Classification Analysis of Cardiovascular Diagnosis with Lifestyle and Demographic Factors: Comparison of Logistic Regression, Lasso Regression, and Ridge Regression

Evangeline Suciadi[1] ,Stevani Alexandra Harmareta[2] ,Albertus Christian[3] ,
Ivan Bagus Purnomo[4]
[1,2,3,4] Information System, Faculty of Engineering & Information
Multimedia Nusantara University, Tangerang Banten 15810, Indonesia.

[1]evangeline.suciadi@student.umn.ac.id,[2]
stevani.alexandra@student.umn.ac.id,[3] albertus.christian@student.umn.ac.id,
[4]ivan.bagus@student.umn.ac.id

https://github.com/albertusc/HeartDiseaseModelTraining

Abstract-Cardiovascular ***disease*** (CVD) is one of the leading causes of death globally, which demands early detection and effective management to reduce its negative impact. This study investigates the use of data mining techniques, specifically Logistic Regression, Lasso Regression, and Ridge Regression, in developing predictive models for CVD risk. The aim of this study is to evaluate the performance of these models in predicting CVD risk considering lifestyle, demographic, and medical variables. The study used a dataset involving these factors to build and test the models, which can then assist healthcare practitioners in identifying high-risk patients and designing targeted interventions.
keywords : cardiovascular disease, regression, detection

Results from the study showed that a combination of the three regression techniques resulted in a robust predictive model, with Logistic Regression often showing the best performance based on accuracy and ease of interpretation. The study also highlighted the importance of including comprehensive lifestyle and demographic variables to improve the accuracy of CVD risk prediction. Furthermore, the application of this model in AI cardiovascular screening shows the potential for the development of more accurate and cost-effective screening, utilizing advances in artificial intelligence technology.

I. INTRODUCTION

1.1 Background

Data mining acts as a powerful tool for healthcare professionals, allowing them to unlock valuable knowledge from massive data sets[1]. Using techniques such as classification, clustering, and association rule learning, researchers can extract patterns and make predictions about patient data[1].

This results in real-world benefits. Data mining helps in identifying optimal treatment plans, predicting disease susceptibility, and optimizing healthcare cost structures. Researchers have successfully applied these techniques to a variety of diseases, including diabetes, asthma, and cardiovascular disease[2]. A variety of data mining tools are available, ranging from simpler models such as Naive Bayes to more complex artificial intelligence techniques such as neural networks. These tools empower researchers to create models capable of analyzing medical data, uncovering relationships, and making predictions[3].

Cardiovascular disease (CVD) is a serious global health problem, causing more than 17 million deaths each year[3]. CVD encompasses a number of disorders of the heart and blood vessels, including coronary heart disease, stroke, and other diseases that have a serious impact on an individual's quality of life and impose a large burden on the healthcare system.

Heart disease is a health problem that continues to increase in prevalence worldwide and contributes significantly to the burden of morbidity and mortality. According to the World Health Organization (WHO), heart disease is the leading cause of death in the world. Therefore, early detection and effective management of heart disease is essential to reduce its negative impact[4].

Although much research has been done in the development of predictive models of cardiovascular risk, there is still a need to continuously improve the accuracy of these models. One aspect that needs attention is the inclusion of lifestyle and demographic factors that are often underestimated in existing predictive models, even though these factors have a significant influence on the risk of cardiovascular disease.

In this context, predictive analysis of cardiovascular risk is an important step in the prevention and management of heart disease[4]. Using statistical techniques such as Logistic Regression, Lasso Regression, and Ridge Regression, researchers can identify factors that contribute to an increased risk of cardiovascular disease and develop models that can predict the likelihood of an individual experiencing a cardiovascular event[4][5].

By combining these three methods, researchers can develop robust and accurate predictive models for cardiovascular risk that consider a wide range of risk factors, from demographics to clinical biomarkers and patient lifestyle[6]. These models can be used by healthcare practitioners to identify patients at high risk and design targeted interventions to reduce that risk[6].

It is also hoped that this model can be implemented for cardiovascular AI screening. Cardiovascular *screening* AI is the process of screening for cardiovascular disease (CVD) risk using artificial intelligence (AI). AI is trained with large health data to recognize patterns and predict the likelihood of a person developing cardiovascular disease.

Advances in artificial intelligence (AI) technology have opened the door for the development of more accurate and cost-effective models for CVD screening. Using advanced machine learning algorithms, AI can rapidly process medical data and generate more precise CVD risk predictions. This provides an opportunity to improve the effectiveness of screening and intervention, thereby reducing the overall negative impact of cardiovascular disease.

1.2 Problem Formulation

a. How do lifestyle, demographic, and medical testing variables and demographic distribution affect cardiovascular risk (CVD)?
b. How can these predictive models assist healthcare practitioners in identifying patients at high cardiovascular risk and designing targeted interventions?
c. How does logistic regression, lasso regression, and ridge regression compare in cardiovascular disease (CVD) risk detection?

1.3 Problem Limitation

a. The data used to build the diagnosis model may not be representative of the entire population.
b. The diagnosis classification model may not be generalizable to all populations.
c. This research will focus on the use of logistic regression, Lasso regression, and Ridge regression as analytical methods for predicting in populations that have a high risk of developing cardiovascular disease. These methods were chosen for their ability to handle *multicollinearity* and overfitting issues, as well as in selecting the most relevant variables in the model.

1.4 Research Objectives

Recent research in cardiovascular health has shown that the risk of cardiovascular disease is not only determined by medical testing factors alone, but is also significantly influenced by lifestyle and demographic factors. Therefore, in an effort to improve the accuracy of predictive models related to cardiovascular risk, it is imperative to consider the contribution of these factors.

In identifying lifestyle and environmental factors that contribute to cardiovascular risk, studies have shown that habits such as unhealthy diet, lack of physical activity, smoking, and exposure to air pollution and other environmental stressors can significantly increase the risk of cardiovascular disease. Therefore, including these variables in predictive models may help in identifying high-risk individuals and enable the adoption of more appropriate and personalized prevention strategies.

By developing more comprehensive models that incorporate information on lifestyle, environmental and demographic factors, we can improve the accuracy of cardiovascular risk prediction at the individual level. Through this approach, we can pay greater attention to primary prevention and more targeted interventions, thereby reducing the overall burden of cardiovascular disease in the population. Thus, the development of predictive models that take these aspects into account could potentially have a positive impact on the prevention and management of cardiovascular disease in the future.

This modeling will be implemented in AI screening with the promising goal of increasing access to cardiovascular disease (CVD) screening for the public at large. With AI technology that can be implemented on software and *smartphone* apps, CVD screening becomes more accessible to individuals everywhere. This opens up opportunities for earlier detection of CVD, which in turn increases the chances for successful treatment. With early detection through AI screening, interventions can be made faster, helping to prevent serious complications and improve the quality of life of patients affected by CVD.

1.5 Research Benefits

a. **Improving the classification accuracy of cardiovascular diagnoses:** A comprehensive classification model can help identify patients at high risk of CVD more accurately than traditional models. This enables more effective and efficient interventions to prevent disease.
b. **Improving patient quality of life:** CVD prevention can significantly improve patients' quality of life by reducing complications and morbidity.
c. **Develop cost-effective AI models:** AI models can be trained on large datasets for pattern learning and CVD detection. These models can be implemented in accessible and affordable software and applications. In addition, low-cost AI screening can help improve access to CVD screening for the wider community, including in rural and remote areas.

1.6 Hypothesis

a. Lifestyle factors, such as smoking, physical activity, and alcohol consumption, have a significant influence on cardiovascular risk.
b. A comprehensive cardiovascular diagnosis classification model can improve patients' quality of life and reduce the burden on the healthcare system.
c. Lasso and Ridge regression tend to produce more stable and accurate cardiovascular disease (CVD) diagnosis models than logistic regression. This is because they are able to handle multicollinearity and overfitting more effectively, which can improve the consistency and accuracy of CVD risk prediction.

II. THEORETICAL FOUNDATION

2.1 Literature Study

I. Logistic Regression

Logistic regression is a statistical method used to predict a categorical dependent variable (usually with two possibilities) based on its relationship with one or more independent variables. The categorical dependent variable is often referred to as the target variable or outcome variable[5][6]. Here are the steps of how the logistic regression algorithm works:

1. Logistic regression uses a mathematical function called the logistic function, also known as the logit function. This function maps a linear predicted value (resulting from the independent variable) into a probability between 0 (unlikely) and 1 (certain).
2. Constructed by finding the coefficient (weight) for each independent variable. This coefficient affects the amount of contribution each variable makes to the final prediction.
3. The model is trained using historical data. The algorithm will continuously adjust the coefficients to minimize the difference between the predicted probability and the actual value of the dependent variable in the training data.
4. Once the model is trained, new data that has not been seen before can be entered. The model will predict the probability of an event occurring (according to the categorical dependent variable) based on the characteristics of the new data. The linear output of the model with probabilities is associated with a sigmoid function:

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

Sigmoid Function Formula

- x= input value
- y= predicted output
- b0 = bias or intercept term
- b1 = coefficient for input(x)

This function has the shape of an S curve and maps any real value to a probability value between 0 and 1. The output of the logistic regression model before the sigmoid function is a linear value that represents a linear combination of the input features. The sigmoid function converts these linear values to probabilities between 0 and 1. Output values closer to 1 indicate a higher probability of being classified as 1, while values closer to 0 indicate a higher probability of being classified as 0.

II. Lasso Regression

Lasso regression (*Least Absolute Shrinkage and Selection Operator*) is a regularization method used to overcome multicollinearity problems and improve the prediction accuracy of regression models[7]. Lasso adds a penalty to the loss function based on the absolute value of the regression coefficients [6]. This penalty pushes insignificant coefficients to zero, resulting in a model that is more compact and easier to interpret[7][8]. Here is how Lasso works:

1. LASSO shrinks the regression coefficients of the predictor variables to near or even zero.
2. During the optimization process, regression coefficients that do not

contribute significantly to the model will be depreciated towards zero. This means that some coefficients will become zero and the corresponding variables will be eliminated from the model.

3. To determine the optimal value of (β:), cross-validation techniques, such as k-fold cross-validation, are usually used. This technique helps in evaluating the performance of the model on datasets that are not involved in training the model:

$$\text{minimize}\left(||\mathbf{y} - \mathbf{X}\beta||_2^2 + \alpha||\beta||_1\right)$$

Figure 2. Lasso formula

y = vector of observed values

X = matrix of observed values

β: = coefficient of the vector

Estimated

α: penalty parameter

Finding the optimal value for α is essential to strike a good balance between model complexity and performance. Techniques such as cross-validation can be used to evaluate model performance on unseen data for different α values, helping you choose the most appropriate one for your specific scenario.

In summary, the Lasso regression objective function balances data fitting with a penalty for large coefficients, which is controlled by the regularization parameter α. This helps achieve a balance between model complexity and generalizability.

III. Ridge Regression

Ridge regression, also known as Tikhonov regularization, is a technique used to analyze data subject to multicollinearity. Like Lasso, Ridge adds a penalty to the loss function, but in a slightly different way[9][10]. Here is the working mechanism of Ridge Regression:

1. Ridge adds a penalty to the loss function which is the square of the regression coefficient values.
2. The goal of Ridge Regression is to minimize this modified loss function. The loss function is usually the Sum of Squared Residuals (SSR) plus the penalty of squaring the coefficients.
3. This process causes the regression coefficients to be depreciated, but never become completely zero as in Lasso. This means that all variables remain included in the model, but with adjusted effects.
4. Just like Lasso, the optimal β value is often determined using cross-validation to ensure that the model performs best on unseen data. The following is the SSE formula used in ridge regression:

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda\sum_{j=1}^{P}{\beta_j}^2$$

Figure 3. SSE formula

y = vector of observed values

β: = coefficient of the estimated vector

α: penalty parameter

Ridge regression is particularly useful when the dataset has many correlated predictor variables[10]. By penalizing the coefficients, Ridge helps reduce the complexity of the model without eliminating variables completely, which can be beneficial if all variables have potential relevance to the response.

2.2 Literature Review (*Research Gap*)

Table 1. Literature Review

Penelitian	Algoritma	Populasi	Label	Features	Accuracy	Jumlah Samples
A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL) *Mythili T, Dev Mukherji, Nikita Padalia, dan Abhiram Naidu (2021)*	Decission Tree, dan Random Forest, Logistic Regressiont	920	0 = tidak memiliki penyakit jantung 1 = memiliki penyakit jantung	Umur, Jenis Kelamin, Tekanan Darah, Tingkat Kolesterol, Tingkat Gula Darah, Berat Badan, riwayat penyakit	KNN, DT, RF,Logistic Regression, melalui SVM memperoleh akurasi 100%	736
Heart disease prediction using machine learning algorithms *Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain & Preeti Nagrath (2020)*	Logistic Regression, dan KNN, Random Forest Classifier	303 Pasien terdiagnosa penyakit jantung	0 = tidak memiliki penyakit jantung 1 = memiliki penyakit jantung	Umur, Jenis Kelamin, Tekanan Darah, Tingkat Kolesterol, Tingkat Gula Darah, Berat Badan, tekanan darah saat istirahat, Hasil tes stres thallium	logistik regresi dan KNN dan mendapatkan akurasi rata-rata 87,5% pada model	212
Pengklasifikasian Penyakit Jantung dengan Metode Decision Tree Indrayatna (2020)	KNN, Decision Tree	303 pasien penyakit jantung	0 = tidak memiliki penyakit jantung 1 = punya penyakit jantung	Umur, jenis kelamin, jenis nyeri dada, tekanan dara saat istirahat, jumlah pembuluh darah yang terdeteksi,tdepresi ST, kemiringan puncak ST	Evaluasi model dengan confusion matrix menghasilkan akurasi 75,4098%	242
Classification models for heart disease prediction using feature selection and PCA *Hassani (2020)*	CHI-PCA Random Forest, Decision Tree, Logistic Regression, MPC, Naive Bayes, Gradient-Boosted Trees (GBT)	920 pasien dari Cleveland	0 = punya penyakit jantung : 1,2,3,4 = tingkatan penyakit jantung	ID Pasien, usia, jenis kelamin, nyeri dada, status perokok/tidak, rokok perhari, riwayat oenyakit pada keluarga, detak jantung tertinggi, ada atau tidaknya penyakit jantung	CHI-PCA dengan RF memiliki kinerja maksimum 22 kinerja maksimum, dengan akurasi 98,7%	300
HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System *Fitriyani (2020)*	DBSCAN, SMOTE-ENN, dan XGBOOST	270 pasien penyakit jantung dari dataset I dan 297 pasien dari dataset II	0 = tidak memiliki penyakit jantung 1 = punya penyakit jantung	Umur, Jenis Kelamin, Tekanan Darah, Tingkat Kolesterol, Tingkat Gula Darah, Berat Badan, tekanan darah saat istirahat, Hasil uji stress thallium	Dataset I memiliki akurasi 88% sedangan dataset II memiliki akurasi 90%	216 dari dataset I dan 237 dari dataset II

2.3 Research Methodology

The stages carried out by the researcher can be explained through Figure.

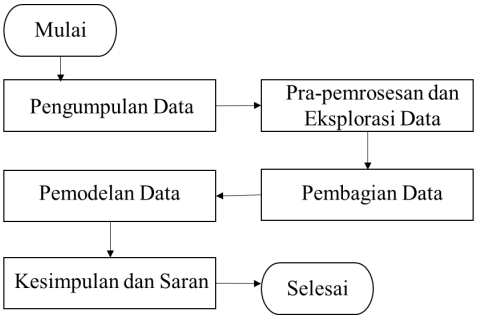

Figure 4: Stages of Research Methodology

The detailed explanation of each stage will be explained in the following sub-sections.

2.3.1 Data Collection

In this study, the dataset used came from a large health survey run by the CDC called BRFSS. They have been interviewing adults in the US since 1984, starting with 15 states and now covering the entire country. It is the largest ongoing health survey in the world, with over 400,000 interviews each year. The data used is from 2020 and includes factors that can influence heart disease and has a total of 31,9795 data points.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1548183 entries, 0 to 1548182
Data columns (total 28 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   HeartDisease      1548183 non-null  int64
 1   BMI               1548183 non-null  float64
 2   Smoking           1548183 non-null  int64
 3   AlcoholDrinking   1548183 non-null  int64
 4   Stroke            1548183 non-null  int64
 5   PhysicalHealth    1548183 non-null  float64
 6   MentalHealth      1548183 non-null  float64
 7   DiffWalking       1548183 non-null  int64
 8   Sex               1548183 non-null  int64
 9   Race              1548183 non-null  int32
 10  Diabetic          1548183 non-null  int64
 11  PhysicalActivity  1548183 non-null  int64
 12  GenHealth         1548183 non-null  int64
 13  SleepTime         1548183 non-null  float64
 14  Asthma            1548183 non-null  int64
 15  KidneyDisease     1548183 non-null  int64
 16  SkinCancer        1548183 non-null  int64
 17  Age               1548183 non-null  int64
 18  cp                1548183 non-null  int64
 19  trestbps          1548183 non-null  int64
 20  chol              1548183 non-null  int64
 21  fbs               1548183 non-null  int64
 22  restecg           1548183 non-null  int64
 23  thalach           1548183 non-null  int64
 24  exang             1548183 non-null  int64
 25  oldpeak           1548183 non-null  float64
 26  slope             1548183 non-null  int64
 27  ca                1548183 non-null  int64
dtypes: float64(5), int32(1), int64(22)
memory usage: 324.8 MB
```

Figure 5. Dataset info

The dataset shown is a collection of medical data that includes 1,548,183 entries with 28 different variables. Each column in this dataset represents a medical attribute or characteristic of the individuals involved in the dataset.

```
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   HeartDisease      1548183 non-null  int64
 1   BMI               1548183 non-null  float64
 2   Smoking           1548183 non-null  int64
 3   AlcoholDrinking   1548183 non-null  int64
 4   Stroke            1548183 non-null  int64
 5   PhysicalHealth    1548183 non-null  float64
 6   MentalHealth      1548183 non-null  float64
 7   DiffWalking       1548183 non-null  int64
 8   Sex               1548183 non-null  int64
 9   Race              1548183 non-null  int32
 10  Diabetic          1548183 non-null  int64
 11  PhysicalActivity  1548183 non-null  int64
 12  GenHealth         1548183 non-null  int64
 13  SleepTime         1548183 non-null  float64
 14  Asthma            1548183 non-null  int64
 15  KidneyDisease     1548183 non-null  int64
 16  SkinCancer        1548183 non-null  int64
 17  Age               1548183 non-null  int64
 18  cp                1548183 non-null  int64
 19  trestbps          1548183 non-null  int64
 20  chol              1548183 non-null  int64
 21  fbs               1548183 non-null  int64
 22  restecg           1548183 non-null  int64
 23  thalach           1548183 non-null  int64
 24  exang             1548183 non-null  int64
 25  oldpeak           1548183 non-null  float64
 26  slope             1548183 non-null  int64
 27  ca                1548183 non-null  int64
 28  age_group         1548183 non-null  category
dtypes: category(1), float64(5), int32(1), int64(22)
memory usage: 326.3 MB
```

Figure 6. Dataset Info

After tidying up, the dataset is a large data set consisting of 1,548,183 rows and 29 columns with an additional column in the form of an age group where the patient's age is grouped into three, namely the 21-39 age group, the 40-60 age group, and the over 61 age group.

```
# Lifestyle and Behavior
lifestyle_features = ['Smoking', 'AlcoholDrinking', 'PhysicalActivity', 'SleepTime']

# Medical Tests and Indicators
medical_tests = ['Stroke', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca']

# Demographic and Personal Information
demographic_info = ['Sex', 'Race', 'Age']

# Functional Abilities
functional_abilities = ['DiffWalking']

feature_cols = demographic_info + lifestyle_features + medical_tests
```

Figure 7. Clustering Code

After further sorting, the variables selected were those that were part of medical_tests, demographic_info, and lifestyle features. The following is an explanation of the features used:

Table 2. Identification of Lifestyle Factor Variables

Column Name	*Definition*
Smoking	*indicator of whether a person smokes or not (1 for "Yes" and 0 for "No")*
Alcohol Drinking	*an indicator of whether a person has alcohol consumption habits (1 for "Yes" and 0 for "No")*
Physical Activity	*Indicator of whether or not the person frequently engages in sports activities (1 for "Yes" and 0 for "No")*
SleepTime	*measurement of one's sleep time*

Table 3. Variable Identification of Demographic Factors

Column Name	*Definition*
Sex	*an indicator of a person's gender (1 for "male" and 0 for "female")*
AgeCategory	*grouping people into age ranges*
Race	*one's racial background*

Table 4. Identification of medical test Variables

Column Name	*Definition*
Heart Disease	*To express the final result, 1 is the presence of heart disease and 0 indicates the absence of heart disease.*
Stroke	*Whether or not there is a history of stroke*
cp	*4 types of chest pain*
trestbps	*blood pressure at rest*
chol	*cholesterol level in mg/dl*
fbs	*blood sugar level in mg/dl*
restecg	*the result of the electrocardiograph at rest (with values 0,1,2)*
thalach	*maximum heart rate achieved*
exang	*chest pain when exercising or experiencing emotional stress*
oldpeak	*Exercise-induced ST-segment depression is considered a reliable ECG finding for the diagnosis of obstructive coronary atherosclerosis*
slope	*peak training ST segment slope*
ca	*number of major blood vessels (0-3) stained with fluoroscopy*

Table 5: Identification of Dependent Variables

Column Name	*Definition*
Heart Disease	*states the presence or absence of cardiovascular disease in the patient (1 for "yes" and 0 for "no")*

These tables show which variables were selected for further data processing and modeling. These variables were selected because they contain factors in accordance with the objectives of this study, namely to identify the classification of cardiovascular diagnoses (CDV) through lifestyle factors and demographic distribution. Based on research by Widayata [11], lifestyle factors include smoking habits, alcohol consumption, sleep time, physical health, mental health, and body weight. Demographic factors consist of gender, age, and race. As for the medical test factor, it consists of various kinds such as the type of chest pain, history of stroke, number of

major blood vessels, and many more. The variables of lifestyle factors, demographics and medical tests will be used as independent variables, while the *Heart Disease variable* becomes the dependent variable.

2.3.2 Data Pre-processing and Exploration

Data pre-processing and exploration is a critical stage in data analysis that involves several important steps. First, data cleaning is performed to identify and address incomplete, duplicate, or inaccurate data. The next step is the handling of missing values, where techniques such as imputation or deletion of incomplete data can be applied accordingly. In addition, variable transformation is also required to change the format or scale of the variables to match the assumptions of the model to be used. This data pre-processing and exploration provides important insights in understanding the structure of the data and prepares a solid foundation for further analysis. The following are the results of the pre-processing and data exploration stages:

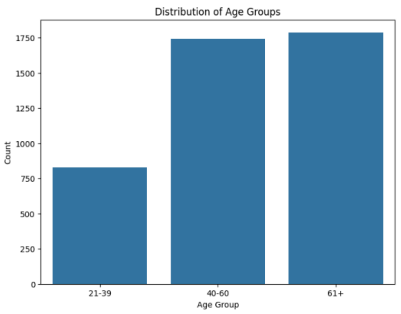

Figure 8. age distribution chart

The following is the age distribution of the people who took part in the survey. The largest age group was 61+ years old, while the smallest age group was 21-39 years old.
Figure 7: Heart disease distribution chart

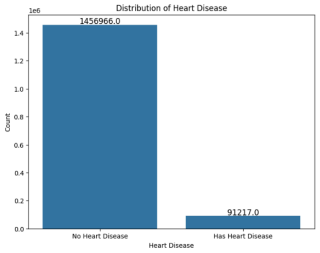

Figure 9: Heart disease distribution chart

The following is the distribution of heart disease or another word is cardiovascular. According to the visualization, there are 145 thousand patients who do not have cardiovascular disease, while there are 91 thousand patients who have heart disease.

```
Missing values in each column:
HeartDisease        0
BMI                 0
Smoking             0
AlcoholDrinking     0
Stroke              0
PhysicalHealth      0
MentalHealth        0
DiffWalking         0
Sex                 0
Race                0
Diabetic            0
PhysicalActivity    0
GenHealth           0
SleepTime           0
Asthma              0
KidneyDisease       0
SkinCancer          0
Age                 0
```

Figure 10. missing value of each variable

The checking results show that there are no missing values in the dataset. Next is the outlier checking stage by removing non-numeric variables in the dataset. Outlier checking using the *interquartile* value mechanism can be seen in Figure 4.

```
# Menghitung IQR dan menghapus outlier dari dataset

Q1 = df[columns_of_interest].quantile(0.25)
Q3 = df[columns_of_interest].quantile(0.75)
IQR = Q3 - Q1

outlier_threshold = 1.5

outlier_mask = ~((df[columns_of_interest] < (Q1 - outlier_threshold * IQ

df_no_outliers = df[outlier_mask]

sns.boxplot(data=df_no_outliers[columns_of_interest])
plt.show()
```

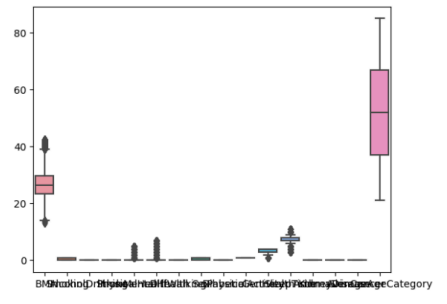

Figure 11. outliers for each variable

The box plot displays the normalized data. Using IQRs and thresholds, it is possible to identify data points that substantially deviate from the rest of the data and potentially create skewed data results.

```
# Pemeriksaan Jumlah Penyakit Jantung
heart_disease_counts = df_no_outliers['HeartDisease'].value_counts()

print("Heart Disease Counts:")
print(heart_disease_counts)
```

```
Heart Disease Counts:
0    89831
1     3868
Name: HeartDisease, dtype: int64
```

figure 12. data distribution

The dataset has 3868 patients with cardiovascular disease, while 89831 patients do not have it.

2.3.3 Data Split (about train test split,)

At this stage, the process of dividing data into train data and test data is carried out. The goal is to measure the goodness of the model formed. The 80:20 ratio is used in this research, which means 80% of the data will be used as train data and the remaining 20% as test data. This data division is expected to be able to train machine learning models to improve accuracy.

2.3.4 **Data** Modeling

Logistic Regression

```
param_grid_logistic = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear', 'saga', 'newton-cg', 'lbfgs'],
}
```

```
from sklearn.model_selection import GridSearchCV, KFold
modelLogistic = LogisticRegression(max_iter=1000)
kf = KFold(n_splits=5, shuffle=True, random_state=42)
```

```
grid_search_logistic = GridSearchCV(modelLogistic, param_grid_logistic,
grid_search_logistic.fit(X_train, y_train)
```

```
:2]:
GridSearchCV
estimator: LogisticRegression
LogisticRegression
```

```
best_model = grid_search_logistic.best_estimator_
```

```
y_pred = best_model.predict(X_test)
```

Figure 13. logistic regression modeling

Figure 14. lasso regression modeling

Figure 15. ridge regression modeling

This grid search process implemented in logistic, ridge, and lasso regression analysis explores various combinations of these hyperparameters and evaluates each combination using 5-fold cross validation. The technique divides the training data into five folds, trains the model on four folds, and tests it on the remaining fold. This process is repeated five times, to ensure robust model performance across different subsets of data. In the end, the grid search selects the hyperparameter combination that yields the best performance according to the chosen scoring matrix (ROC AUC in this case). This approach helps to find logistic regression models that tend to generalize well to unseen data, making accurate predictions on future instances.

```
Lasso Regression Accuracy: 0.9797421731123389
Classification Report - Lasso Model:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98       295
           1       0.96      1.00      0.98       248

    accuracy                           0.98       543
   macro avg       0.98      0.98      0.98       543
weighted avg       0.98      0.98      0.98       543


Ridge Regression Accuracy: 0.9797421731123389
Classification Report - Ridge Model:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98       295
           1       0.96      1.00      0.98       248

    accuracy                           0.98       543
   macro avg       0.98      0.98      0.98       543
weighted avg       0.98      0.98      0.98       543


Accuracy: 0.9797421731123389
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98       295
           1       0.96      1.00      0.98       248

    accuracy                           0.98       543
   macro avg       0.98      0.98      0.98       543
weighted avg       0.98      0.98      0.98       543
```

Figure 16. accuracy report of each model

Based on the accuracy results of the three regression models evaluated, it can be concluded that all three models (Logistic, Lasso, and Ridge Regression) show relatively good performance in predicting the presence of heart disease based on accuracy, precision, recall, and F1-score. All of them have the same accuracy of 98%.

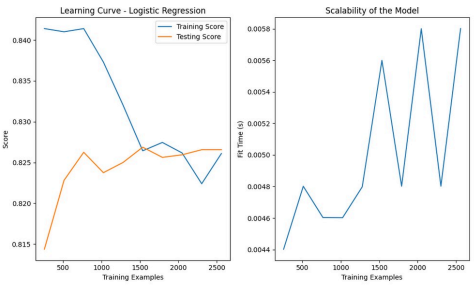

Figure 17. learning curve graph of logistic regression modeling

The figure above displays two graphs related to the logistic regression model. The first graph "Learning Curve", shows that the test scores decrease and then start to rise again, while the training scores increase, indicating that the model is overfitting. The second graph "Model Scalability", illustrates that the time taken to fit the model increases with significant

fluctuations as the number of training samples increases, indicating higher computational complexity when the model is faced with more data.

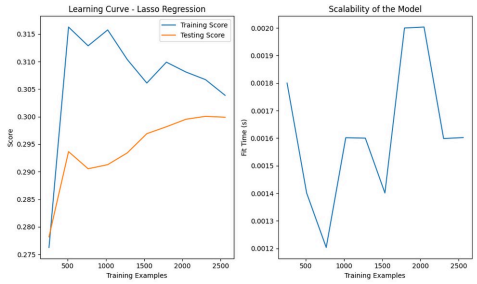

Figure 18. learning curve graph of lasso regression modeling

The figure shows 2 graphs related to the lasso regression model. The "Learning curve" graph shows that the training score decreases slightly and the testing score increases as the number of training samples increases, indicating that the model starts to stabilize after overfitting initially. The "Model Scalability" graph depicts the time taken to fit the model to the number of training samples, it also shows that the computational complexity of the model increases as the data grows.

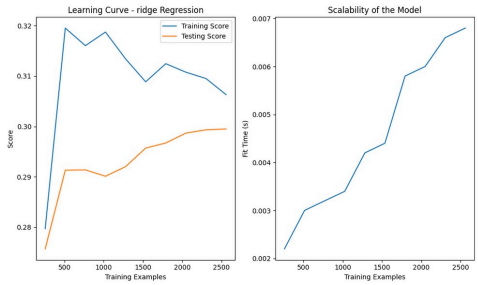

Figure 19. learning curve graph of ridge regression modeling

The graph shows the Learning Curve graph of the ridge model where the training value line increases drastically then according to slowly, the testing value line continues to increase. This identifies that the model is underfitting.

Accuracy Comparison Table

Model	*Train Size (ratio)*	*Test Size (ratio)*	*Accuracy*
Logistic Regression	*80%*	*20%*	*0.981188*
	70%	*30%*	*0.991413*
	60%	*40%*	*0.975731*
LASSO Regression	*80%*	*20%*	*0.907921*
	70%	*30%*	*0.858653*
	60%	*40%*	*0.861317*
Ridge Regression	*80%*	*20%*	*0.986139*
	70%	*30%*	*0.97358*
	60%	*40%*	*0.979198*

This table summarizes the performance of three different machine learning models-Logistic Regression, Lasso, and Ridge-at various training and test set sizes for the classification task. The Logistic Regression model showed outstanding performance, especially with a 70% training set size, achieving 99.14% accuracy. The model consistently maintained high accuracy across other set sizes. The Ridge model also showed strong accuracy, peaking at 98.61% when the training set size was 80%, and generally performed strongly across all configurations tested. In contrast, the Lasso model performed poorly compared to the other two models, with the highest accuracy at 90.79% when the largest training set was at 80%. Nonetheless, this still shows a general trend where a larger training set slightly improves accuracy, a pattern observed across all models. This comparison highlights the effectiveness of Logistic Regression and Ridge in handling classification tasks, with the Logistic Regression model slightly outperforming Ridge in most configurations. Lasso, while still useful, may not be as effective in this particular scenario, possibly due to its regularization method which may cause mismatches in this context.

Accuracy Comparison Table Using Dummy Testing

Model	*Train Size (ratio)*	*Test Size (ratio)*	*Accuracy*
Logistic Regression	*80%*	*20%*	*0.974702*
	70%	*30%*	*0.977183*
	60%	*40%*	*0.980655*
LASSO Regression	*80%*	*20%*	*0.626488*
	70%	*30%*	*0.624008*

	60%	*40%*	*0.629464*
Ridge Regression	*80%*	*20%*	*0.974702*
	70%	*30%*	*0.977183*
	60%	*40%*	*0.980655*

This table provides a comparative analysis of three machine learning models - Logistic Regression, LASSO, and RIDGE using dummy testing across different sizes of training and testing sets on a classification task. The results show that the Logistic and RIDGE models perform consistently well, achieving high accuracy levels that increase slightly with larger training sets. Specifically, the accuracy for these models ranged from 97.47% to 98.07% as the training set size increased from 60% to 80%. In contrast, the LASSO model showed much lower accuracy, with the figure fluctuating around 62.5%, regardless of the training/testing set ratio. This stark difference in performance underscores the effectiveness of the Logistic and RIDGE models in handling this particular classification task compared to the LASSO model, which may have a poor fit to the data. Notably, the trend across all models shows that larger training sets slightly improve model accuracy, which is a common observation in machine learning scenarios.

III. SUGGESTIONS AND CONCLUSIONS

3.1 Conclusions and Suggestions

Evaluation of the three regression models showed that they performed well in predicting the presence of heart disease. Logistic Regression dominated with the highest accuracy, reaching 99.14% in the 70% training set and 30% test set configuration, showing consistent superiority across all training and test set sizes. Meanwhile, Ridge Regression also performed very well, recording a peak accuracy of 98.61% with a training set of 80%. On the other hand, Lasso Regression, despite having a lower performance with its best accuracy of 90.79% at 80% training set, still provided decent results. Therefore, Logistic Regression and Ridge Regression are considered as the most suitable models to be used in heart disease screening AI due to their high reliability and consistency.

To further improve model performance, several strategies can be applied. First, ensemble learning techniques that combine multiple regression models can provide more accurate predictions. Secondly, applying feature engineering techniques to select and process appropriate data features can improve the effectiveness of the model. Third, optimal hyperparameter tuning can improve model performance. Finally, the use of a wider dataset for training can strengthen model predictions.

Researchers can conduct further studies to collect and analyze large amounts of patient data to train and validate predictive models of CVD risk. In addition, exploring more advanced AI techniques such as deep learning may also help to further improve prediction accuracy.

LITERATURE

[1] S. Vanakovarayan *et al.*, "Heart disease prediction using linear regression techniques," *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Nov. 2023. doi:10.1109/icscan58655.2023.10394924

[2] S. Naveen, S. K. Ravindran, S. G, and S. N. Ameen, "Effective heart disease prediction framework using random forest and logistic regression," *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, May 2023. doi:10.1109/vitecon58111.2023.10157078

[3] L. C. Godoy *et al.*, "Predicting left main stenosis in stable ischemic heart disease using logistic regression and boosted trees," *American Heart Journal*, vol. 256, pp. 117-127, Feb. 2023. doi:10.1016/j.ahj.2022.11.004

[4] A. Khang, G. Rana, R. K. Tailor, and V. Abdullayev, *Data-centric AI solutions and emerging technologies in the healthcare ecosystem*, Aug. 2023. doi:10.1201/9781003356189

[5] D. Stojanov, E. Lazarova, E. Veljkova, P. Rubartelli, and M. Giacomini, "Predicting the outcome of heart failure against chronic-ischemic heart disease in elderly population - machine learning approach based on logistic regression, case to Villa Scassi Hospital Genoa, Italy," *Journal of King Saud University - Science*, vol. 35, no. 3, p. 102573, Apr. 2023. doi:10.1016/j.jksus.2023.102573

[6] W. Shi *et al.*, "Systematic review, meta-analysis and meta-regression to determine the effects of patient education on health behavior change in adults diagnosed with coronary heart disease," *Journal of Clinical Nursing,* vol. 32, no. 15-16, pp. 5300-5327, Sep. 2022. doi:10.1111/jocn.16519

[7] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, p. 100130, Nov. 2023. doi:10.1016/j.health.2022.100130

[8] G. K. G., "Analysis of accuracy in heart disease diagnosis system using decision tree classifier over logistic regression based on recursive feature selection," *ECS Transactions*, vol. 107, no. 1, pp. 15661-15674, Apr. 2022. doi:10.1149/10701.15661ecst

[9] K. M. Kumar and P. Uma, "Accuracy analysis of heart disease prediction using logistic regression in comparison with the linear regression algorithm," *Journal of Pharmaceutical Negative Results*, vol. 13, no. SO4, Jan. 2022. doi:10.47750/pnr.2022.13.s04.199

[10] J. Wang and L. Li, "Letter to the editor 'systematic review, meta-analysis and meta-regression to determine the effects of patient education on health behavior change in adults diagnosed with coronary heart disease,'" *Journal of Clinical Nursing*, Feb. 2024. doi:10.1111/jocn.17078

[11] A. Widayati, F. Fenty, Y. Linawati, and P. D. Christasani, "Knowledge and Profile of Healthy Lifestyle in Rural Adults in Yogyakarta Special Region," Indonesian Journal of Clinical Pharmacy, vol. 9, no. 2, p. 118, Jun. 2020, doi: https://doi.org/10.15416/ijcp.2020.9.2.118.