NITTE (Deemed to be University) | NMAM INSTITUTE OF TECHNOLOGY

Nitte (DU) established under Section 3 of UGC Act 1956 | Accredited with 'A+' Grade by NAAC

Department of Computer Science and Engineering

Report on Mini Project

Medical Appointment Analysis

Course Code: 20CSE81

Course Name: Big Data Analysis

Semester: VI SEM

Section: C

Submitted To:

Mr. Sampath Kini
Assistant Professor Gd-II
Department of Computer Science and Engineering

Submitted By:

Rohit Anil Rao- 4NM20CS146
Sandeep Shetty- 4NM20CS153

Date of submission: 07/05/2023

Signature of
Course Instructor

NITTE EDUCATION TRUST | N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India
(ISO 9001:2015 Certified), Accredited with 'A' Grade by NAAC
☎: 08258 - 281039 - 281263, Fax: 08258 - 281265

Department of Computer Science and Engineering
B.E. CSE Program Accredited by NBA, New Delhi from 1-7-2018 to 30-6-2021

CERTIFICATE

"Medical Appointment Analysis" is a bonafide work carried out by Rohit Anil Rao (4NM20CS146) and Sandeep Shetty (4NM20CS153) in partial fulfillment of the requirements for the award of a Bachelor of Engineering Degree in Computer Science and Engineering 2022-2023.

It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report. The Mini project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the Bachelor of Engineering Degree.

Signature of Guide Signature of HOD

ABSTRACT

In recent years, the use of big data analytics in the healthcare industry has gained widespread attention. In this project, we investigate a medical appointment dataset to gain insights into the factors that affect patient attendance at medical appointments. The dataset contains information on over 100,000 medical appointments across the world, including patient demographics, medical history, appointment scheduling information, and whether the patient showed up for the appointment or not. We use data exploration, cleaning, visualization, and statistical analysis techniques to gain a better understanding of the dataset and identify patterns and relationships between different variables. Our findings suggest that several factors, such as age, gender, the day of the week, and the time of the appointment, have a significant impact on patient attendance. We also identify potential areas for further research, such as the impact of reminder notifications on patient attendance. Our project highlights the potential of big data analytics in improving healthcare outcomes and patient experiences.

TABLE OF CONTENTS

INTRODUCTION

The healthcare industry generates a vast amount of data, including electronic health records, administrative claims, and patient-generated data. This data can provide valuable insights into patient behavior, disease patterns, and treatment outcomes, leading to improved healthcare services and better patient outcomes. One area where big data analytics can make a significant impact is in improving patient attendance at medical appointments.

Missed medical appointments are a significant issue in healthcare systems worldwide. They can lead to delayed or inadequate care, increased healthcare costs, and decreased patient satisfaction. For healthcare providers, missed appointments can also result in lost revenue and wasted resources. Therefore, understanding the factors that influence patient attendance is crucial for healthcare providers to deliver high-quality care and optimize their resources.

In this project, we investigate a medical appointment dataset to gain insights into the factors that affect patient attendance. The dataset contains information on over 100,000 medical appointments in Brazil, including patient demographics, medical history, appointment scheduling information, and whether the patient showed up for the appointment or not. We use various data analysis techniques, including data exploration, cleaning, visualization, and statistical analysis, to identify patterns and relationships between different variables.

Our project aims to identify the factors that influence patient attendance and provide recommendations for healthcare providers to optimize their appointment systems. By analyzing the dataset, we aim to answer the following research questions:

- What are the demographic and health-related factors that influence patient attendance at medical appointments?
- How does appointment scheduling information, such as the day of the week and the time of day, affect patient attendance?
- Can reminder notifications improve patient attendance at medical appointments?
- Are there any other factors that healthcare providers should consider when optimizing their appointment systems to improve patient attendance?

Through our project, we aim to demonstrate the potential of big data analytics in healthcare and provide recommendations for healthcare providers to improve their services and patient outcomes.

PROBLEM STATEMENT

By analyzing the Medical Appointment dataset, insights can be gained on various aspects of the restaurant In this notebook we will try to analyze why would some patient not show up for his medical appointment and whether there are reasons for that using the data we have.
We will try to find some correlation between the different attributes we have and whether the patient shows up or not. The dataset we are going to use contains 110.527 medical appointments and its 14 associated variables (PatientId, AppointmentID, Gender, ScheduledDay, AppointmentDay, Age, Neighborhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handcap', SMS_received, No_show)

OBJECTIVES

The main objectives of this project are as follows:

- What is the percentage of no-show?
- What factors are important for us to know in order to predict if a patient will show up for their scheduled appointment?
    - Is the time gender related to whether a patient will show or not?
    - Are patients with scholarship more likely to miss their appointment?
    - Are patients who don't receive SMS more likely to miss their appointment?
    - Is the time difference between the scheduling and appointment related to whether a patient will show?
    - Does age affect whether a patient will show up or not?
    - What is the percentage of patients missing their appointments for every neighborhood

HARDWARE / SOFTWARE Requirements

Hardware Requirements:

- A computer with at least 8GB RAM and a multi-core processor
- Sufficient storage space to store the dataset and model files

Software Requirements:

- Python 3.x installed on the computer
- Jupyter Notebook installed on the computer
- Relevant Python libraries such as Pandas, NumPy, Matplotlib, Scikit-learn, and Seaborn installed in the Python environment

METHODOLOGY

The methodology for medical appointment analysis can be divided into the following steps:

1. Data Collection:

   The first step in any data analysis project is to collect relevant data. In the case of medical appointment, data can be obtained from various sources such as the internet, API, and other third-party data sources. The data can include information such as PatientId**,** AppointmentID**,** Gender, ScheduledDay, AppointmentDay, Age, Neighborhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handcap', SMS_received, No_show.

2. Data Cleaning:

   After collecting the data, it is important to clean and pre-process it. This involves identifying and correcting errors or inconsistencies in the data, such as missing values, incorrectspellings, and outliers. The data can be cleaned using tools such as Python or R programming languages.

3. Data Integration:

   The medical appointment analysis project may involve integrating data from multiple sources. In this case, it is important to ensure that the data is in the same format and can be merged seamlessly.

4. Data Transformation:

   Data transformation involves converting the data into a format that is suitable for analysis. This may involve standardizing the data, normalizing it, or scaling it to ensure that it is comparable.

5. Data Sampling:

   If the dataset is too large, it may be necessary to take a sample of the data to perform analysis. This can be done randomly or based on certain criteria.

6. Data Visualization:

   Data visualization techniques can be used to gain insights into the data and identify patterns and trends. Various charts, graphs, and plots can be used to visualize the data.

IMPLEMENTATION

1. The following Libraries are to be installed and imported to run the project:

```
In [1]:
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

sns.set_style('darkgrid')

%matplotlib inline
```

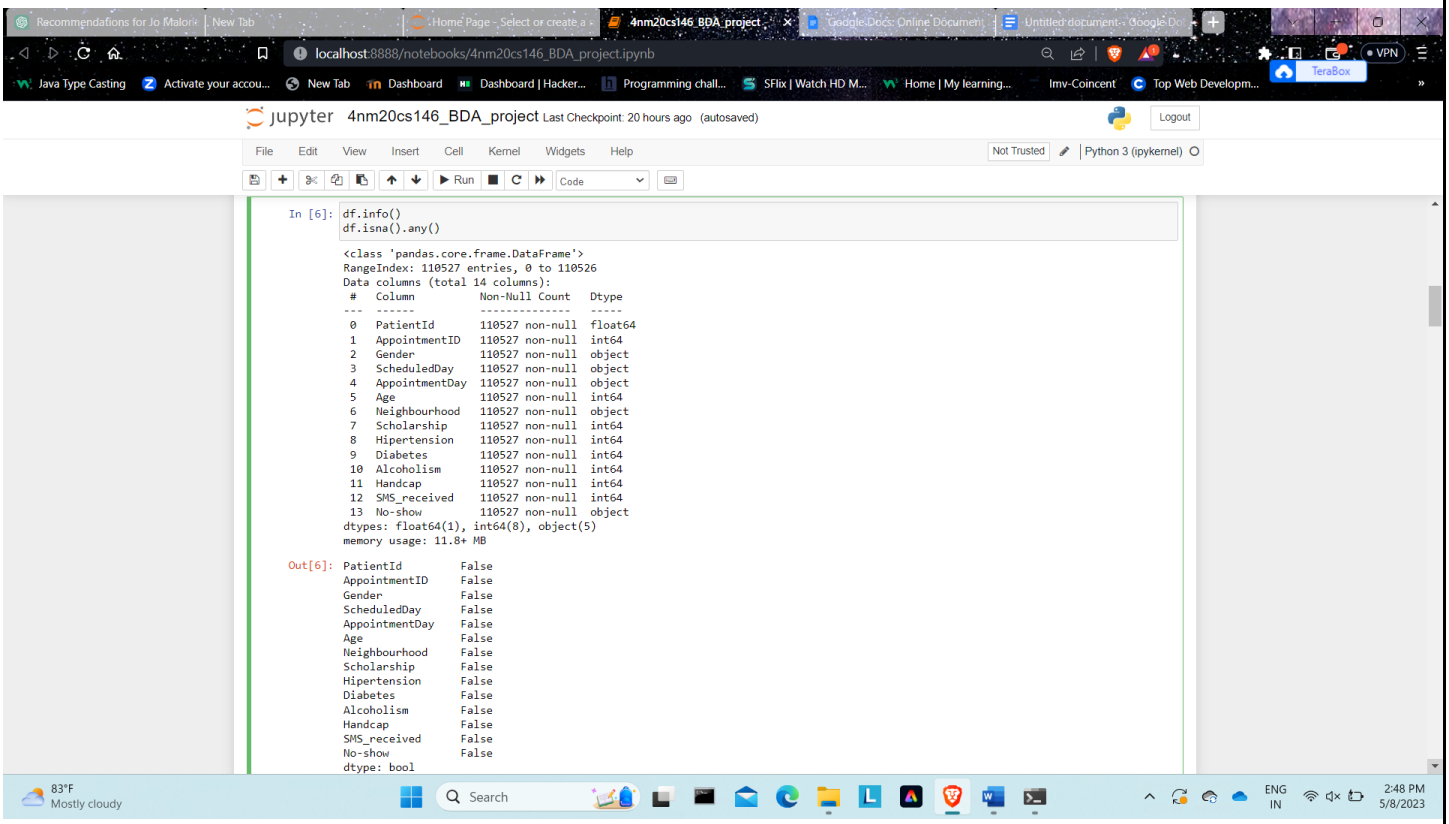
2. The required dataset is being loaded and the first 10 lines are being printed:

```
In [3]: df = pd.read_csv('noshowappointments.csv')
        df.head()
```

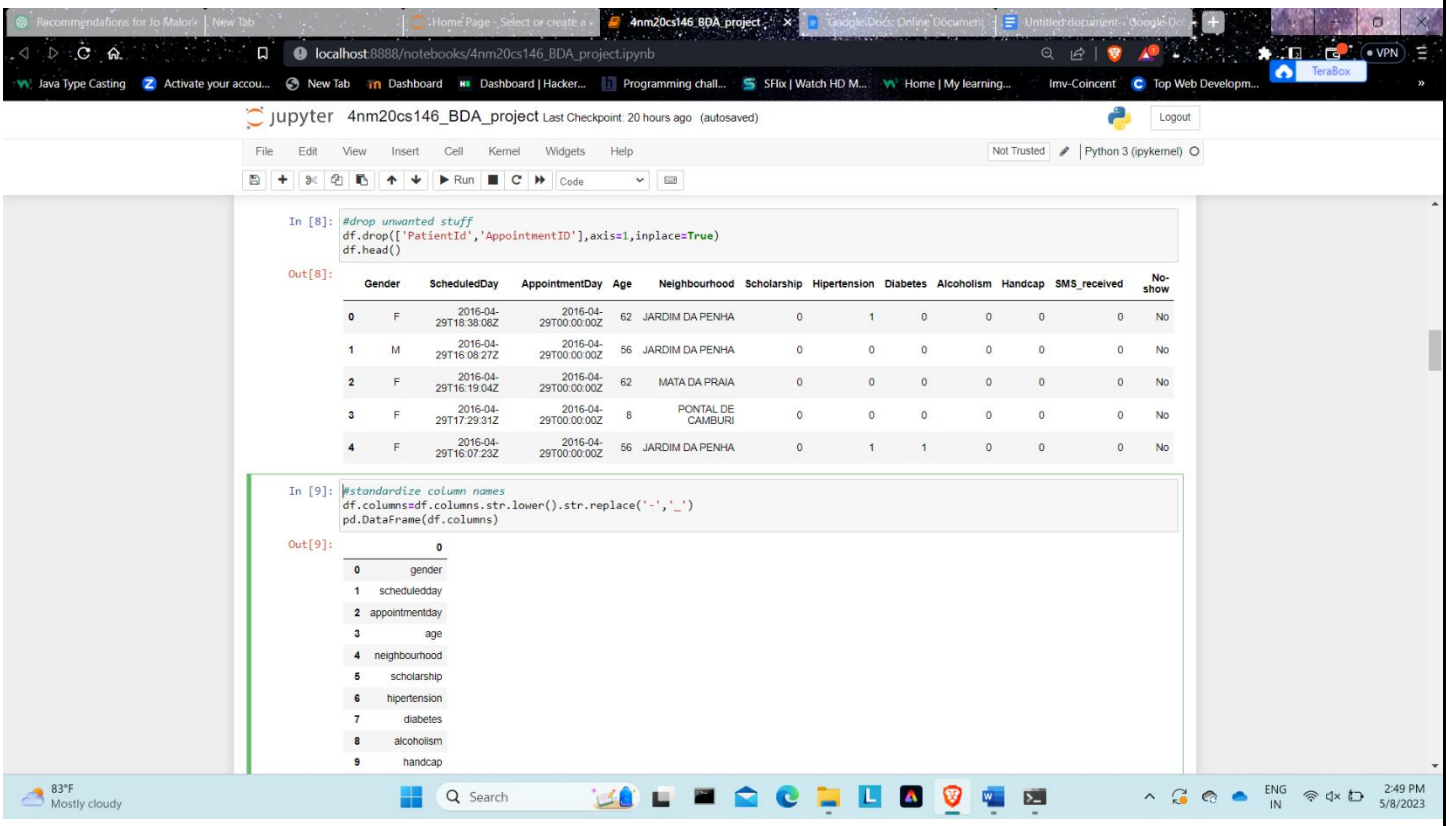
Out[3]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SM
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	

3. The number of missing values present in each column can be printed using the below code.



4. In the next step, the data can be cleaned so that the data can be converted into a way which which will be more suited to train Machine Learning models.



jupyter 4nm20cs146_BDA_project Last Checkpoint: 20 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [8]:
```
#drop unwanted stuff
df.drop(['PatientId','AppointmentID'],axis=1,inplace=True)
df.head()
```

Out[8]:

	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
1	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

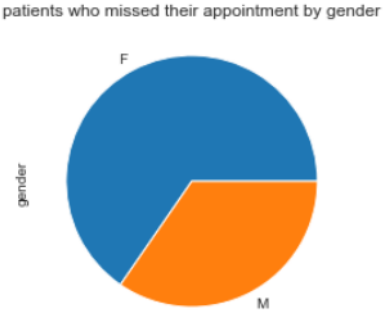
In [9]:
```
#standardize column names
df.columns=df.columns.str.lower().str.replace('-','_')
pd.DataFrame(df.columns)
```

Out[9]:

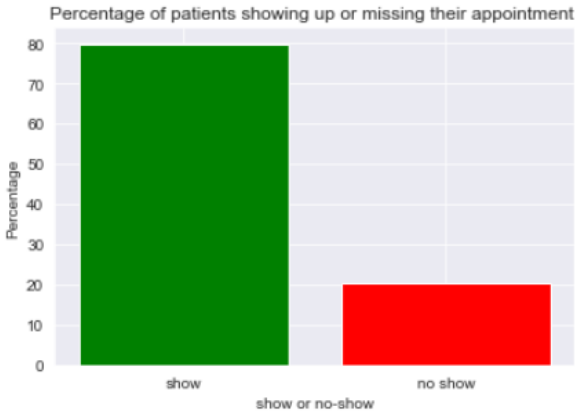
	0
0	gender
1	scheduledday
2	appointmentday
3	age
4	neighbourhood
5	scholarship
6	hipertension
7	diabetes
8	alcoholism
9	handcap

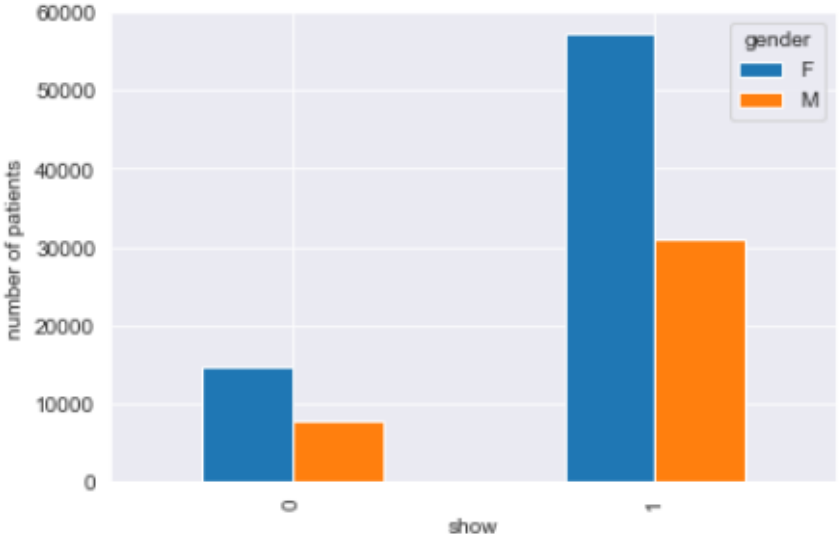
5. Below are some of the attributes which have been represented in the form of charts and graphs.

	gender
F	13.204013
M	6.989242

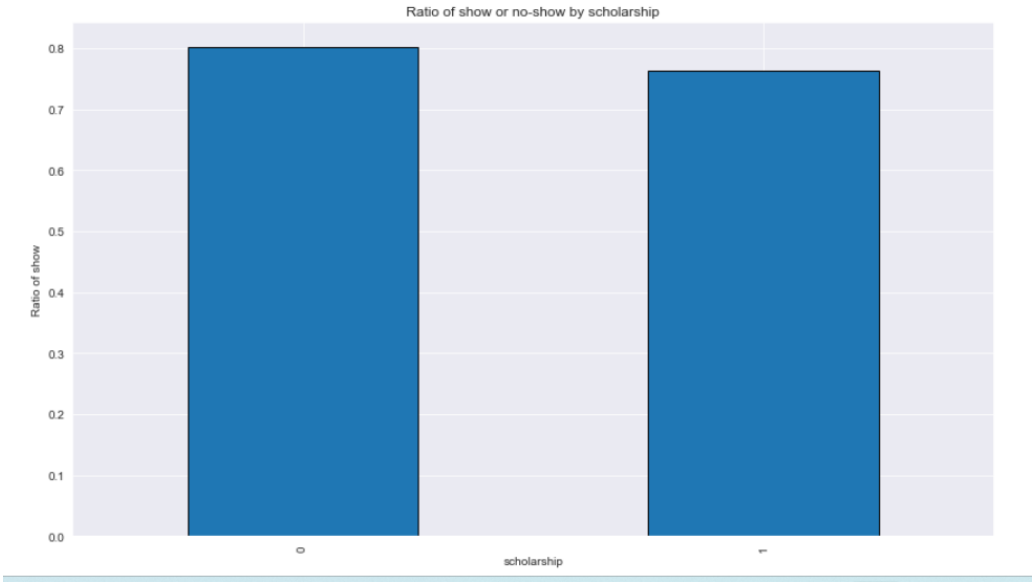


	show
show	
0	22319
1	88208

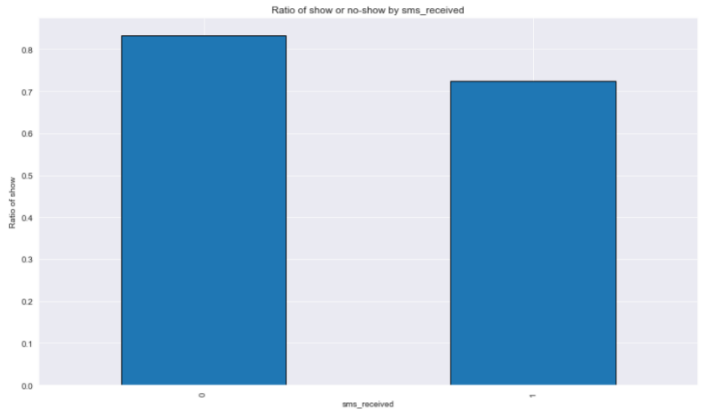




scholarship	
0	0.801928
1	0.762637

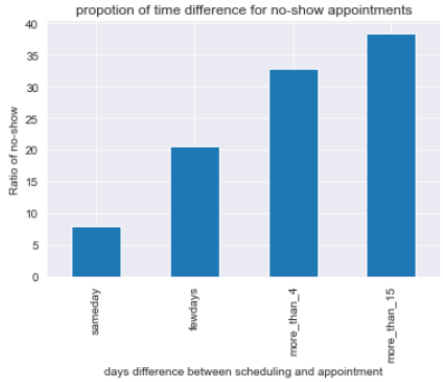


sms_received	show
0	0.832967
1	0.724255



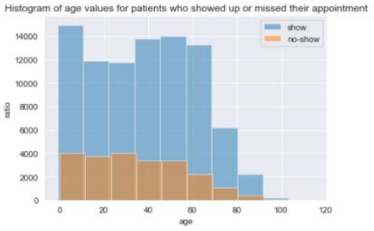
Out[29]:

	day_diff2
more_than_15	38.460505
more_than_4	32.922622
fewdays	20.565438
sameday	8.029034



Out[32]:

	age
count	22319.000000
mean	34.317667
std	21.965941
min	0.000000
25%	16.000000
50%	33.000000
75%	51.000000
max	115.000000



CONCLUSION AND FUTURE SCOPE

In recent years, medical data analysis has become an increasingly important field of research with the potential to revolutionize healthcare outcomes. The growth of big data and machine learning technologies has provided researchers and healthcare professionals with the tools they need to explore and analyze vast amounts of medical data. As a result, medical data analysis has the potential to transform healthcare by enabling more personalized treatment plans and improving clinical decision-making.

One of the key challenges associated with medical data analysis is data quality. Collecting and organizing medical data is a complex and challenging process, as medical data can come from a wide range of sources, including electronic health records, medical imaging, genomics, and wearable devices. Additionally, medical data often contains missing data or outliers, which can be difficult to handle. To address these challenges, data preprocessing techniques can be used to clean and transform medical data, enabling it to be used for analysis.

Once the data has been cleaned and preprocessed, researchers can use a range of descriptive statistics and data visualization techniques to explore patterns and trends in the data. This can provide valuable insights into disease prevalence, risk factors, and treatment outcomes. In addition, machine learning techniques such as decision trees, random forests, and neural networks can be applied to medical data to develop predictive models that can be used to make accurate and effective clinical decisions.

One of the key benefits of medical data analysis is the ability to develop personalized treatment plans for patients. By analyzing patient data, healthcare professionals can identify the most effective treatment options for individual patients, based on their medical history, genetics, and other factors. This can help to improve patient outcomes and reduce healthcare costs.

In addition to its potential benefits, medical data analysis also raises a number of ethical and privacy concerns. Collecting and analyzing medical data requires careful consideration of patient privacy and confidentiality. As such, it is important for researchers and healthcare professionals to develop ethical guidelines for medical data analysis and to ensure that patient data is handled responsibly.

In conclusion, medical data analysis has the potential to revolutionize healthcare outcomes by enabling personalized treatment plans and improving clinical decision-making. While there are many challenges associated with collecting and analyzing medical data, advancements in machine learning and predictive modeling techniques have opened up new avenues for research and innovation. As such, continued investment in medical data analysis is critical for improving patient outcomes and advancing the field of healthcare. However, it is important to carefully consider the ethical and privacy implications of medical data analysis and to develop responsible data handling practices.

REFERENCES

- Github: https://github.com/HadeerArafa/No-ShowAppointments