicator of whether the student has a *Kartu Indonesia Pintar* (KIP), which indicates receipt of educational assistance from the government.
2.	KKS	An indicator of whether the student's family has a Prosperous Family Card (KKS), which indicates the family's welfare status.
3.	PKH	An indicator of whether the student's family is part of the Family Hope Program (PKH), a government social

		assistance program.
4.	SKTM	An indicator of whether the student's family has a Certificate of Disadvantage (SKTM), which indicates the family's economic status.
5.	Penghasilan	Information about the amount of income of the student's parents.
6.	Tanggungan	The number of family members who are dependent on the student's parents.

RESULTS AND DISCUSSION

After obtaining the data to be used for research, the next stage is data preprocessing. Figure 1 is the sample data used in the research.

	Nama	KIP	KKS	PKH	SKTM	Penghasilan	Tanggungan	Status
0	Alea Kinanti Putri	Tidak	Tidak	Tidak	Tidak	> Rp 6.000.000	4	Tidak Layak
1	Muhammad Sahal	Tidak	Tidak	Tidak	Ya	Rp 3.000.000 - Rp 5.999.999	5	Tidak Layak
2	Muhammad Jibril Adriansyah	Tidak	Tidak	Tidak	Tidak	Rp 1.000.000 - Rp 2.999.999	4	Tidak Layak
3	Fara Nisa Putri Anzani	Tidak	Tidak	Tidak	Tidak	Rp 1.000.000 - Rp 2.999.999	2	Tidak Layak
4	Ayasha Alzea Humaira	Tidak	Tidak	Tidak	Tidak	> Rp 6.000.000	3	Tidak Layak

Figure 1. Research Data Sample

The first stage of data processing will look at the missing values from the dataset. Missing values in the data can occur due to errors in data collection and input or inability to obtain complete information. Dealing with missing values is very important as it will affect the results of the analysis. In this research, the process of checking missing values is carried out using the pandas library. In this research, there are several missing values from the dataset. The missing values can be seen in Figure 2.

```
missing_values = dataset.isnull().sum()

# Menampilkan jumlah nilai yang hilang
print("Jumlah Nilai yang Hilang (Missing Values):")
print(missing_values)
```

```
Jumlah Nilai yang Hilang (Missing Values):
Nama            0
KIP             0
KKS            79
PKH            69
SKTM           79
Penghasilan    52
Tanggungan     91
Status          0
dtype: int64
```

Figure 2. Missing Values in Dataset

In Figure 2, there are 79 students who did not fill in the KKS column, 69 students who did not fill in the PKH column, 79 students did not fill in the SKTM column, 52 students did not fill in the Income column, and 91 students did not fill in the Dependent column. To resolve the missing values, data rows that had missing values were deleted because the amount of missing data exceeded the acceptable threshold for data imputation, which could cause distortion of the results if forced to be filled in. Figure 3 shows the condition of the data after the missing values were addressed.

```
dataset = dataset.dropna()
missing_values = dataset.isnull().sum()
print("Missing Values:")
print(missing_values)
```

```
Missing Values:
Nama            0
KIP             0
KKS             0
PKH             0
SKTM            0
Penghasilan     0
Tanggungan      0
Status          0
dtype: int64
```

Figure 3. Missing Value Processing Result

Figure 3 is the result after the process of handling missing values by deleting data that has empty values. This data deletion causes the data of students who are eligible to receive PIP to decrease by 129 data to 229 data, and the data of students who are not eligible to receive PIP to decrease by 15 data to 36 data. Figure 4 is the result of visualization of MI Nurul Qolbi student data for the 2022-2023 school year who are eligible and not eligible for PIP funding assistance.

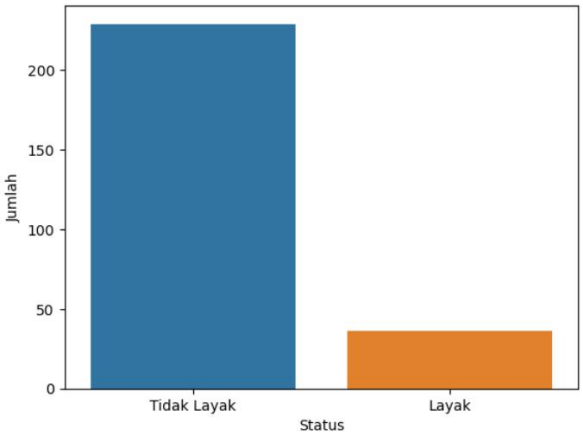

Figure 4. Distribution of Status After Handling Missing Value

Figure 4 shows the data after handling missing values. The remaining data includes 229 PIP-eligible students and 36 PIP-ineligible students. Figure 4 shows the comparison between various feature variables and target variables in the dataset. Figure 5 is a comparison of each feature variable and target variable.

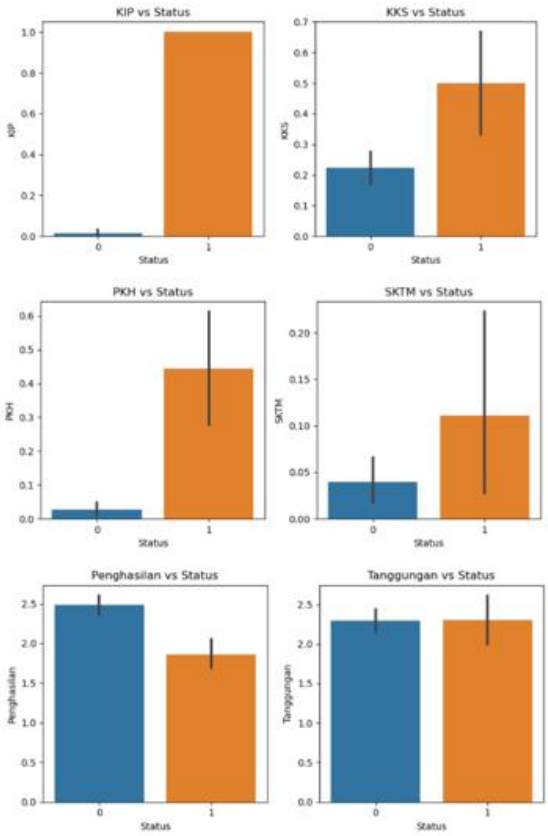

Figure 5. Comparison between feature variables and target variables

In Figure 5, the feature variable that strongly influences the eligibility status of students based on the dataset used is KIP. All students who are eligible to receive PIP assistance funds have KIP, and almost all students who are not eligible to receive PIP assistance funds do not have KIP. Other than KIP, the variables that are quite influential are KKS, PKH, and SKTM. Figure 5 shows that on average, students who are eligible to receive PIP mostly have these three variables. Meanwhile, feature variables such as Income and Dependents are not significantly influential, because the average number of both columns is almost equal.

After processing the data, the next step is to do the labeling. The K-Nearest Neighbor method is carried out using data with numeric types to perform distance calculations (Anwar Pauji et al., 2022), therefore an encoding process is carried out to change the data type according to the needs of the prediction model. The results of changing the data type or encoding can be seen in Figure 5.

	Nama	KIP	KKS	PKH	SKTM	Penghasilan	Tanggungan	Status
0	Alea Kinanti Putri	0	0	0	0	4	1	0
1	Muhammad Sahal	0	0	0	1	3	1	0
2	Muhammad Jibril Adriansyah	0	0	0	0	2	1	0
3	Fara Nisa Putri Anzani	0	0	0	0	2	3	0
4	Ayasha Alzea Humaira	0	0	0	0	4	2	0

Figure 5. Sample Data After Encoding

After changing the data type, the next step is data normalization. Normalization is changing the numerical values of the features from the dataset to a more general value by changing the range of different values, but not changing the meaning of the data. The goal is to make the resulting model more stable (Ibnu Daqiqil, 2021). Normalization is done with the min-max method, which divides the difference between the previous data and the minimum value of the column in the data by the difference between the maximum value and the minimum value of the column (Widaningsih & Yusuf, 2022). The results of data processing are shown in Figure 6.

	KIP	KKS	PKH	SKTM	Penghasilan	Tanggungan	Status
0	0.0	0.0	0.0	0.0	1.000000	0.000000	0
1	0.0	0.0	0.0	1.0	0.666667	0.000000	0
2	0.0	0.0	0.0	0.0	0.333333	0.000000	0
3	0.0	0.0	0.0	0.0	0.333333	0.666667	0
4	0.0	0.0	0.0	0.0	1.000000	0.333333	0

Figure 6. Sample Data After the Data Normalization Process

After the data preprocessing stage, the next step is to perform K-Nearest Neighbors (KNN) modeling using Python. The accuracy results obtained are the average accuracy of each test performed. Furthermore, experiments will be carried out to determine the best k value. Figure 7 is a graph of the accuracy results of each k value with 10-fold cross-validation to determine the optimal k value.

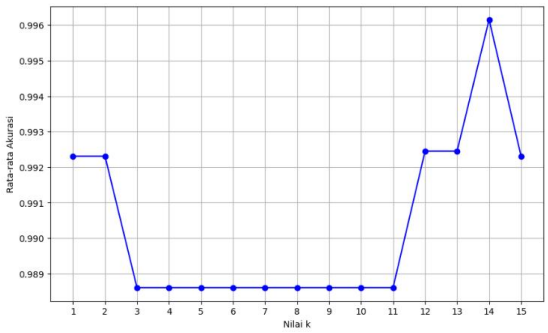

Figure 7. 10-Fold Cross-Validation Accuracy Chart

The k value with the highest accuracy shown in Figure 7 is k=14 with an accuracy of 99.7%. Values of k=1, k=2, k=12, k=13, and k=15 have good accuracy after k=14 at 99.2%. Meanwhile, the k value in the range of 3-11 produces stable accuracy but lower than the other k values which is 98.86%. In this study, the k value used is an odd value to avoid balanced voting results and to ensure classification stability. The accuracy results are shown in Table 2 below.

Table 2. Accuracy Result of The Best K Value

Nilai (k)	CV_AVG_ACC (%)
3	98.86 %
5	98.86 %
7	98.86 %
9	98.86 %
11	98.86 %
13	99.24 %
15	99.23 %

Based on Table 2, the value of k is the number of nearest neighbors to be used and CV_AVG_ACC(%) is the average accuracy percentage using 10-fold cross validation. The values of k=3, k=5, k=7, k=9, and k=11 show a high and stable accuracy rate with an accuracy value of 98.86%. However, when k=13, the accuracy increases to 99.24%, and k=15 produces almost the same accuracy of 99.23%. Although the difference between the accuracy for k=13 and k=15 is very small, k=13 is chosen as the optimal value in this study because it is an odd k value with better accuracy results. In Table IV, we will experiment with k = 13 to see the accuracy obtained by the split ratio. Table 3 is the average obtained by the split ratio.

Table 3. Average Accuracy of Split Ratio

Ratio	Data Used		Accuracy(%)
	Testing	Training	
10:90	27	238	100%
20:80	53	212	98.11%
30:70	80	185	98.75%
40:60	106	159	99.05%
Average Accuracy			**98.98%**

Table 3 shows the results of the accuracy calculation using the split ratio. With a 10:90 ratio between testing and training, 100% accuracy is obtained. The 20:80 ratio resulted in 98.11% accuracy, the 30:70 ratio resulted in 98.75% accuracy, and the 40:60 ratio resulted in 99.05% accuracy. The average accuracy using the split ratio is 98.98%. In Table 4, an experiment will be conducted using the value of k = 13 to see the level of accuracy obtained from each iteration using 10-fold cross validation.

Table 4, Accuracy of Each Iteration of K-Fold Cross Validation

K-Fold	Accuracy
1	100 %
2	100 %
3	100 %
4	100 %
5	96.30 %
6	100 %
7	96.15 %
8	100 %
9	100 %
10	100 %
AVG	**99.24 %**

Table 4 shows the accuracy of each iteration performed with 10-fold cross validation. The k-fold column is the iteration performed and the accuracy is the result of 10-fold cross-validation accuracy. Each iteration produces stable accuracy and the final accuracy result is obtained by calculating the average of the 10 iterations performed which is 98.86%. The results of 10-fold cross-validation show that the KNN model provides consistent performance on segmented training data. However, 100% accuracy on some iterations may signal potential overfitting, where the model fits the training data perfectly but may not perform the same on previously unseen data. After performing data processing and modeling, the next step is to create a website using the Streamlit framework with the Python programming language.

Figure 8. System Interface

Figure 8 is a display of the system that has been created. The data needed is student data to be predicted, users will enter data by selecting one of several options provided by the system, then the system will predict new data using the K-Nearest Neighbor model that is available. Figure 5.1 is the implementation of the output on the system that has been made.

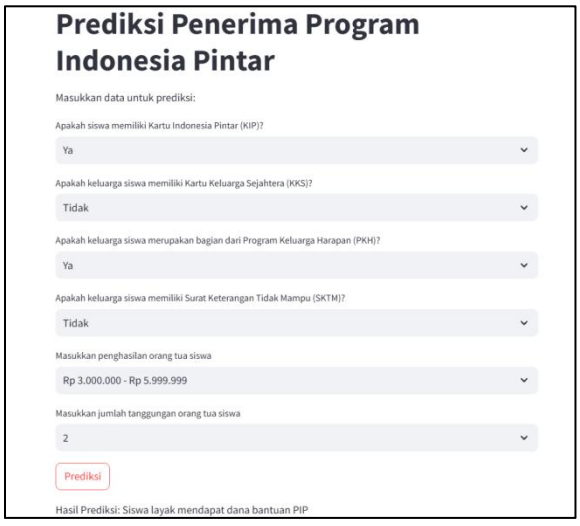

Figure 9. Implementasi Input

In Figure 9, it contains information that students have an Indonesia Smart Card (KIP), do not have a Prosperous Family Card (KKS), are part of the Family Hope Program (PKH), do not have a Certificate of Disability (SKTM), parents' income is in the range of Rp 3,000,000 - Rp 5,999,999, and parents have 2 dependents. After all the data is filled in, click the prediction button on the bottom left to display the results of the student eligibility prediction. The results will be displayed at the very bottom of the dashboard page, namely students eligible for PIP assistance funds.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the results of research that has been conducted in predicting the eligibility of students receiving the Smart Indonesia Program (PIP). This research successfully developed a prediction model for student eligibility to receive the Smart Indonesia Program (PIP) using the K-Nearest Neighbor (KNN) algorithm. Variables such as Indonesia Smart Card (KIP), Prosperous Family Card (KKS), Family Hope Program (PKH), and Certificate of Disability (SKTM) have a significant influence on the eligibility of PIP recipients. The KIP variable shows a strong correlation with beneficiary eligibility, while variables such as income and dependents do not contribute significantly. After various stages of data processing, including handling missing values, encoding, and normalization, the model achieved a very high level of accuracy. The optimal k value chosen was k=13, which resulted in an accuracy of 99.24%. The split ratio process also shows that the model is able to maintain a high level of accuracy, with an average accuracy of 98.98% over a wide range of training and testing data ratios.

However, it is important to note that the high accuracy obtained may indicate potential overfitting, especially since a large portion of the data was used for training. Although the results of this study show that the KNN model is highly effective in predicting students' eligibility for PIP on this dataset, the validity of the model for other datasets cannot be confirmed.

Suggestion

Based on the conclusions obtained from this research, to improve the validity and generalization of the KNN model developed, it is recommended to conduct further testing using a larger and more diverse dataset. In addition, other methods such as Random Forest or Support Vector Machine can be tried to see if they can produce higher accuracy or overcome potential overfitting that may occur in KNN. The selection of different methods could also be more appropriate in conditions where the data is unbalanced or more complex variables need to be considered. These additional evaluations are important to ensure that the resulting predictive models can be effectively applied across different schools and situations.

REFERENCES

Anwar Pauji, Aisyah, S., Surip, A., Saputra, R., & Ali, I. (2022). Implementasi Algoritma K-Nearest Neighbor Dalam Menentukan Penerima Bantuan Langsung Tunai. *KOPERTIP : Jurnal Ilmiah Manajemen Informatika Dan Komputer*, *4*(1), 21–27. https://doi.org/10.32485/kopertip.v4i1.114

Baharuddin, M. M., Azis, H., & Hasanuddin, T. (2019). Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca. *ILKOM Jurnal Ilmiah*, *11*(3), 269–274. https://doi.org/10.33096/ilkom.v11i3.489.269-274

Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *6*(2), 118–127.

Penulis. (2023). PIP, Peningkatan Mutu Pembelajaran dan Pengentasan Kemiskinan Siswa Madrasah. Retrieved April 12, 2024, from Kementerian Agama Republik Indonesia website: https://kemenag.go.id/kolom/pip-peningkatan-mutu-pembelajaran-dan-pengentasan-kemiskinan-siswa-madrasah-X25JQ

Ibnu Daqiqil. (2021). *Machine Learning: Teori, Studi Kasus, dan Implementasi Menggunakan Python*. UR PRESS. https://doi.org/10.5281/zenodo.5113507

Khoirunnisa, Susanti, L., Rokhmah, I. T., & Stianingsih, L. (2021). Prediksi Siswa SMK Al-Hidayah yang Masuk Perguruan Tinggi dengan Metode Klasifikasi. *Jurnal Informatika*, *8*, 26–33.

Kushartanto, A. I., & Aldisa, R. T. (2023). Data Mining Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes dalam Prediksi Penerimaan Beasiswa. *Journal of Computer System and Informatics (JoSYC)*, *5*(1), 196–207. https://doi.org/10.47065/josyc.v5i1.4566

Nata, A., & Royal, S. (2022). Analisis Sistem Pendukung Keputusan Dengan Model Klasifikasi Berbasis Machine Learning Dalam Penentuan Penerima Program Indonesia Pintar. *Journal of Science and Social Research*, (3), 697–702. Retrieved from http://jurnal.goretanpena.com/index.php/JSSR

Noviana, D., Susanti, Y., & Susanto, I. (2019). Analisis Rekomendasi Penerima Beasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN) Dan Algoritma C4.5. *Seminar Nasional Penelitian Pendidikan Matematika Universitas Muhammadiyah Tangerang*, 79–87. https://doi.org/http://dx.doi.org/10.31000/cpu.v0i0.1685

Pebdika, A., Herdiana, R., & Solihudin, D. (2023). Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima PIP. *Jurnal Mahasiswa Teknik Informatika*, *7*(1), 452–458. https://doi.org/10.36040/jati.v7i1.6303

yambodo, D., Nugroho, A., & Zaman, B. (n.d.). Prediksi Ketepatan Waktu Studi Mahasiswa Bidik Misi Menggunakan K-Nearest Neighbour. *Jurnal Pendidikan Teknologi Informasi (JUKANTI)*, (5).

lapdik. (2022). Pusat Layanan Pembiayaan Pendidikan kementerian Pendidikan, Kebudayaan, Riset Dan Teknologi. Retrieved April 12, 2024, from Program Indonesia Pintar website: https://pip.kemdikbud.go.id/

lapdik. (2023a). SIPINTAR Enterprise, Sistem Informasi Indonesia Pintar, Data Penyaluran Nasional. Retrieved April 12, 2024, from Pusat Layanan Pembiayaan Pendidikankementerian Pendidikan, Kebudayaan, Riset Dan Teknologi website: https://pip.kemdikbud.go.id/penyaluran/nasional

lapdik. (2023b). SIPINTAR Enterprise, Sistem Informasi Indonesia Pintar, Data Penyaluran Provinsi. Retrieved April 12, 2024, from Pusat Layanan Pembiayaan Pendidikankementerian Pendidikan, Kebudayaan, Riset Dan Teknologi website: https://pip.kemdikbud.go.id/penyaluran

ımadini, LorencisLubis, E. E., Priansyah, A., N, Y. R. W., & Meutia, T. (2023). PENERAPAN Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor. *JURNAL MAHASISWA AKUNTANSI SAMUDRA (JMAS)2023*, *4*(4), 223–235. https://doi.org/10.33059/jmas.v4i4.7074

ndhani, A. M. (2023). *Petunjuk Teknis Pelaksanaan Program Indonesia Pintar Untuk Siswa Madrasah Tahun Anggaran 2023*. Retrieved from www.pendis.kemenag.go.id

niah, A., & Mirantika, N. (2020). Perbandingan Metode K-Nearest Neighbor dan Naive Bayes untuk Rekomendasi Penentuan Mahasiswa Penerima Beasiswa pada Universitas Kuningan. *Buffer Informatika*, *6 (1)*, 1–10. Retrieved from https://journal.uniku.ac.id/index.php/buffer

laningsih, S., & Yusuf, S. (2022). Penerapan Data Mining Untuk Memprediksi Siswa Berprestasi Dengan Menggunakan Algoritma K Nearest Neighbor. *Jurnal Teknik Informatika Dan Sistem Informasi*, *9*(3), 2598–2611. Retrieved from http://jurnal.mdp.ac.id

ıarno, J. (2023). Penerapan Algoritma K-Nearest Neighbour Untuk Prediksi Penerima Beasiswa. *Teknologiterkini.Org*, *3*(3), 1–16.

ıdi Saputra, A., & Primadasa, Y. (2018). Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour Implementation of Classification Method to Predict Student Graduation Using K-

Nearest Neighbor Algorithm. *Techno Com*, *17*(4), 395–403. https://doi.org/10.33633/tc.v17i4.1864