Hyperparameter Tuning Feature Selection with Genetic Algorithm and Gaussian Naïve Bayes for Diabetes Disease Prediction

Ilham Firman Ashari[*1], Meida Cahyo Untoro[*2], Eliza Maharani Sutowo[*3], Della Salsabila[*4], Dhifaf Athiyah Zhabiyan[*5]

[*]*Informatics Engineering, Institut Teknologi Sumatera*
Jl. Terusan Ryacudu, Indonesia
[1]firman.ashari@if.itera.ac.id
[2]cahyo.untoro@if.itera.ac.id

Abstract— **Diabetes Mellitus is a disease that occurs due to disorders of carbohydrate, fat and protein metabolism associated with a lack of performance of insulin secretion. Diabetes is a degenerative disease that requires appropriate and serious treatment efforts. The effects lead to various complications of other serious diseases such as heart disease and stroke. Erectile dysfunction, kidney failure, nervous system damage, etc. Because there are so many impacts caused by diabetes, it is important to study this disease. The benefit of this study is to prevent the occurrence of severe complications and can help medical personnel in predicting this disease early and reduce the cost burden that arises due to this problem. The purpose of this study is to determine the level of accuracy resulting from the use of feature selection with genetic algorithms and nave Bayes. In this study, predictions will be made using hyperparameter tuning with genetic algorithms and Naive Bayes optimization by performing feature selection. After conducting related research, it was found that the accuracy of 17 features using a genetic algorithm was better than modeling with 10 features. By using 17 features and hyperparameter tuning with genetic algorithm and naive Bayes modeling, the accuracy is 93.2%. By using 17 features without feature selection, the accuracy is 91.2%, there is an increase in accuracy of 1.5%.**

Keywords— **Diabetes Melitus, hyperparameter, feature selection, Genetic algorithm, Naïve Bayes.**

Abstrak— Diabetes Mellitus merupakan penyakit yang terjadi akibat gangguan metabolisme karbohidrat, lemak, dan protein yang berhubungan dengan kurangnya kinerja sekresi insulin. Diabetes merupakan penyakit degeneratif yang memerlukan upaya penanganan yang tepat dan serius. Efeknya mengakibatkan berbagai komplikasi penyakit serius lainnya seperti penyakit jantung dan stroke. Disfungsi ereksi, gagal ginjal, kerusakan sistem saraf, dll. Karena begitu banyak dampak yang ditimbulkan oleh penyakit diabetes, maka penting untuk mempelajari penyakit ini. Manfaat dari penelitian ini adalah untuk mencegah terjadinya komplikasi yang parah dan dapat membantu tenaga medis dalam memprediksi penyakit ini secara dini serta mengurangi beban biaya yang timbul akibat masalah ini. Tujuan dari penelitian ini adalah mengetahui tingkat akurasi yang dihasilkan dari penggunakan seleksi fitur dengan algoritma genetika dan Naïve Bayes. Pada penelitian ini dilakukan prediksi menggunakan hyperparameter tuning dengan algoritma genetika dan optimasi Naive Bayes dengan melakukan seleksi fitur. Setelah dilakukan penelitian terkait, ditemukan bahwa akurasi 17 fitur menggunakan algoritma genetika lebih baik daripada pemodelan dengan 10 fitur. Dengan menggunakan 17 fitur dan hyperparameter tuning dengan algoritma genetika dan pemodelan Naive Bayes menghasilkan akurasi sebesar 93,2%. Dengan menggunakan 17 fitur tanpa seleksi fitur didapatkan akurasi sebesar 91,2%, terjadi peningkatan akurasi sebesar 1,5%.

Kata Kunci— **Diabetes Melitus, hyperparameter, *seleksi fitur*, *algoritme Genetika*, Naïve Bayes.**

I. INTRODUCTION

Diabetes is a serious chronic disease that occurs when the pancreas does not produce enough insulin, or the body cannot effectively use the insulin it produces. The disease is characterized by high blood sugar levels that exceed normal limits [1][2]. There are three types of diabetes: insulin-dependent, non-insulin-dependent diabetes (NIDDM), and gestational diabetes (GDM) [2]. Diabetes is a major cause of heart disease and one of the tragedies that kills mothers during childbirth. In addition, diabetes is also a risk factor for disease transmission in newborns [3]. Type diabetes is a degenerative disease that requires appropriate and serious treatment. Its effects result in a variety of complications of other serious illnesses, such as heart disease and stroke. Erectile dysfunction, renal failure, nervous system damage, etc. [4].

In developing countries, the increase in Diabetes Mellitus sufferers can be caused by genetic factors, age, demographics of a country, and lifestyle changes, such as overeating and lack of exercise activity. In addition, there are several risk factors for the possibility of Diabetes Mellitus, namely gender, high levels of physical activity, body mass index (BMI), hypertension, alcohol and cigarette consumption habits, and low knowledge about early detection of Diabetes Mellitus [5]. According to the International Diabetes Federation (IDF), up to 382 million people have diabetes. This number is expected to increase to 592 million by 2035. About 175 million of them remain undiagnosed and are estimated to be at risk of progressive onset and complications without prior prevention efforts [6].

Because there are so many impacts caused by diabetes, it is important to study this disease. Technology needs to be used

to make it easier to solve problems [7]. Similar research has been done before. Several previous studies only focus on classifying and predicting diabetes mellitus using Naïve Bayes, KNN, Decision Tree algorithm, C4.5, Naive Bayes, SVM, Logistic Regression [8][9][10]. Although it is the most widely used method, Naïve Bayes still has a weakness in which the probability results are not running optimally and are often wrong on attributes. To overcome the weakness of Naive Bayes, a data weighting method is applied to increase the accuracy of Naive Bayes [11].

Another research that uses machine learning methods that have low recall rate disorders is overcome by using the feature selection method combined with the Naive Bayes method so that the AUC value increases from 0.71 to 0.98 [12]. Feature selection is a way to make classification more efficient and effective. This method is better because by reducing the amount of data analyzed or by identifying appropriate features as material for the learning process [13]. There are two types of feature selection process in machine learning. They are the wrapper method which uses several algorithms as a measurement of classification accuracy and the filter method. [14]. Feature selection is a method used to optimize classifier performance. The way this feature works is by reducing the large feature space and by eliminating the less significant attributes [15].

Genetic algorithm is a heuristic algorithm based on the mechanism of natural genetics and natural selection [16][17]. The Genetic algorithm starts from a group of individuals called a population [6]. Genetic algorithms are commonly used to solve problems related to optimization [18]. In this study, predictions about diabetes will be made using Naive Bayes modelling. However, optimization of the Naive Bayes model will be carried out using a Genetic algorithm to perform feature selection. The feature selection is carried out to improve the accuracy of the modelling that is carried out so that it is better and more efficient.

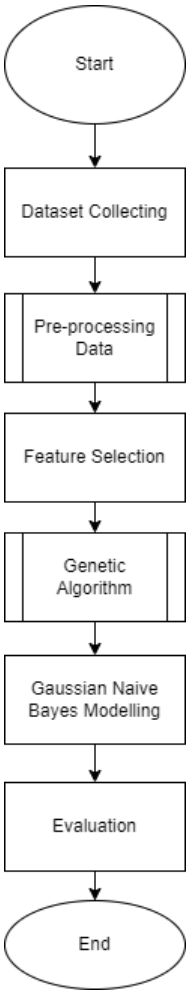


Fig 1. Overview of research flow

II. METHODOLOGY

The method used to predict diabetes and model the data in this study is to implement the Genetic Algorithm and Naive Bayes optimization to perform feature selection. The focus of this research is on feature selection to build an efficient model. The overall research flow can be seen in Figure 1.

A. Dataset

In this research, the dataset used is a dataset with public access from the Kaggle platform entitled Early-Stage Diabetes Risk Prediction Dataset. The dataset needs to be filtered so that the processing results run optimally [19]. In the dataset there are 520 data from subjects who have ages in the range of 16 years to 90 years. This dataset has 16 attributes, namely Age, Gender, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital Thrush, Visual Blurring, and Itching. The Table I is a description of each attribute possessed by the dataset.

From the existing attribute data, it can be seen the correlation between each attribute. This can be seen from the correlation matrix shown in Figure 2.

B. Data Pre-processing

In the data pre-processing stage, data preparation and processing are carried out before the dataset is modelled. Figure 3 is the stages contained in the pre-processing stage of the data.

1) Data Transformation

In the process of data transformation, the value of the attribute data will be changed to a new scale. This process will be applied to attributes that have string data values, such as 'Male', 'Female', 'Yes', 'No', 'Positive', and 'Negative'. This process will change the string attribute data into numeric form 1 or 0. Data in numeric or binary form makes it easier for computers to perform data processing. In this study, the values of 'Male', 'Yes', and 'Positive' will be converted into numeric

TABLE I
ATTRIBUTE DESCRIPTION (16 FEATURES)

Attribute	Attribute Description	Value
Age	Subject's age in years	16 - 90 years
Gender	Subject gender	female: 192 male: 327
Polyuria	Subjects urinate frequently	yes: 262 not: 258
Polydipsia	The subject often or drinks a lot	yes: 287 not: 233
Sudden Weight Loss	The subject suddenly loses weight	yes: 303 not: 217
Weakness	Subject feels weak or tired	yes: 215 not: 305
Polyphagia	The subject often or eats a lot	yes: 283 not: 237
Genital Thrush	Subjects experiencing genital thrush	yes: 404 not: 116
Visual blurring	Subject is blurred	yes: 287 not: 233
Itching	The subject has itching	yes: 267 not: 253
Irritability	Subject is irritated	yes: 394 not: 126
Delayed healing	The subject has a long healing time	yes: 281 not: 239
Partial paresis	The subject is partially paralysed	yes: 296 not: 224
Muscle stiffness	Subjects experience muscle stiffness	yes: 325 not: 195
Alopecia	Subject experiencing baldness or hair loss	yes: 341 not: 179
Obesity	Subject is obese	yes: 432 not: 88

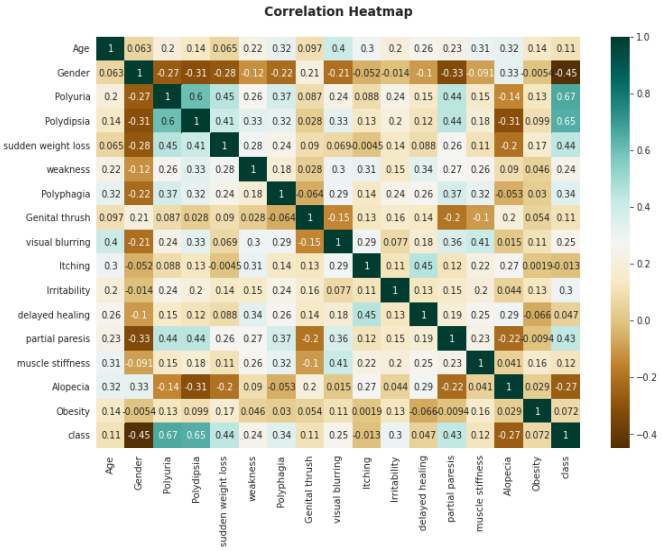


Fig 2. Correlation matrix

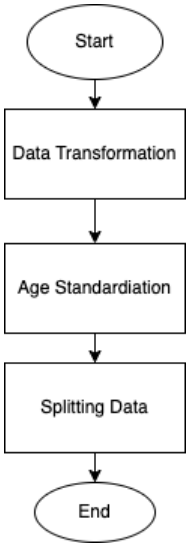


Fig 3. Pre-processing

data 1. Attribute data that has values of 'Female', 'No', and 'Negative' will be converted into numeric data 0.

2) Age Attribute Standardization

The standardization process is carried out on the 'Age' attribute because the value of this attribute has a range and is very diverse. Therefore, it is necessary to standardize the 'Age' attribute data which can speed up data training and improve modeling accuracy. Standardization is carried out using the Z-Score formula as follows.

$$Z = (Yi - Y)/SD \quad (1)$$

Z is Z-score, Yi is raw score, Y is initial mean, and SD is standard deviation.

3) Split Data into Train and Test Data

Training and test data is needed to perform the feature selection and modeling process using the Naive Bayes algorithm. Therefore, the modified and standardized data is divided into training data and test data. Training data is data used to conduct model training, while test data is not used as a model, but data that is tested by the model. The results of testing the test data on the training data model will be evaluated later to determine whether the model is overfitting. In this study, the training data ratio of 7:3 was used for the test data.

C. Feature Data Selection

Feature selection in this study aims to improve the accuracy of the model built with the Naive Bayes algorithm and make

the model more efficient. At this feature selection stage, feature selection is carried out on attributes. This is done by using hyperparameter to perform the Genetic algorithm process, to produce optimal functionality for modeling with Naive Bayes. Figure 4 shows the process flow of the Genetic algorithm [18].

D. Naive Bayes Modelling

At this stage of Naive Bayes modeling, the help of the GaussianNB() function from the genetic_selection module will be used. Modeling will be carried out using certain attributes generated from the feature selection stage on attributes using the Genetic algorithm. Furthermore, the modeling results will be tested with test data so as to produce an accuracy value from the predictions made by the model.

III. Results and Discussion

In this study, the assistance of Google Collab was used, without activating the GPU feature, and the programming language used was Python based on Jupyter Notebook. The following are the results of implementation and discussion.

A. Data Pre-Processing

The first step taken after the desired or required dataset in this research is data pre-processing. In this process, the data used must be explored, what attributes are there and what are their uses or effects on the labeling of each data. The dataset used in this study was obtained from the Kaggle website and accessed using the Pandas library. Files are stored in existing Google Drive. The code to read csv dataset can be seen in Figure 5.

Figure 6 shows an example of the dataset used. There are 16 features that are owned and related to the diagnosis of diabetes. Of the 16 features, all of them use Boolean values for 14 features. For age, an integer value is used which indicates the age of the patient. For gender or gender use the string value male or female. The last is the class or label of each data. The dataset parameters used in the study can be seen in Figure 6.

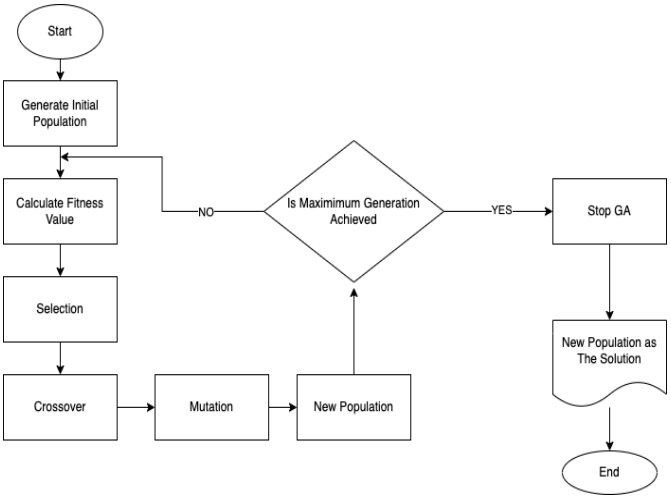


Fig 4. Genetic algorithm process

```
drive.mount('/content/gdrive')
```

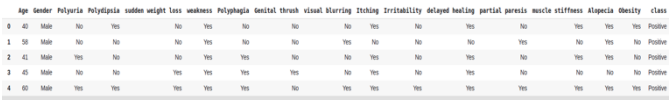
```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content
```

```
df = pd.read_csv('/content/gdrive/MyDrive/jurnal-ksi/eksplorasi/dataset/diabetes_data_upload.csv')
df
```

Fig 5. Reading datasets

In the data pre-processing stage, the first step is to change the Male value to 1 and the Female value to 0. Followed by changing the Yes value to 1 and the No value to 0 and the Positive value to 1 and the Negative value to 0. Finally, changing the age value in Age column with standard z-score values. The code to represent class parameters and values consisting of 16 features and 1 class can be seen in Figure 7.

Figure 8 shows the data that has been converted and equalized so that it will be easier to process by Genetic algorithms and Naive Bayes models. From the code in Figure 7, the resulting data is shown in Figure 8.



	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

Fig 6. Dataset

```
df['Gender'] =df['Gender'].map({'Male' : 1 , 'Female':0})
df['Polyuria'] =df['Polyuria'].map({'Yes' : 1 , 'No':0})
df['Polydipsia'] =df['Polydipsia'].map({'Yes' : 1 , 'No':0})
df['sudden weight loss'] =df['sudden weight loss'].map({'Yes' : 1 , 'No':0})
df['weakness'] =df['weakness'].map({'Yes' : 1 , 'No':0})
df['Polyphagia'] =df['Polyphagia'].map({'Yes' : 1 , 'No':0})
df['Genital thrush'] =df['Genital thrush'].map({'Yes' : 1 , 'No':0})
df['visual blurring'] =df['visual blurring'].map({'Yes' : 1 , 'No':0})
df['Itching'] =df['Itching'].map({'Yes' : 1 , 'No':0})
df['Irritability'] =df['Irritability'].map({'Yes' : 1 , 'No':0})
df['delayed healing'] =df['delayed healing'].map({'Yes' : 1 , 'No':0})
df['partial paresis'] =df['partial paresis'].map({'Yes' : 1 , 'No':0})
df['muscle stiffness'] =df['muscle stiffness'].map({'Yes' : 1 , 'No':0})
df['Alopecia'] =df['Alopecia'].map({'Yes' : 1 , 'No':0})
df['Obesity'] =df['Obesity'].map({'Yes' : 1 , 'No':0})
df['class'] =df['class'].map({'Positive' : 1 , 'Negative':0})

df['Age'] = (df['Age']-df['Age'].mean())/df['Age'].std()

df = df.astype(float)
scaled_df=df
df.head(100)
```

Fig 7. Data Transformation

	Age	Gender	Polyuria	Polydipsia	...	muscle stiffness	Alopecia	Obesity	class
0	-0.660731	1.0	0.0	1.0	...	1.0	1.0	1.0	1.0
1	0.820572	1.0	0.0	0.0	...	0.0	1.0	0.0	1.0
2	-0.578436	1.0	1.0	0.0	...	1.0	1.0	0.0	1.0
3	-0.249258	1.0	0.0	0.0	...	0.0	0.0	0.0	1.0
4	0.985161	1.0	1.0	1.0	...	1.0	1.0	1.0	1.0
...	...	...	...	...	...	...	...	...	...
95	0.655983	0.0	1.0	1.0	...	0.0	0.0	0.0	1.0
96	-1.483677	0.0	1.0	1.0	...	1.0	0.0	0.0	1.0
97	-1.401382	0.0	1.0	1.0	...	1.0	0.0	0.0	1.0
98	-1.072204	0.0	1.0	1.0	...	0.0	1.0	0.0	1.0
99	-0.743025	0.0	1.0	1.0	...	0.0	0.0	0.0	1.0

Fig 8. Data Transformation Result

After pre-processing the data, the next step is to divide the data into two parts (7:3): training data and test data. The distribution of the test and training data codes can be seen in Figure 9.

B. Feature Selection

Next is to select the feature or data column that will be used. This aims to improve the accuracy of the Naive Bayes model that will later be formed. To select this feature, use the hyperparameter setting. In this genetic algorithm, an estimator in the form of a decision tree is used to select the most accurate results from each process in this genetic algorithm. The function for feature selection can be seen in Figure 10.

The estimation algorithm used is a decision tree classifier to estimate population values using sample values. The population that has degenerated is 100 populations with a probability of crossing over 0.5 and a mutation probability of 0.2 and with a number of generations of 1000. The selection used is the selection of tournament rankings with a ranking range limit of 3.

C. Naive Bayes Modelling

To perform this Naive Bayes modelling, the help of the sklearn.naive_bayes module is used with the GaussianNB function. The input data used is training data whose features match those generated by feature selection using Genetic algorithms. The code used for modelling and calculating accuracy can be seen in Figure 11.

D. Experiment

In this study, experiments were carried out starting from the feature selection process to the model training process using Naive Bayes. This is done repeatedly to get the most optimal features and is able to produce models with high accuracy, es pecially when tested with data testing. Table III is a breakdown of each experiment carried out.

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.33, random_state=42)
```

Fig 9. Sharing test and train data

```
# dataframe = df.copy()
# x = dataframe.drop(['class'], axis=1)
# y = dataframe['class'].astype(float)
x = X_train
y = y_train
estimators = DecisionTreeClassifier()
# estimators = linear_model.LogisticRegression(solver="liblinear", multi_class="ovr")
models = GeneticSelectionCV(
    estimators, cv=5, verbose=0,
    scoring="accuracy", max_features=10,
    n_population=100, crossover_proba=0.5,
    mutation_proba=0.2, n_generations=1000,
    crossover_independent_proba=0.5,
    mutation_independent_proba=0.04,
    tournament_size=3, n_gen_no_change=50,
    caching=True, n_jobs=-1)
models = models.fit(x, y)
print('Feature Selection:', x.columns[models.support_])
```

Fig 10. Feature selection function

TABLE III
EXPERIMENT RESULT

No.	Hyperparameter With Genetic Algorithm	Number and Features Generated	Naive Bayes Accuracy
1	There is no genetic algorithm for feature selection	The number of Feature: 16 Feature: 'Age', 'Gender' , 'Polyuria', 'Polydipsia' , 'sudden weight loss' ,'weakness', 'Polyphagia' ,'Genital thrush', 'visual blurring', 'Itching', 'Irritability', 'delayed healing', 'partial paresis', 'muscle stiffness' , 'Alopecia', 'Obesity'	Accuracy: 91.7 %
2	estimators, cv=5, verbose=0, scoring="accuracy", max_features=16, n_population=100, crossover_proba=0.5, mutation_proba=0.2, n_generations=50, crossover_independent_proba=0.5, mutation_independent_proba=0.04, tournament_size=3, n_gen_no_change=10, caching=True, n_jobs=-1	The number of Feature: 16 Feature: 'Age', 'Gender' , 'Polyuria', 'Polydipsia' , 'sudden weight loss' ,'weakness', 'Polyphagia' ,'Genital thrush', 'visual blurring', 'Itching', 'Irritability', 'delayed healing', 'partial paresis', 'muscle stiffness' , 'Alopecia', 'Obesity'	Accuracy: 93.2 %
3	estimators, cv=5, verbose=0, scoring="accuracy", max_features=16, n_population=100, crossover_proba=0.5, mutation_proba=0.2, n_generations=50, crossover_independent_proba=0.5, mutation_independent_proba=0.04, tournament_size=3, n_gen_no_change=10, caching=True, n_jobs=-1	The number of Feature: 13 Feature: 'Age', 'Gender', 'Polydipsia', 'sudden weight loss', 'weakness', 'Genital thrush', 'visual blurring', 'Itching','Irritability','delayed healing', 'partial paresis','muscle stiffness', 'Alopecia'	Accuracy: 88.2 %

No.	Hyperparameter With Genetic Algorithm	Number and Features Generated	Naive Bayes Accuracy
4	estimators, cv=10, verbose=0, scoring="accuracy", max_features=16, n_population=100, crossover_proba=0.5, mutation_proba=0.2, n_generations=1000, crossover_independent_proba=0.5, mutation_independent_proba=0.05, tournament_size=3, n_gen_no_change=50, caching=True, n_jobs=-1	The number of Feature: 11 Feature: 'Age', 'Gender', 'Polyuria', 'Polydipsia', 'sudden weight loss', 'Genital thrush','Irritability', 'delayed healing', 'muscle stiffness','Alopecia', 'Obesity'	Accuracy: 87.91 %
5	estimators, cv=5, verbose=0, scoring="accuracy", max_features=8, n_population=100, crossover_proba=0.5, mutation_proba=0.2, n_generations=100, crossover_independent_proba=0.5, mutation_independent_proba=0.04, tournament_size=3, n_gen_no_change=10, caching=True, n_jobs=-1	The number of Feature: 9 Feature: 'Age', 'Gender', 'Polyuria', 'sudden weight loss', 'Genital thrush','visual blurring', 'Itching','Irritability', 'Alopecia'	Accuracy: 87.91 %
6	estimators, cv=5, verbose=0, scoring="accuracy", max_features=8, n_population=100, crossover_proba=0.5, mutation_proba=0.2, n_generations=100, crossover_independent_proba=0.5, mutation_independent_proba=0.04, tournament_size=3, n_gen_no_change=10, caching=True, n_jobs=-1	The number of Feature: 8 Feature: 'Age', 'Gender', 'Polyuria', 'sudden weight loss', 'visual blurring', 'Irritability', 'delayed healing', 'Alopecia'	Accuracy: 87.08 %
7	estimators, cv=5, verbose=0, scoring="accuracy", max_features=10, n_population=100, crossover_proba=0.5, mutation_proba=0.2, n_generations=1000, crossover_independent_proba=0.5, mutation_independent_proba=0.04, tournament_size=3, n_gen_no_change=50, caching=True, n_jobs=-1	The number of Feature: 10 Feature: 'Age', 'Gender', 'Polyuria', 'sudden weight loss', 'Polyphagia', 'Genital thrush', 'visual blurring', 'Irritability', 'delayed healing', 'Alopecia'	Accuracy: 87.06 % STD: 4.69 %

```
nb=GaussianNB()
nb.fit(X_train_new,y_train)
accuracies = cross_val_score(estimator=nb, X=X_train_new ,y=y_train,cv=10)
print("accuracy is {:.2f} %".format(accuracies.mean()*100))
print("std is {:.2f} %".format(accuracies.std()*100))
```

Fig 11. Naïve Bayes modelling

From the six experiments, each process was carried out iteratively from the Genetic algorithm to the model evaluation stage by testing the data. The dataset generated in the sixth trial was selected using 10 features: Age, Gender, Polyuria, Sudden Weight Loss, Polydipsia, Genital Candidiasis, Visual Blurring, Hypersensitivity, Delayed Healing, and Polyphagia.

E. Model Evaluation

The last step is to re-confirm the resulting model by testing it using testing data. For modeling, a data set with 10 selected features was used as generated by the 6th experimental stage. The initial comparison uses test data if the data used is initial data without feature selection. The resulting accuracy is 92%. Table IV is the confusion matrix of the model using normal data.

When compared with a model that uses 10 selected features, a model with an accuracy increase of 1% is produced. The confusion matrix table can be seen in Table V.

The precision, recall, and f-score values of the Naive Bayes model that were trained using 10 selected features from the initial dataset are shown in Table VI. Table VI shows the accuracy results.

After applying the Genetic algorithm for feature selection, it generates the 10 best features or attributes to use in modeling with Naive Bayes. The ten characteristics were Age, Gender, Polyuria, Sudden Weight Loss, Polyphagia, Thrush, Blurred Vision, Irritability, Delayed Healing, and Alopecia. These ten features are then used to optimize the model using Naive Bayes. Table VII are the results of modelling using these 10 features.

IV. CONCLUSIONS

From this study, it can be concluded that modeling using 10 features selected using the Genetic algorithm has a higher accuracy value than modeling using 17 features. The results of modeling using 17 features with hyperparameters and feature selection using the Genetic algorithm, modeling with Naïve Bayes produces an accuracy of 93.2%. Using 17 features without feature selection resulted in an accuracy of 91.7%. There is an increase in accuracy of 1%. This is because the existing parameters have been tuned with a genetic algorithm.

TABLE IV
CONFUSION MATRIX WITH NORMAL MATRIX

	0	1
0	54	7
1	7	104

TABLE V
CONFUSION MATRIX WITH 10 FEATURES

	0	1
0	55	6
1	6	105

TABLE VI
ACCURACY RESULTS WITH 10 FEATURES

	Precision	Recall	F1-score	Support
0	0.90	0.90	0.90	61
1	0.95	0.95	0.95	111
Accuracy			0.93	172
Macro avg	0.92	0.92	0.92	172
Weighted avg	0.93	0.93	0.93	172

TABLE VII
MODELLING RESULT WITH 10 FEATURES

Feature	Label	Precision	Recall	F-1 score
Age, Gender, Polyuria, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesity	0.0	0.90	0.90	0.90
	1.0	0.95	0.95	0.95
Accuracy		0.93		

REFERENCES

[1] M. A. Wiratama and W. M. Pradnya, "Optimasi algoritma data mining menggunakan backward elimination untuk klasifikasi penyakit diabetes," *J. Nas. Pendidik. Tek. Inform.*, vol. 11, no. 1, p. 1, 2022. DOI: 10.23887/janapati.v11i1.45282.

[2] F. Handayanna, Rinawati, E. Arisawati, and L. S. Dewi, "Prediksi penyakit diabetes menggunakan Naive Bayes dengan optimasi parameter menggunakan algoritme Genetika," *KNiST (Konferensi Nas. Ilmu Sos. Teknol.*, March 2017, pp. 71–76.

[3] K. Mohammad Burhan Hanif, "Sistem aplikasi prediksi penyakit diabetes menggunakan feature selection korelasi Pearson dan klasifikasi Naive Bayes," *Pengemb. Rekayasa dan Teknol.*, vol. 16, no. 2, pp. 199–205, 2020.

[4] S. Katno and D. Anistyani, "Uji aktivitas hipoglikemik ekstrak etanol daun teh (*Camellia Sinensis L.*) pada tikus putih jantan galur wistar," in *Prosiding Seminar Nasional "Peranan dan Kontribusi Herbal dalam Terapi Penyakit Degeneratif*", Semarang, December 17th, 2011, pp. 108-113.

[5] G. Kusnadi, E. A. Murbawani, and D. Y. Fitranti, "Faktor risiko diabetes melitus tipe 2 pada petani dan buruh," *J. Nutr. Coll.*, vol. 6, no. 2, p. 138, 2017. DOI: 10.14710/jnc.v6i2.16905.

[6] W. D. Septiani and U. Rohwadi, "Optimasi algoritma Genetika pada algoritme C4.5 untuk deteksi dini penyakit diabetes," *J. Akrab Juara*, vol. 6, pp. 221–229, 2021.

[7] M. C. Untoro, M. Praseptiawan, I. F. Ashari, and A. Afriansyah, "Evaluation of Decision Tree, K-NN, Naive Bayes, and SVM with MWMOTE on UCI dataset," *J. Phys. Conf. Ser.*, vol. 1477, no. 3, 2020. DOI: 10.1088/1742-6596/1477/3/032005.

[8] A. Ridwan, "Penerapan algoritme Naïve Bayes untuk klasifikasi penyakit diabetes melitus," *J. Siskom-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020. DOI: 10.47970/siskom-kb.v4i1.169.

[9] N. M. Putry and B. N. Sari, "Komparasi algoritme KNN dan Naïve Bayes untuk klasifikasi diagnosis penyakit diabetes melitus," *Evolusi J. Sains dan Manaj.*, vol. 10, no. 1, pp. 45–57, 2022.

[10] Noviandi, "Implementasi algoritme Decision Tree C4.5 untuk prediksi penyakit diabetes," *J. Inohim*, vol. 6, no. 1, pp. 1–5, 2018. [Online]. Available: http://www.kaggle.com/uciml/pima-indians-diabetes-database.

[11] I. S. Bakti and Ivandari, "Model prediksi penyakit diabetes menggunakan Bayesian classification dan information *gain* untuk seleksi fitur dan adaptive boosting untuk pembobotan data," *J. Inform. Comput. Technol.*, vol. 14, no. 1, pp. 1–13, 2019. DOI: 10.47775/ictech.v14i1.54

[12] T. Zheng, W. Xie, L. Xiu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," *International Journal of Medical Informatics*, vol. 97, p. 120-127, 2017. DOI: 10.1016/j.ijmedinf.2016.09.014.A.

[13] S. A. Putri and D. Larasati, "Penerapan feature selection pada Bayesian network untuk prediksi cacat perangkat lunak," *Pilar Nusa Mandiri*, vol. 13, no. 2, pp. 275–280, 2017.

[14] O. Somantri and D. Apriliani, "Support vector machine berbasis feature selection untuk sentiment analysis kepuasan pelanggan terhadap pelayanan warung dan restoran kuliner kota Tegal," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 537, 2018. DOI: 10.25126/jtiik.201855867.

[15] O. Somantri and M. Khambali, "Feature selection klasifikasi kategori cerita pendek menggunakan Naïve Bayes dan algoritme Genetika," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 6, no. 3, pp. 301–306, 2017. DOI: 10.22146/jnteti.v6i3.332.

[16] R. N. Putri and D. Setiawan, "Prediksi penyakit Systemic Lupus Erythematosus menggunakan algoritme Genetika," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 12, no. 1, pp. 19–31, 2021. DOI: 10.31849/digitalzone.v12i1.5973.

[17] D. Setiawan, R. N. Putri, and R. Suryanita, "Implementasi algoritme Genetika untuk prediksi penyakit autoimun," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 4, no. 1, pp. 8–16, 2019. DOI: 10.36341/rabit.v4i1.595.

[18] I. F. Ashari, A. G. Manalu, and R. Setiawan, "Analysis of security guard scheduling system using Genetic algorithm and tournament selection (case study: Institut Teknologi Sumatera)," vol. 5, no. 2, pp. 202–207, 2021.

[19] I. F. Ashari, R. Banjarnahor, and D. R. Farida, "Application of data mining with the K-Means clustering method and Davies Bouldin Index for grouping IMDB movies," vol. 6, no. 1, pp. 7–15, 2022.

Ilham Firman Ashari, born in Batusangkar, 1993. After studying for a master's degree at Institut Teknologi Bandung in the field of information security, he became a lecturer in Informatics at the Institut Teknologi Sumatra. In addition to being a lecturer, he is also actively carrying out the Tridharma Perguruan Tinggi and serves as chairman of the cybersecurity and pervasive science group. His research focuses are information security, IoT, machine learning, and computing.