University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Distributed & Scalable Data Engineering(DSCI-6007-03)

TECHNICAL REPORT

Spring 2023

CONTENTS

Cardiovascular disease prediction using AWS Tool

Executive Summary

- Cardiovascular disease (CVD) is a leading cause of death worldwide, and early identification of individuals at risk can help prevent its development and progression.

- We describe the approach and methods used to develop a CVD prediction model to reduce the number of features from 12 to 6 for easy identification of disease by using AWS tools.

Team Members:

Hawar Dzaee
Team Leader, Application developer

Divya Soma
Data Scientist

Anjali Yadagiri
Data Modeling

Kaja Mohan Manikanta Reddy
Data visualization

HIGHLIGHTS OF THE PROJECT

- Our CVD prediction project with AWS tools included the use of the Medical dataset, the generation of Dimensionality Reduction using Sage Maker, and the deployment of the model using AWS Sage Maker, with a training accuracy of 0.95. set.
- The medical dataset has a large number of features, and we reduced the dimension reduction in the PCA algorithms by applying unsupervised learning.

ABSTRACT

- The development of a CVD prediction model using AWS tools. We utilized the Medical dataset, preprocessed the data using AWS Glue, and developed and trained a Dimensionality reduction using Sage Maker.
- The model was evaluated using various metrics and achieved an accuracy of 0.95 on the testing set. Finally, we deployed the model using AWS Sage Maker. Overall, this project demonstrates the potential of AWS tools for developing and deploying machine learning

GitHub link is attached below –https://github.com/SomaDivya1999/TEAM-04.git

METHODOLOGY

- We take the dataset from Kaggle the objectives were factual information then examinations are taken from results of Medical examination.
- Dimensionality reduction use SageMaker's built-in PCA (Principal Component Analysis) algorithm to reduce the number of features in the dataset while retaining most of the information.

- We would be evaluating the performance of the model to reduce the time, dimension, storage and further interference.

BUSINESS UNDERSTANDING

Cardiovascular disease is a leading cause of mortality worldwide, and its prevention is a critical public health concern. In this context, data analytics has emerged as an essential tool for predicting the risk of cardiovascular disease, enabling early intervention and better management of the disease.

To make sense of this data, we have used principal component analysis (PCA), a technique commonly used for dimensionality reduction. By reducing the dataset from five dimensions to three, you can improve the interpretability of the data and identify the most important features for predicting cardiovascular disease.

Reducing the dimensionality of the dataset will have several benefits for your business.

First, it will simplify the data and make it more manageable for analysis. With fewer dimensions to consider, it will be easier to identify patterns and relationships in the data and to develop predictive models that accurately capture the risk of cardiovascular disease. Second, reducing the dataset's dimensionality will help you identify the most important features for predicting cardiovascular disease.

By focusing on the six most essential dimensions, you can develop a more targeted and practical approach to preventing and managing cardiovascular disease. This could include interventions such as lifestyle changes, medication, or surgical procedures, depending on the individual risk profile of each patient.

Overall, your decision to use PCA to reduce the dimensionality of your cardiovascular disease dataset is a sound one. By simplifying the data and identifying the most critical features, you can develop more effective strategies for preventing and managing this serious health condition.

PROJECT REQUIREMENTS

The technology used in this project is AWS ecosystem, more generally, AWS Sage maker. We used Sage maker notebook instance to run our Python code. For Data preprocessing, we used Panda's library. For interacting with AWS services through our notebook instance, we took advantage of boto3, such as an S3 bucket. We used NumPy to prepare the data for serialization. The 'sage maker' library allowed us to start a session, fetch the IAM role and necessary dependencies for the algorithm, train the model, and finally deploy the model.

DATA UNDERSTANDING

For this project, we will be extracting the customer data sets from Kaggle. It is a collection of Medical examinations that can be used for the purpose of Data predictions.

In [2]:
```
# read the csv file
cardio_df = pd.read_csv("cardio_train.csv", sep=";")
```

In [3]:
```
cardio_df.head()
```

Out[3]:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

DATA PREPARATION

Drop irrelevant column (id). Change the ‘age in days’ column to ‘age in years’

- Check for null values.
- Dropping outliers, using 3 z-score as a threshold.
- Drop CVD column as needed for supervised learning
- Drop categorical data.PCA works best with numerical data
- Dataset overview
- scaling the data

```
In [8]: cardio_df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 55699 entries, 0 to 69999
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   age          55699 non-null  float64
 1   gender       55699 non-null  int64
 2   height       55699 non-null  int64
 3   weight       55699 non-null  float64
 4   ap_hi        55699 non-null  int64
 5   ap_lo        55699 non-null  int64
 6   cholesterol  55699 non-null  int64
 7   gluc         55699 non-null  int64
 8   smoke        55699 non-null  int64
 9   alco         55699 non-null  int64
 10  active       55699 non-null  int64
 11  cardio       55699 non-null  int64
dtypes: float64(2), int64(10)
memory usage: 5.5 MB
```

```
In [9]: cardio_df.describe()
```

Out[9]:

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	
count	55699.000000	55699.000000	55699.000000	55699.000000	55699.000000	55699.000000	55699.000000	55699.000000	55699.0	55699.0	55699.000000	5569
mean	53.255880	1.294206	163.892476	72.919929	125.643656	81.011077	1.270148	1.075944	0.0	0.0	0.799494	
std	6.769454	0.455689	7.566255	13.016990	17.752886	10.207992	0.576480	0.264911	0.0	0.0	0.400383	
min	39.128767	1.000000	140.000000	32.000000	-150.000000	-70.000000	1.000000	1.000000	0.0	0.0	0.000000	
25%	48.282192	1.000000	158.000000	64.000000	120.000000	80.000000	1.000000	1.000000	0.0	0.0	1.000000	
50%	53.917808	1.000000	164.000000	71.000000	120.000000	80.000000	1.000000	1.000000	0.0	0.0	1.000000	
75%	58.378082	2.000000	169.000000	80.000000	140.000000	90.000000	1.000000	1.000000	0.0	0.0	1.000000	
max	64.967123	2.000000	188.000000	117.000000	309.000000	602.000000	3.000000	2.000000	0.0	0.0	1.000000	

Data serialization

- Writing the NumPy array df_matrix to the Bytes IO object buffer allows us to store the array's binary data in memory as a bytes object. This can be useful for various purposes such as transmitting the data over a network, saving it to a file, or passing it to other functions or processes. In this specific case, the binary data stored in the Bytes IO object buffer is being used as input to a machine learning algorithm that requires the data to be in a particular format or passed through a specific API. Storing the data as a byte's object in memory can be more efficient than writing it to a file or transmitting it over a network.

MODELLING

- Data And Model Artifact Pathway
- Retrieve The Container Image URI For The "PCA" Algorithm
- Training The Model Using Sage maker Library

EVALUATION

Once the machine learning models are trained, they are trained using dimensionality reduction. We will reduce the features by using PCA algorithm's the evaluation metrics reduce the dimensions, time and storage.

DEPLOYMENT

This application will be deployed using the trained PCA models