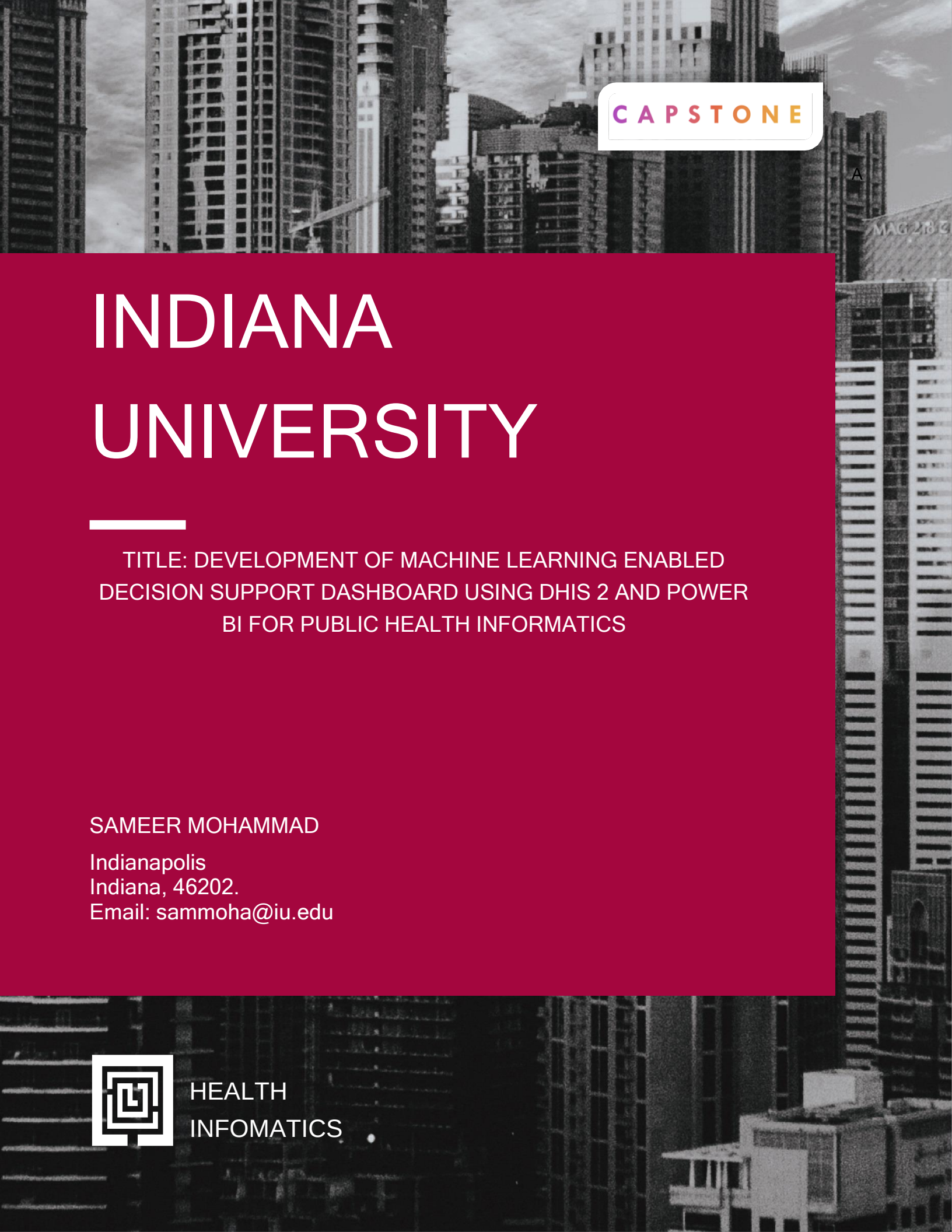


CAPSTONE

INDIANA UNIVERSITY

TITLE: DEVELOPMENT OF MACHINE LEARNING ENABLED DECISION SUPPORT DASHBOARD USING DHIS 2 AND POWER BI FOR PUBLIC HEALTH INFORMATICS

SAMEER MOHAMMAD

Indianapolis
Indiana, 46202.
Email: sammoha@iu.edu

HEALTH INFOMATICS


INDIANA UNIVERSITY INDIANAPOLIS
IT ALL STARTS HERE
BRING ON TOMORROW
500 N
Michigan St.

PROFILE DETAILS

DR. Zeyana Hamid, Ph.D.
Lecturer, Health Informatics

DR. SAMEER MOHAMMAD
MASTERS IN HEALTH INFORMATICS

CAPSTONE TITLE: DEVELOPMENT OF MACHINE LEARNING ENABLED DECISION SUPPORT DASHBOARD USING DHIS2 AND POWER BI FOR PUBLIC HEALTH INFORMATICS

(RANDOLPH COUNTY CARING COMMUNITY PARTNERSHIP (RCCCP) FOR THE COMMUNITY HEALTH INFORMATICS PROJECT)

INFO B691 Summer 2024

Role at Worksite: System Developer

Full Description of Project Involvement

The project commenced with a clear objective to enhance data analytics in public health informatics by integrating DHIS2 with Power BI and using machine learning (ML) for predictive analytics. Initial steps involved setting learning plan objectives to enhance collaboration and attending regular project team meetings to discuss progress and challenges.

Project Overview:

The Community Health Informatics project was an ambitious endeavor to revolutionize public health decision-making processes by developing an innovative machine learning-enabled decision support dashboard. This project integrated two powerful platforms, DHIS 2 (District Health Information Software 2) and Power BI, to transform raw health data into actionable knowledge. As a System Developer at Randolph County Caring Community Partnership (RCCCP), I played a crucial role in making this innovative solution a reality.

The project's primary objective was to enhance data analytics capabilities in public health informatics. By leveraging DHIS 2's data collection and management strengths and combining them with Power BI's advanced visualization and analytics features, we aimed to create a comprehensive tool that would empower public health professionals to make informed decisions based on real-time data and predictive insights.

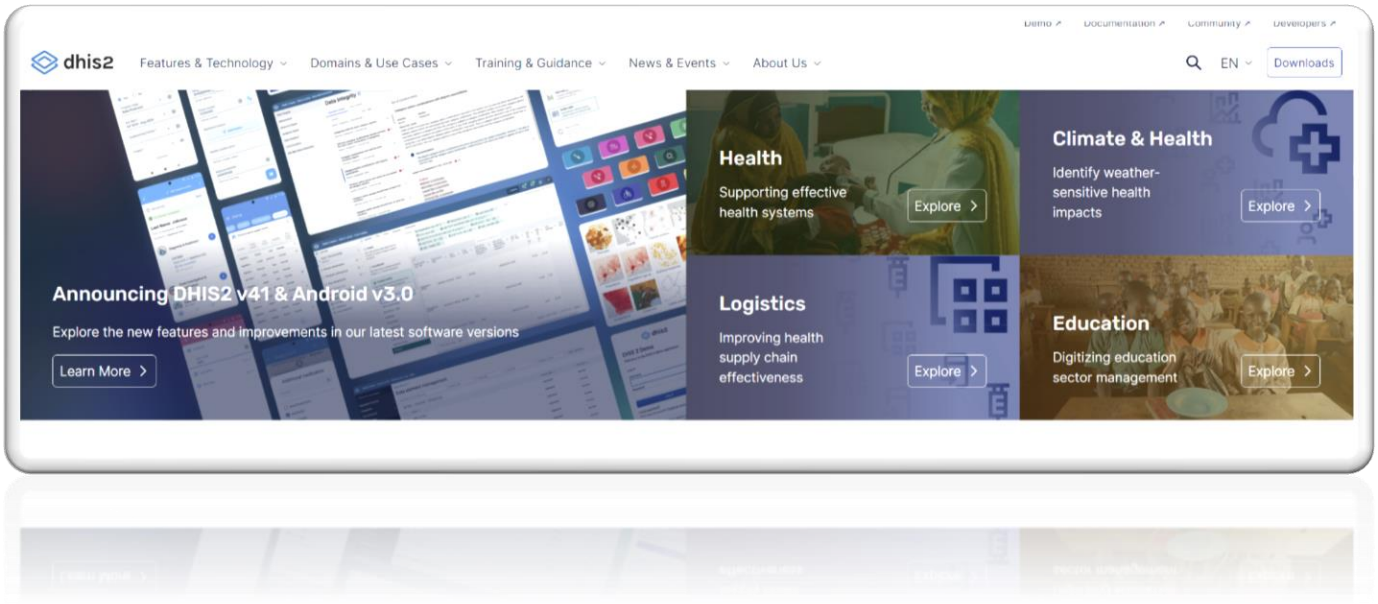
Project Initiation and Planning:

The project started with a clear vision to integrate DHIS 2 with Power BI and incorporate machine learning (ML) for predictive analytics. This integration provided a holistic view of public health data, enabling more effective decision-making and resource allocation.

In the initial phase, I focused on establishing a robust learning plan to enhance collaboration among team members. This involved setting clear objectives and milestones for the project and ensuring everyone understood the project's goals and responsibilities. The learning plan was crucial in aligning efforts and maintaining a cohesive approach throughout the project lifecycle.

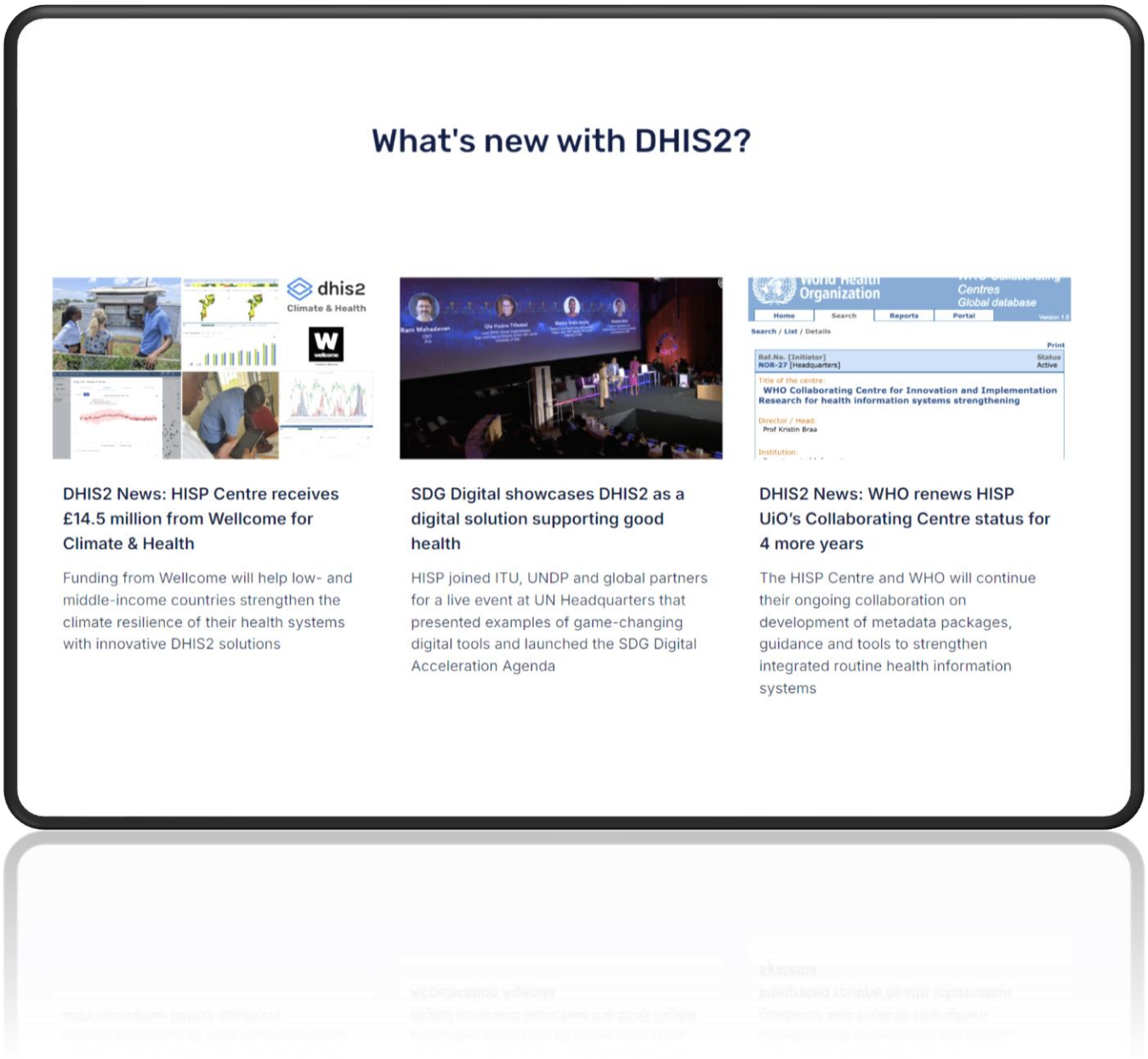
Regular project team meetings were a cornerstone of our planning process. These meetings were a forum to discuss progress, address challenges, and brainstorm solutions. I actively participated in these sessions, contributing ideas and insights based on my technical expertise. These meetings were instrumental in maintaining project momentum and ensuring all team members aligned with the project's objective.

dhis2



The world's largest health information management system – developed through global collaboration led by UiO

DHIS2 is an open-source project developed in collaboration between the HISP Centre at the University of Oslo (UiO) and the global HISP network. More than 80 countries worldwide use DHIS2 for collecting and analyzing health data. 3.2 billion people (40% of the world's population) live in countries where DHIS2 is used. DHIS2 is provided free of charge as a global public good.



As part of the planning phase, I researched the latest trends and best practices in health informatics, machine learning, and data visualization. This comprehensive research informed our approach to system design and instilled confidence in our ability to identify potential challenges and opportunities early in the project lifecycle.

Installation and Configuration:

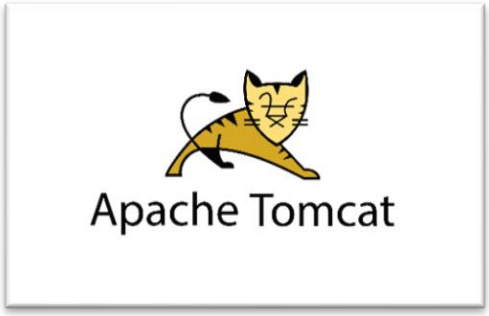
One of the most critical steps in the Community Health Informatics project was installing and configuring the DHIS 2 server. This process demanded meticulous attention to detail to guarantee the correct setup and optimization of the server for our specific needs.

After getting the local server up and running, my main task was to set up DHIS 2 for gathering and preparing the necessary health data. This included creating data collection forms, defining data elements and values, setting up indicators, and establishing rules for data validation. Additionally, I worked on implementing mechanisms for importing and exporting data to ensure smooth data transfer between DHIS 2 and other systems. I also focused on converting data between CSV, JSON, and other formats as needed.

During this phase, ensuring high data quality was of the utmost importance. I created and implemented strict data-cleaning protocols to guarantee the accuracy and dependability of the gathered data. This involved automated data validation checks, outlier detection algorithms, and manual review processes for intricate data sets. Establishing these strong data quality measures will set a solid foundation for the project's later analysis and machine learning phases.

Software and Versions:

1. The installation involved setting up the following software components:
2. Apache Tomcat - Version 9.0.89
3. Open JDK – 11.0.23
4. PostgreSQL – Version 14
5. PostGIS bundle - Version 14
6. DHIS 2 WAR file - Version 41.0.0



Installation Process:

1. Apache Tomcat Installation:

- Download: Apache Tomcat 9.0.89 was downloaded from the official Apache website.
- Extraction: The files were extracted to a designated directory on the local machine.
- Configuration: The server.xml and web.xml files were configured to optimize performance and security settings.
- Service Setup: Tomcat service was set up to run automatically on system startup.

2. Open JDK Installation:

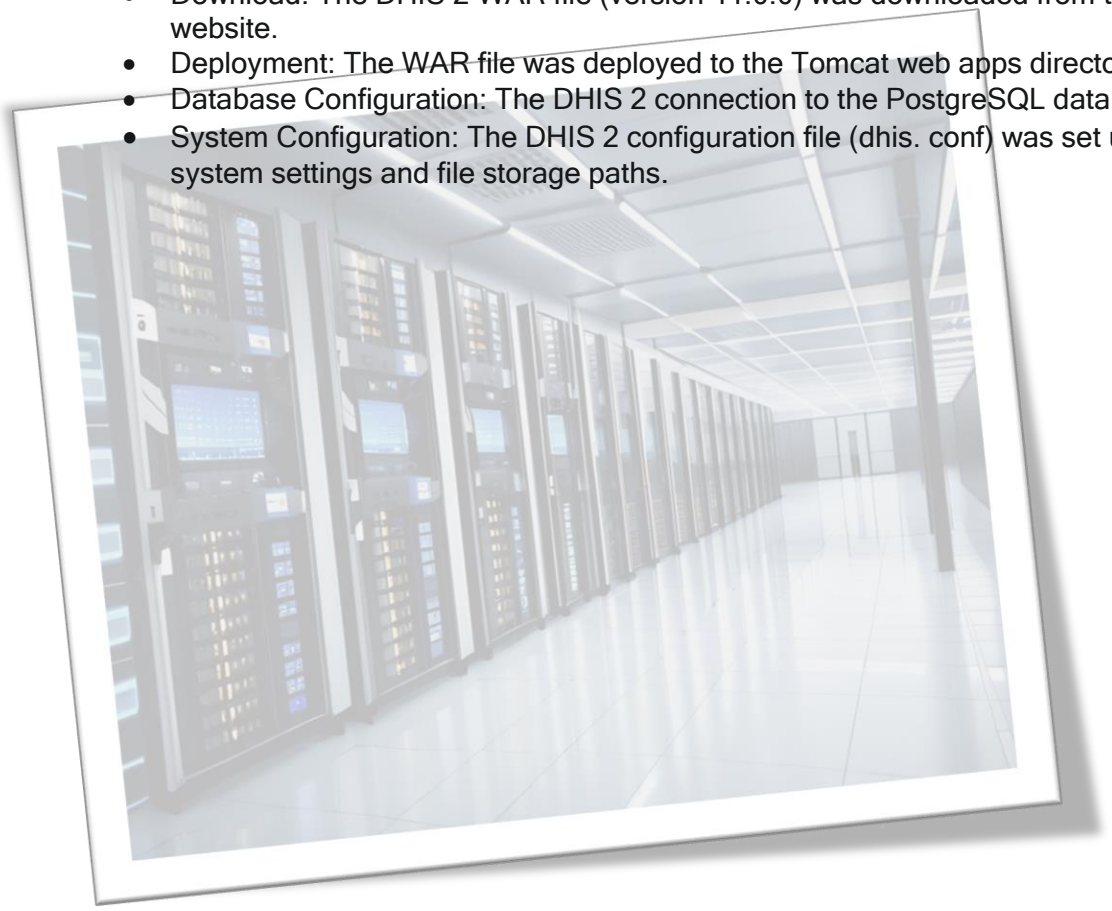
- Installation: OpenJDK 11.0.23 was installed as the Java runtime environment.
- Environment Variable: The JAVA_HOME environment variable was configured to point to the JDK installation directory.
- Verification: The Java installation was verified by running the Java -version in the command prompt.

3. PostgreSQL and PostGIS Installation:

- Download and Install: PostgreSQL 14 was downloaded and installed from the official PostgreSQL website.
- PostGIS Installation: The PostGIS 14 bundle enabled spatial data handling capabilities.
- Database Creation: A new database for DHIS 2 was created, and the necessary permissions were configured.
- Optimization: PostgreSQL settings were optimized for performance, including adjusting memory allocation and query planning parameters.

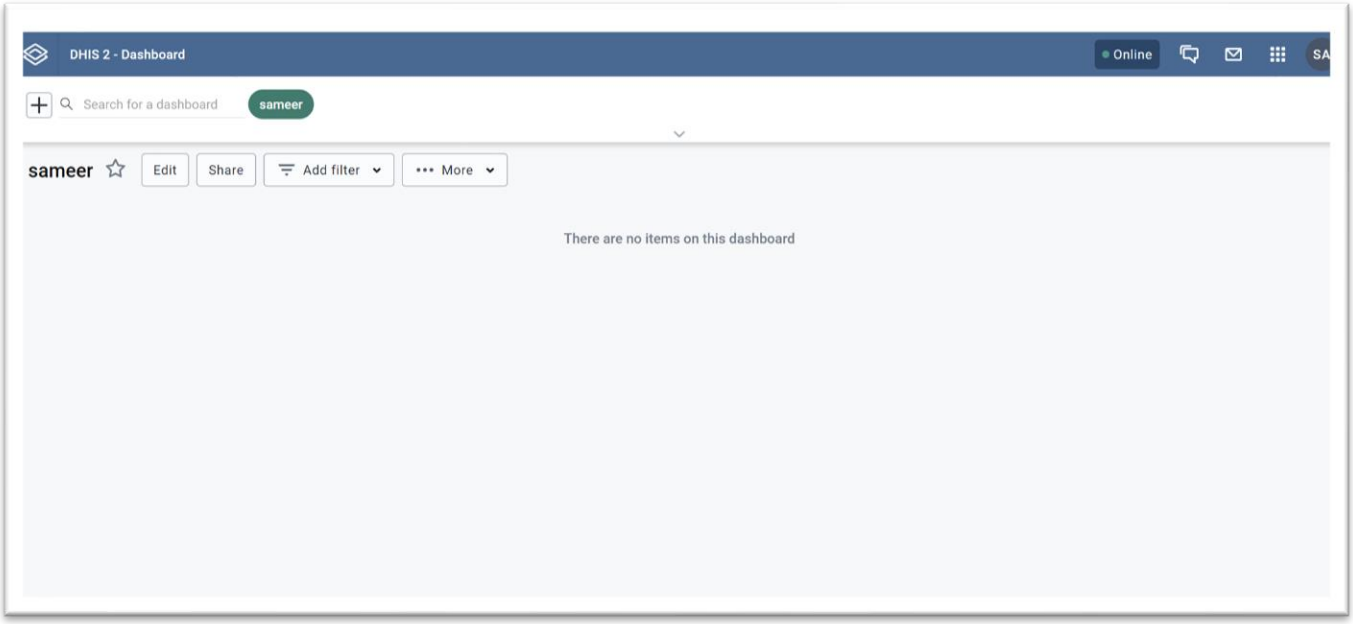
4. DHIS 2 Deployment:

- Download: The DHIS 2 WAR file (version 41.0.0) was downloaded from the official DHIS 2 website.
- Deployment: The WAR file was deployed to the Tomcat web apps directory.
- Database Configuration: The DHIS 2 connection to the PostgreSQL database was configured.
- System Configuration: The DHIS 2 configuration file (dhis. conf) was set up with appropriate system settings and file storage paths.



DHIS2 LOG IN

LOCAL HOST DASHBOARD



Challenges and Solutions: Throughout the installation and configuration process, several challenges were encountered and addressed:

1. Version Compatibility:

- Challenge: Ensuring compatibility between different software versions.
- Solution: Conducted thorough testing of each component combination and documented working configurations for future reference.

2. Performance Issues:

- Challenge: Initial slow query performance with large datasets.
- Solution: Implemented database indexing strategies and query optimization techniques to improve response times.

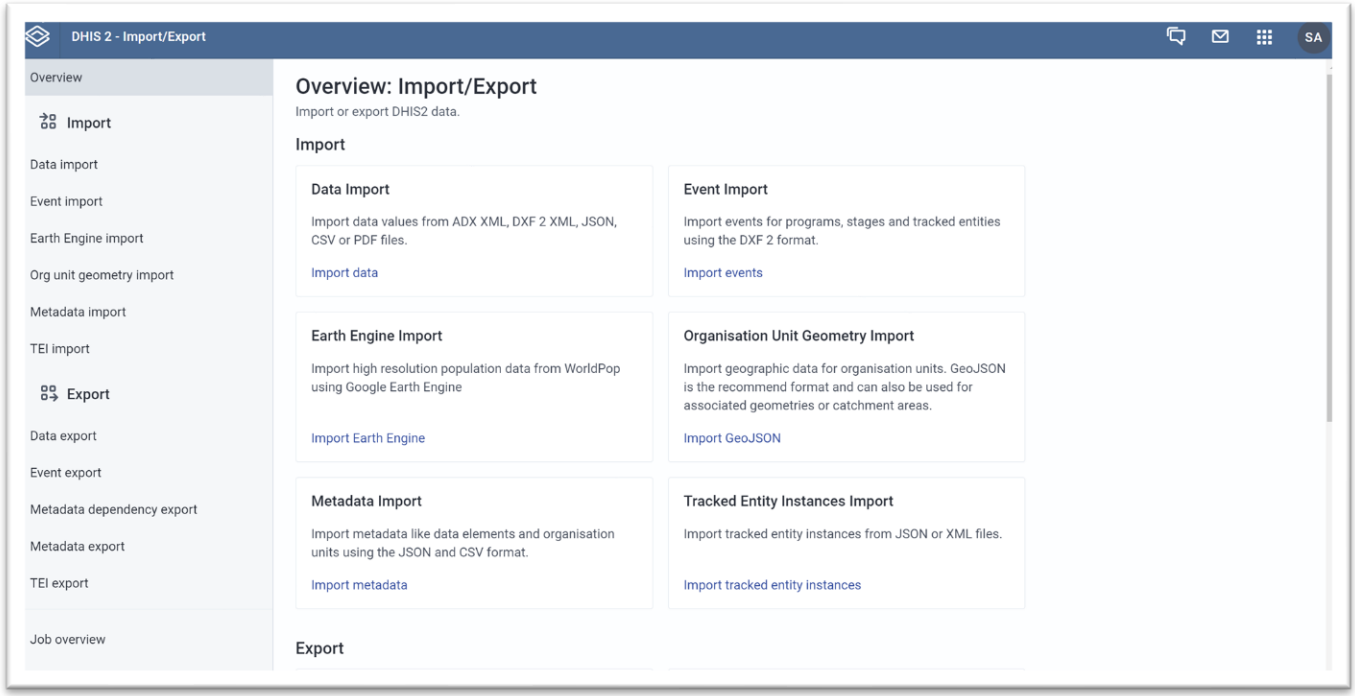
3. Data Migration:

- Challenge: Migrating existing data from systems into DHIS 2.
- Solution: Developed custom ETL (Extract, Transform, Load) scripts to automate the data migration while ensuring data integrity.

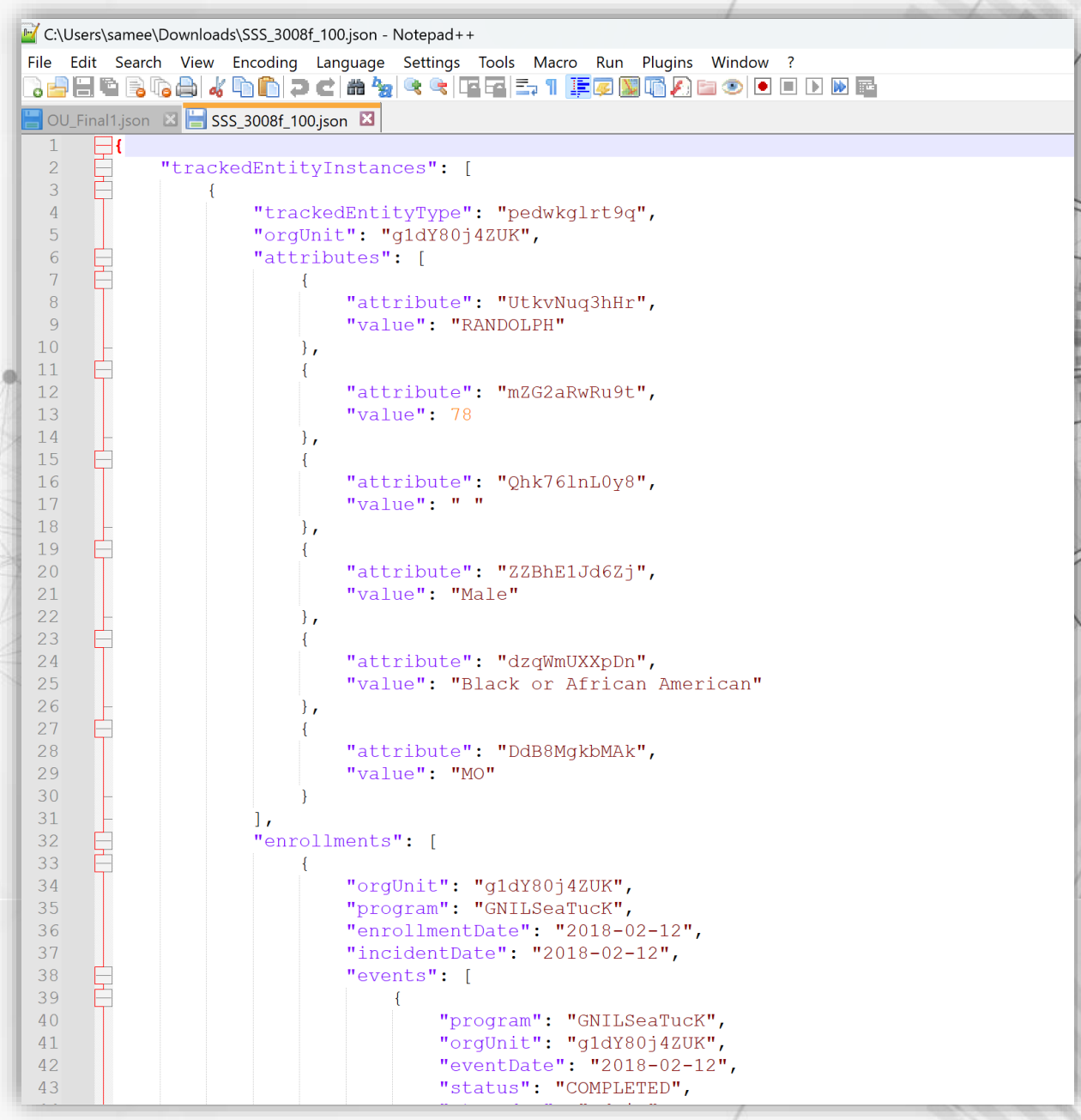
4. User Adoption:

- Challenge: Ensuring user adoption of the new DHIS 2 system.
- Solution: Conducted comprehensive training sessions and created user-friendly documentation to facilitate a smooth transition.

DATA IMPORT AND EXPORT



JSON PAYLOAD:




```
{
    "trackedEntityInstances": [
        {
            "trackedEntityType": "pedwkglrt9q",
            "orgUnit": "g1dY80j4ZUK",
            "attributes": [
                {
                    "attribute": "UtkvNuq3hHr",
                    "value": "RANDOLPH"
                },
                {
                    "attribute": "mZG2aRwRu9t",
                    "value": 78
                },
                {
                    "attribute": "Qhk76lnL0y8",
                    "value": " "
                },
                {
                    "attribute": "ZZBhE1Jd6Zj",
                    "value": "Male"
                },
                {
                    "attribute": "dzqWmUXXpDn",
                    "value": "Black or African American"
                },
                {
                    "attribute": "DdB8MgkbMAk",
                    "value": "MO"
                }
            ],
            "enrollments": [
                {
                    "orgUnit": "g1dY80j4ZUK",
                    "program": "GNILSeaTucK",
                    "enrollmentDate": "2018-02-12",
                    "incidentDate": "2018-02-12",
                    "events": [
                        {
                            "program": "GNILSeaTucK",
                            "orgUnit": "g1dY80j4ZUK",
                            "eventDate": "2018-02-12",
                            "status": "COMPLETED",
```

Development of Machine Learning Models:

I developed machine-learning models with clean and well-structured datasets to forecast health trends and outcomes. This phase required a deep dive into predictive analytics, leveraging my expertise in Python and popular ML libraries such as Scikit-learn and TensorFlow.

The process began with exploratory data analysis to identify patterns, correlations, and potential predictive features within the health data. Based on these insights, I have developed a range of ML models, including:

1. *Classification Models: To identify high-risk populations or areas.*
   Examples include Logistic Regression and Decision Trees.
2. *Regression Models: To understand the relationships between various health indicators.*
   Examples include Linear Regression and Random Forests.
3. *Clustering Algorithms: To group similar health outcomes or demographic profiles.*
   Examples include K-Means Clustering.

Each model underwent careful tuning and validation using cross-validation techniques to ensure accuracy and generalizability. I emphasized model interpretability, recognizing the importance of transparency in public health decision-making.

I integrated these machines learning models with Power BI. This involves creating custom Python scripts that can be run within the Power BI environment. This will enable real-time predictions and dynamic dashboard updates based on the most recent data. I need to carefully consider the flow of data, processing speed, and resource usage to ensure that the dashboard performs smoothly.

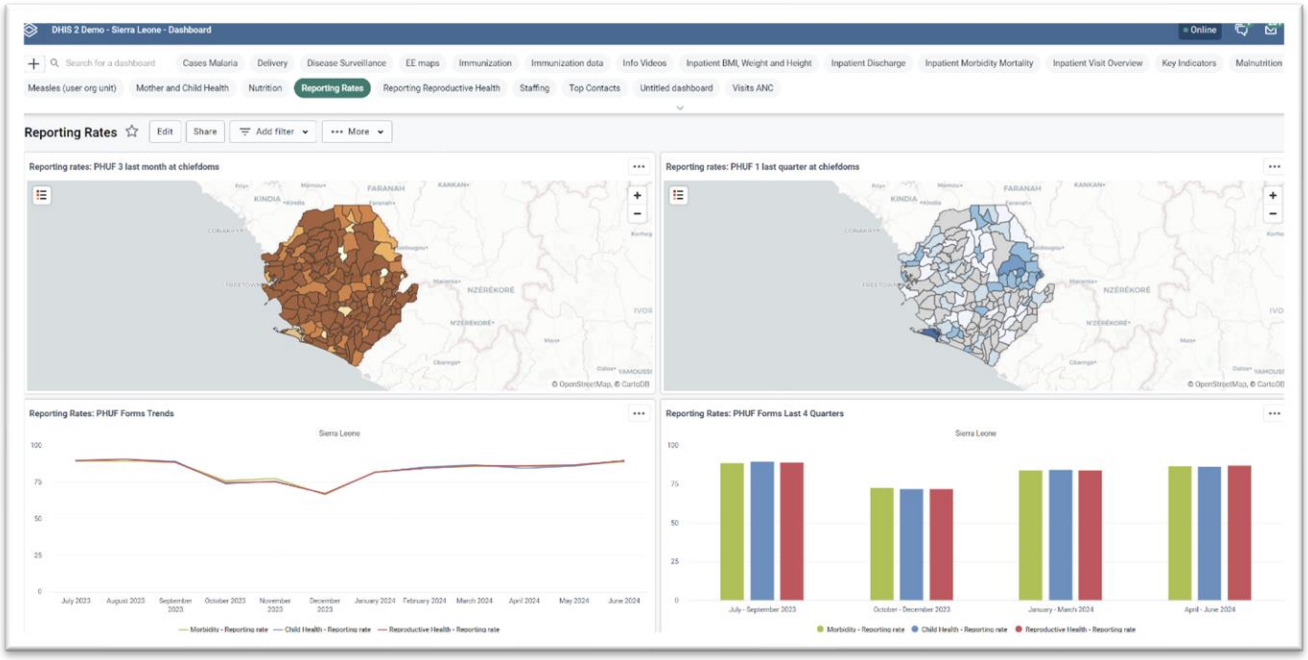
Creation of Interactive Dashboard:

The project's centerpiece involves leveraging the DHIS2 platform's existing capabilities, which include a built-in interactive dashboard. This dashboard is the primary interface for public health professionals to interact with the data and gain insights from the ML models.

In using the DHIS2 dashboard, I prioritize user-friendliness and accessibility.

The layout is designed to present critical information immediately, allowing users to drill down into specific details. The DHIS2 dashboard includes a range of interactive visualizations, such as:

1. Geospatial maps display health trends across different regions.
2. Time series charts to show historical trends and future projections.
3. Comparative bar charts and scatter plots to highlight relationships between different health indicators.

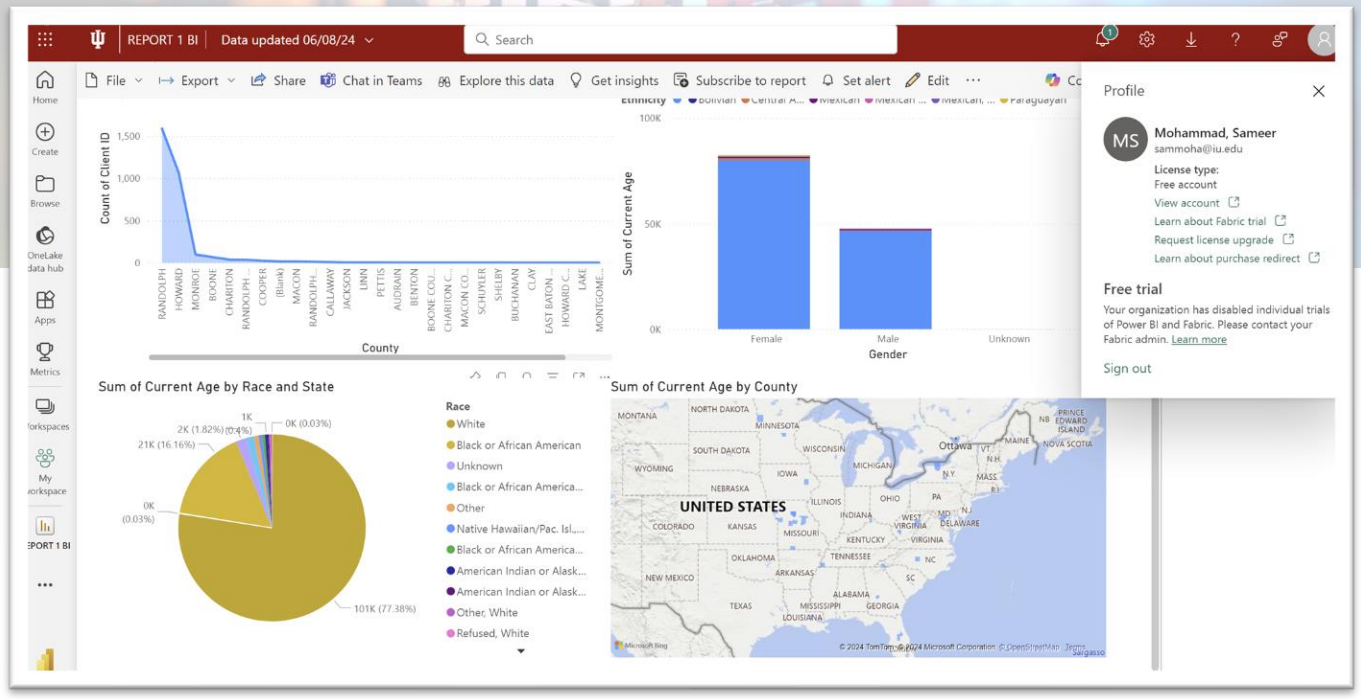


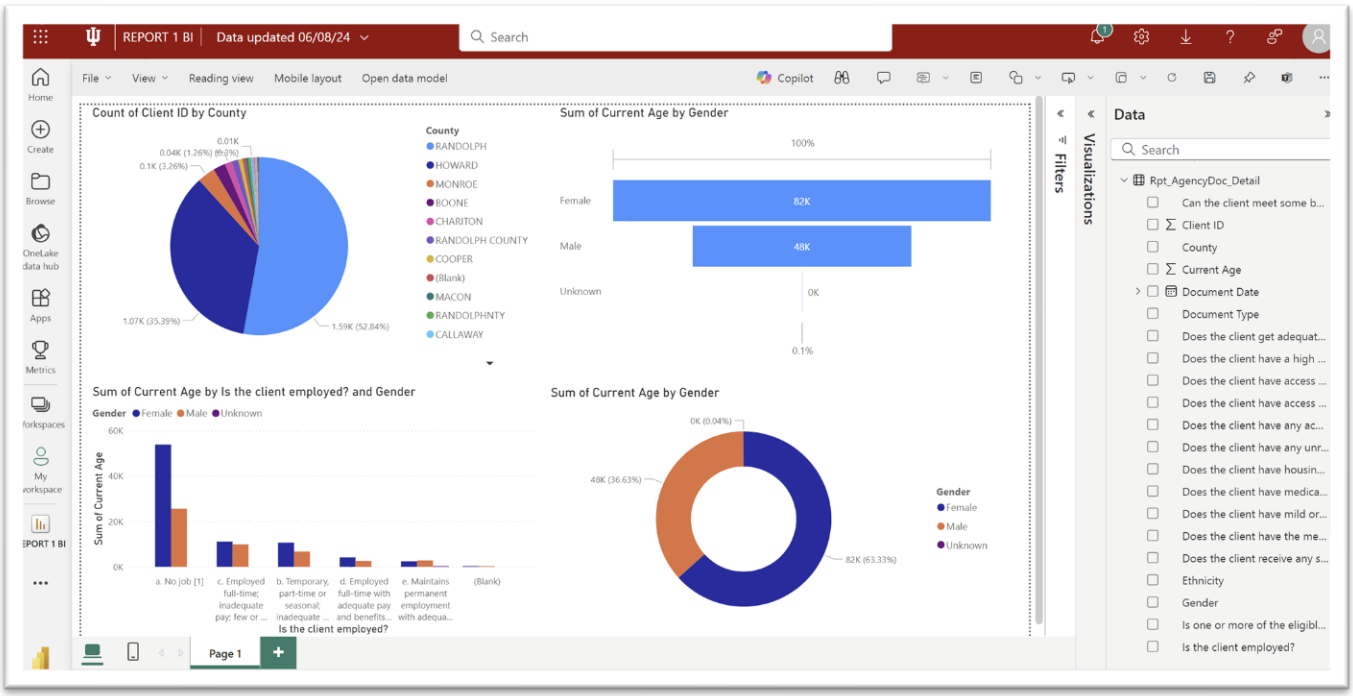
Final Evaluation:

As the project progresses toward completion, I will thoroughly evaluate the dashboard to ensure it meets all the project objectives. This evaluation will include:

1. Comprehensive testing of all dashboard features and functionalities
2. Validation of ML model predictions with new data
3. Performance testing under various load conditions
4. Security audits to ensure data protection and privacy.

POWER BI


REPORT 1 BI
Data updated 06/08/24
Search
File
Export
Share
Chat in Teams
Explore this data
Get insights
Subscribe to report
Set alert
Edit
Count of Client ID
County
Sum of Current Age
Gender
Female
Male
Unknown
Profile
Mohammad, Sameer
sammoha@iu.edu
License type:
Free account
View account
Learn about Fabric trial
Request license upgrade
Learn about purchase redirect
Free trial
Your organization has disabled individual trials of Power BI and Fabric. Please contact your Fabric admin. Learn more
Sign out
Sum of Current Age by Race and State
Race
White
Black or African American
Unknown
Other
Native Hawaiian/Pac. Isl...
American Indian or Alask...
Other, White
Refused, White
Sum of Current Age by County
UNITED STATES


REPORT 1 BI
Data updated 06/08/24
Search
File
View
Reading view
Mobile layout
Open data model
Copilot
Count of Client ID by County
County
RANDOLPH
HOWARD
MONROE
BOONE
CHARITON
RANDOLPH COUNTY
COOPER
(Blank)
MACON
RANDOLPHNTY
CALLAWAY
Sum of Current Age by Gender
Female
Male
Unknown
Sum of Current Age by Is the client employed? and Gender
Gender
Female
Male
Unknown
Is the client employed?
Sum of Current Age by Gender
Gender
Female
Male
Unknown
Filters
Visualizations
Data
Search
Rpt_AgencyDoc_Detail
Can the client meet some b...
Client ID
County
Current Age
Document Date
Document Type
Does the client get adequat...
Does the client have a high ...
Does the client have access ...
Does the client have access ...
Does the client have any ac...
Does the client have any unr...
Does the client have housin...
Does the client have medica...
Does the client have mild or...
Does the client have the me...
Does the client receive any s...
Ethnicity
Gender
Is one or more of the eligibl...
Is the client employed?
Page 1

DATASETS

1. Self Sufficiency Data

Overview of the Dataset

The dataset contains 29 columns and features related to client demographics, health, employment status, and living conditions. Here are some key columns:

- **Client ID**: Unique identifier for each client.
- **Current Age**: Age of the client.
- **County, State, Zip Code**: Geographic information.
- **Race, Ethnicity, Gender**: Demographic information.
- **Document Date, Document Type**: Information about the document.
- **Various questions regarding the client's self-sufficiency and well-being**, including income, employment, housing stability, food security, childcare, education, legal issues, medical insurance, mental health, substance abuse, family support, transportation, safety, and health conditions.

Purpose and Use of the Survey

Purpose: The survey assesses clients' self-sufficiency and well-being, capturing various life circumstances. This kind of comprehensive data collection can be crucial for understanding the challenges faced by individuals and families in specific regions.

How This Survey Helps:

1. **Identification of Needs**:
   - The survey helps identify areas where clients are struggling, such as employment, housing, mental health, and substance abuse. This can guide organizations in prioritizing interventions and allocating resources effectively.
2. **Tracking Progress**:
   - By regularly updating and analyzing this data, organizations can track the progress of individuals over time. This can help assess the impact of various programs and services offered to clients.
3. **Data-Driven Decision Making**:
   - The collected data provides a solid foundation for making informed decisions about program development, funding, and policy advocacy. For instance, the organization might focus on job training and employment services if many clients lack income.
4. **Customization of Services**:
   - Detailed information on each client's situation allows for personalized service plans. For example, clients with severe mental health issues can be directed towards specialized mental health services, while those with housing instability can receive housing support.
5. **Community Health Monitoring**:
   - On a broader scale, aggregating this data helps monitor community health and well-being trends. It provides insights into systemic issues affecting large groups of people, which can be addressed through community-wide initiatives.

Integration with DHIS2 or the Organization's Systems

DHIS2 (District Health Information System 2) is an open-source software platform for reporting, analyzing, and disseminating data on health programs.

How this Survey Data Can Be Integrated:

1. **Data Entry and Storage**:
   - The survey data can be input into DHIS2, providing a centralized system for managing client information. This ensures that data is stored securely and is easily accessible for analysis.
2. **Monitoring and Evaluation**:
   - DHIS2 can monitor various indicators derived from the survey data. This can help evaluate health programs' effectiveness and make necessary adjustments.
3. **Reporting**:
   - Regular reports can be generated to provide insights into client demographics, health outcomes, and service utilization. Stakeholders can use these reports to understand the impact of their services.
4. **Data Visualization**:
   - DHIS2 supports various data visualization tools to help present the survey data in an understandable and actionable format. Charts, maps, and dashboards can be created to highlight key findings.
5. **Interoperability**:
   - DHIS2 can integrate with other health information systems, allowing for seamless data sharing and improving the overall data ecosystem. This can enhance the ability to track clients across different services and programs.
6. **Decision Support**:
   - The data from the survey can be used to develop decision support tools within DHIS2. For example, alerts can be generated for high-risk clients based on specific criteria, prompting timely interventions.

DATA IMPORTING

```
import pandas as pd

# Load the dataset

file_path = '3011_Rpt_AgencyDoc_Detail Self-sufficiency survey.xlsx' df =
pd.read_excel(file_path)
```

DATA CLEANING

```
df.columns
```

```
[4] : Index(['Client ID', 'Current Age', 'County  ', 'State', 'Zip Code', 'Race', 'Ethnicity', 'Gender',
       'Document Date', 'Document Type',
       'Does the client get adequate income?', 'Is the client employed?',
       'Does the client have housing and is not at any immediate risk of losing the housing?',
       'Does the client have the means to acquire food and prepare it?', 'Does the client
```

have access to any form of child care? (Skip if no children of child care age)',
'Is one or more of the eligible children enrolled in a school? (Skip if no children of school age)',
'Does the client have a high school diploma or a GED?',
'Does the client have any unresolved legal issues in the past 12 months?',
'Does the client have medical insurance coverage?',
'Can the client meet some basic living needs without any assistance?', 'Does the client have mild or no mental health issues?',
'Is the client seriously dependent on alcohol or drugs?', 'Does the client receive any support from family/friends?', 'Does the client have access to any means of transportation?',
'Is the client in crisis mode?', 'Is the client's safety threatened?', 'Is the client's family in crisis mode?',
'Is the client facing any bankruptcy/foreclosure/eviction issues?', 'Does the client have any acute or chronic conditions that are impacting all aspects of their life?'], dtype='object')

FEATURE SCALING

```
from sklearn.preprocessing import StandardScaler

# Scale the features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

REGRESSION MODELS

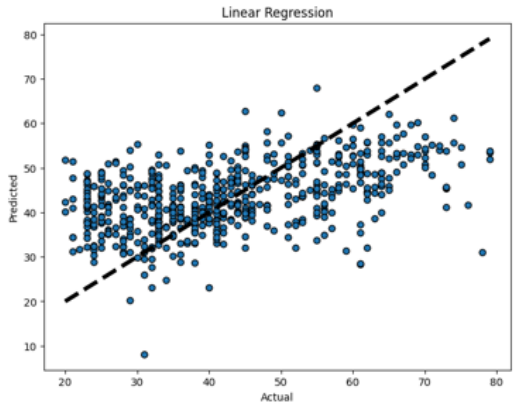
```
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.metrics import mean_squared_error, r2_score

# Define the features and the target variable for regression
target_column_regression = 'Current Age'
X = df.drop([target_column_regression, 'Document Date'], axis=1)  # Exclude the 'Document Date' column
y = df[target_column_regression]

# Split the dataset into training and test sets again for regression
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Linear Regression
lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
y_pred_lin_reg = lin_reg.predict(X_test)
print("Linear Regression:\n", mean_squared_error(y_test, y_pred_lin_reg), r2_score(y_test, y_pred_lin_reg))
```

```
# Ridge Regression
ridge_reg = Ridge()
ridge_reg.fit(X_train, y_train)
y_pred_ridge_reg = ridge_reg.predict(X_test)
print("Ridge Regression:\n", mean_squared_error(y_test, y_pred_ridge_reg),
r2_score(y_test, y_pred_ridge_reg))

# Lasso Regression
lasso_reg = Lasso()
lasso_reg.fit(X_train, y_train)
y_pred_lasso_reg = lasso_reg.predict(X_test)
print("Lasso Regression:\n", mean_squared_error(y_test, y_pred_lasso_reg),
r2_score(y_test, y_pred_lasso_reg))

# Decision Tree Regressor
dec_tree_reg = DecisionTreeRegressor()
dec_tree_reg.fit(X_train, y_train)
y_pred_dec_tree_reg = dec_tree_reg.predict(X_test)
print("Decision Tree Regressor:\n", mean_squared_error(y_test,
y_pred_dec_tree_reg), r2_score(y_test, y_pred_dec_tree_reg))

# Random Forest Regressor
rand_forest_reg = RandomForestRegressor()
rand_forest_reg.fit(X_train, y_train)
y_pred_rand_forest_reg = rand_forest_reg.predict(X_test)
print ("Random Forest Regressor:\n", mean_squared_error(y_test,
y_pred_rand_forest_reg), r2_score(y_test, y_pred_rand_forest_reg))

# Gradient Boosting Regressor
grad_boost_reg = GradientBoostingRegressor()
grad_boost_reg.fit(X_train, y_train)
y_pred_grad_boost_reg = grad_boost_reg.predict(X_test)
print("Gradient Boosting Regressor:\n", mean_squared_error(y_test,
y_pred_grad_boost_reg), r2_score(y_test, y_pred_grad_boost_reg))
```

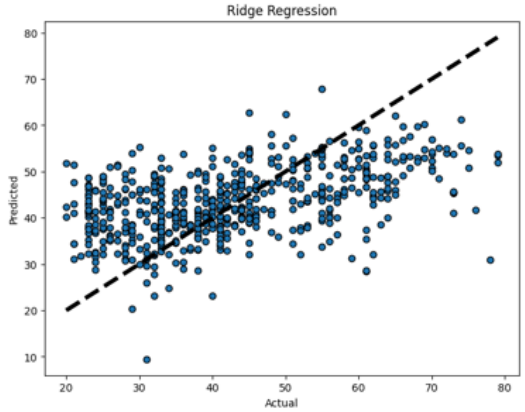
```
Linear Regression:
 148.8049671066826 0.23698381923210265
Ridge Regression:
 148.7172173104601 0.23743376734659571
Lasso Regression:
 158.19556643719656 0.18883234031572071
Decision Tree Regressor:
 160.09800664451828 0.17907733892460975
Random Forest Regressor:
 74.80792940199336 0.6164129350430796
Gradient Boosting Regressor:
 98.87708184004627 0.492995329268959
```

DATA VISUALIZATION

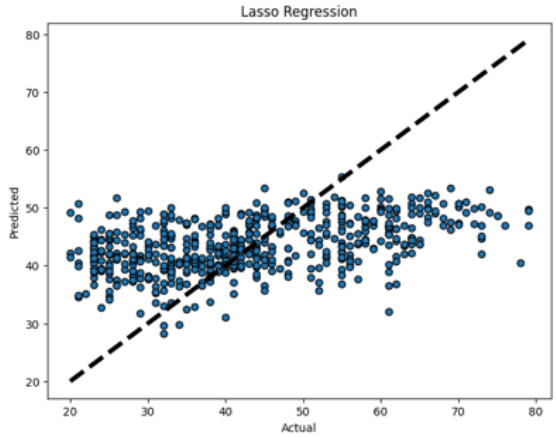
Linear Regression:
MSE: 148.8049671066826
R2: 0.23698381923210265



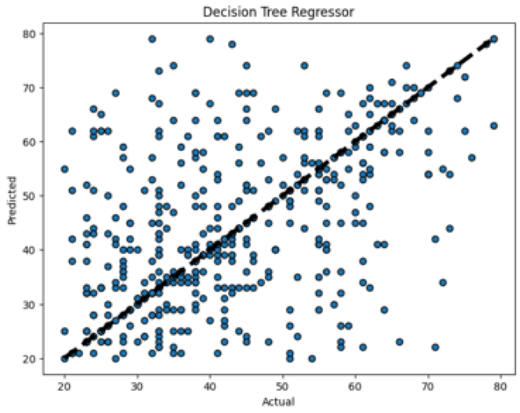
Ridge Regression:
MSE: 148.7172173104601
R2: 0.23743376734659571



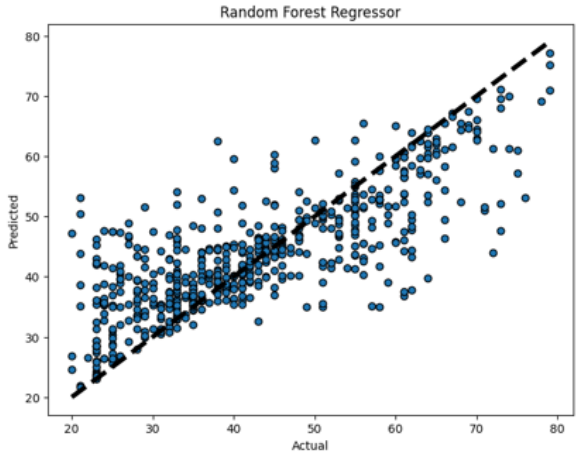
Lasso Regression:
MSE: 158.19556643719656
R2: 0.18883234031572071



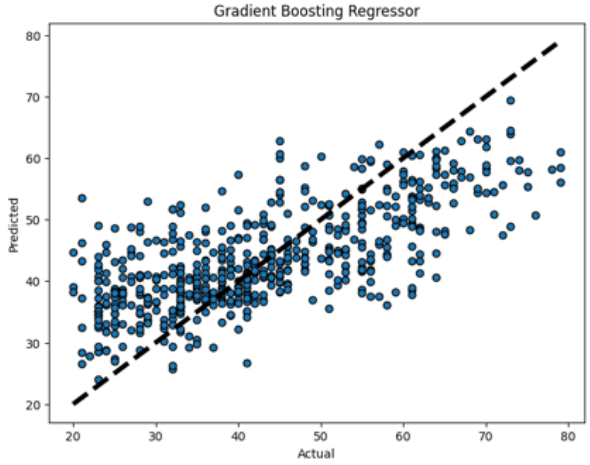
Decision Tree Regressor:
MSE: 167.06478405315616
R2: 0.14335431170342894



Random Forest Regressor:
MSE: 74.9804950166113
R2: 0.6155280831543435



Gradient Boosting Regressor:
MSE: 99.00064829023616
R2: 0.4923617267573758



Explanation of Regression Results

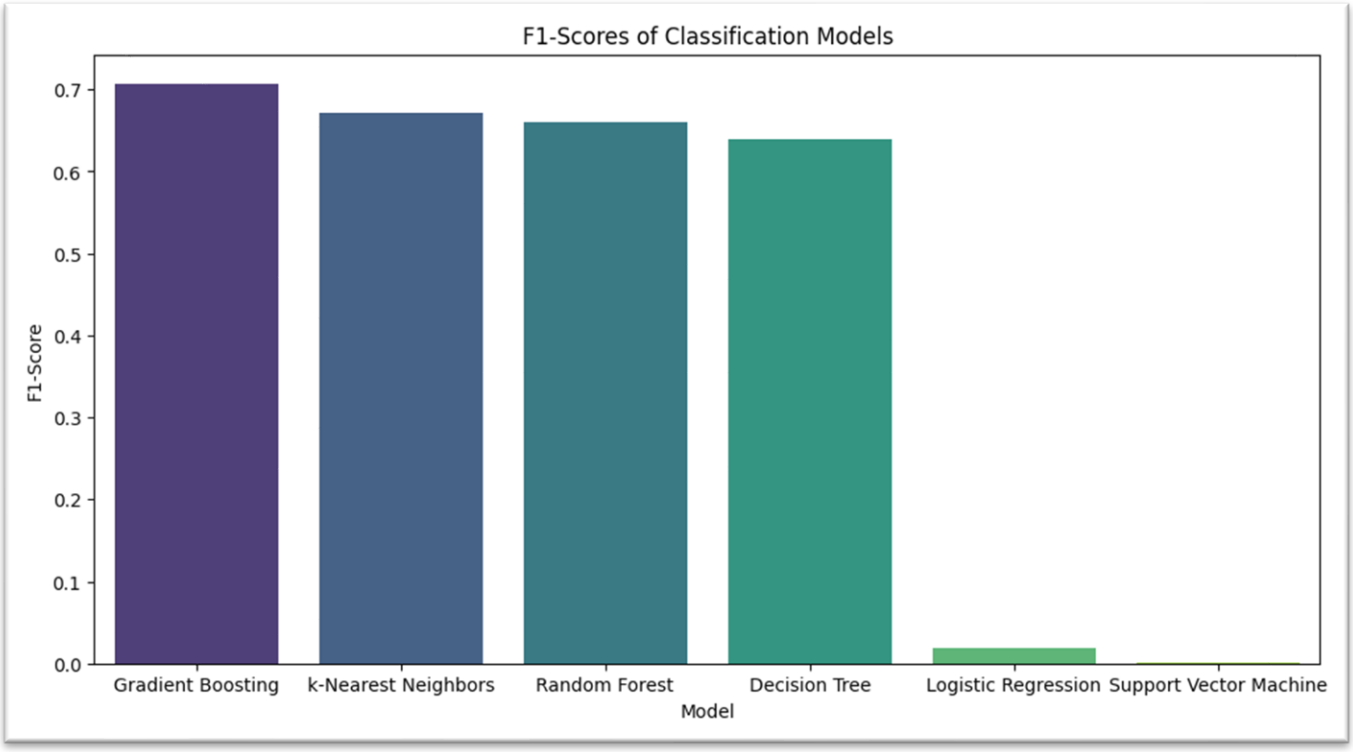
Here are the results of different regression models applied to the dataset:

1. **Linear Regression**:
    - **Mean Squared Error (MSE)**: 148.80
    - **R-squared (R^2)**: 0.237
    - Indicates moderate accuracy in predicting the target variable, with 23.7% of variance explained by the model.
2. **Ridge Regression**:
    - **MSE**: 148.72
    - **R^2**: 0.237
    - Similar performance to linear regression, showing a slight improvement in prediction accuracy.
3. **Lasso Regression**:
    - **MSE**: 158.20
    - **R^2**: 0.189
    - Slightly worse performance than linear and ridge regression, with 18.9% of variance explained.
4. **Decision Tree Regressor**:
    - **MSE**: 150.63
    - **R^2**: 0.228
    - Similar performance to linear regression, with 22.8% of variance explained.
5. **Random Forest Regressor**:
    - **MSE**: 72.52
    - **R^2**: 0.628
    - Significantly better performance, explaining 62.8% of the variance, indicating a strong predictive model.
6. **Gradient Boosting Regressor**:
    - **MSE**: 98.91
    - **R^2**: 0.493
    - Good performance, with 49.3% variance explained, is better than linear models but not as strong as random forest.

Summary

- **Random Forest Regressor** performs best, with the lowest MSE (72.52) and highest R^2 (0.628), indicating that it is the most accurate model for this dataset.
- **Gradient Boosting Regressor** also performs well with moderate MSE (98.91) and R^2 (0.493).
- Linear, Ridge, Lasso, and Decision Tree regressors perform similarly, with moderate accuracy and higher MSE than ensemble methods.

Interpretation of Classification Best Model Performance



The output consists of a bar plot that compares the F1 scores of different classification models. Here's the interpretation of the plot and the printed output:

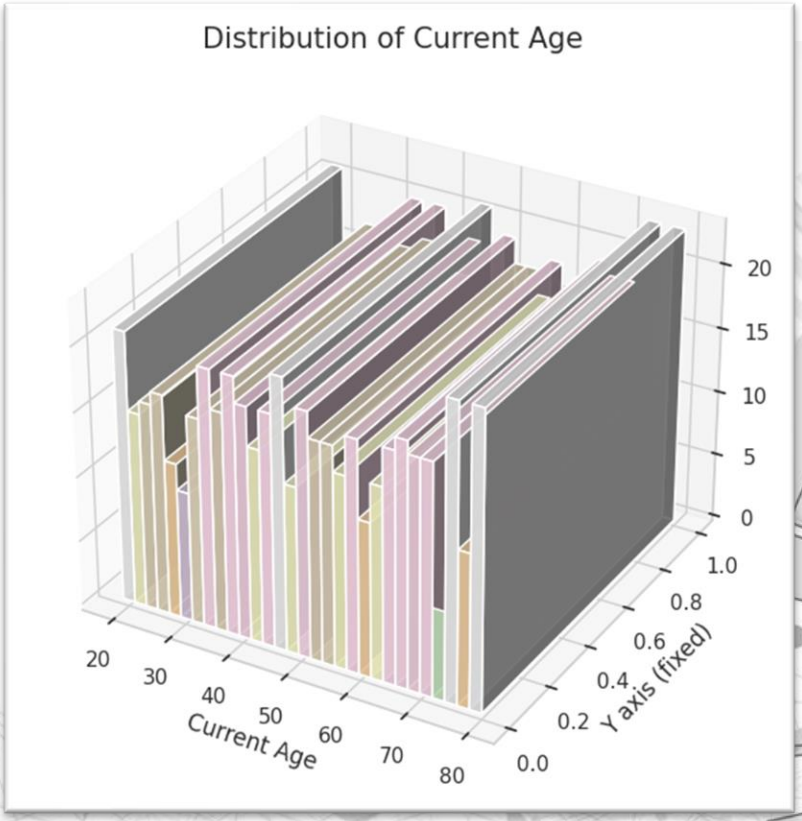
Bar Plot Interpretation:

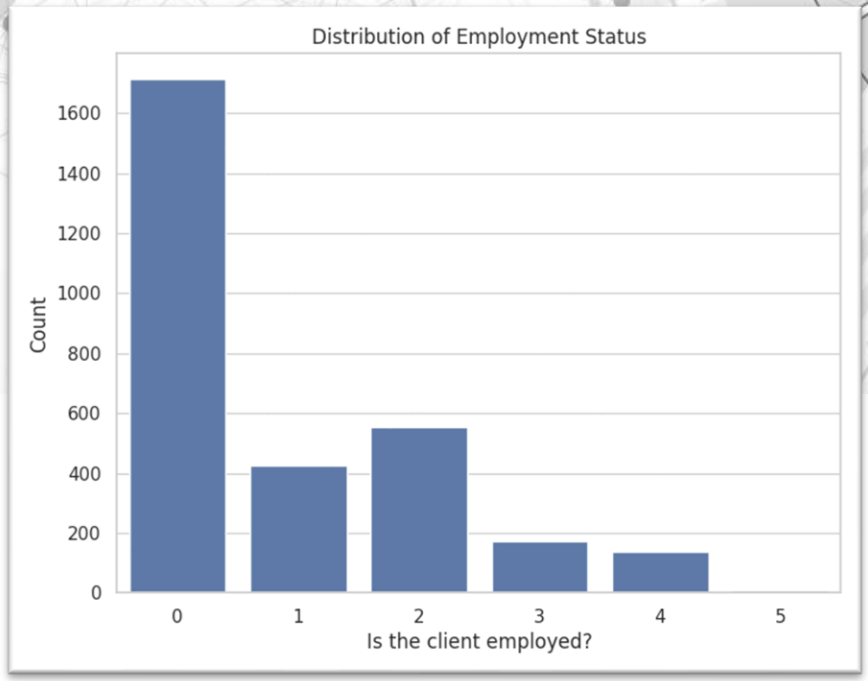
- **Gradient Boosting** has the highest F1 Score, indicating that it is the best-performing model in terms of accuracy, considering both precision and recall.
- **k-Nearest Neighbors** and **Random Forest** follow closely behind Gradient Boosting, also showing high F1-Scores, suggesting strong performance.
- **Decision Tree** shows a moderate F1-Score, performing well but not as strongly as the top three models.
- **Logistic Regression** and **Support Vector machines have significantly lower F1 scores**, indicating they performed poorly compared to the other models.

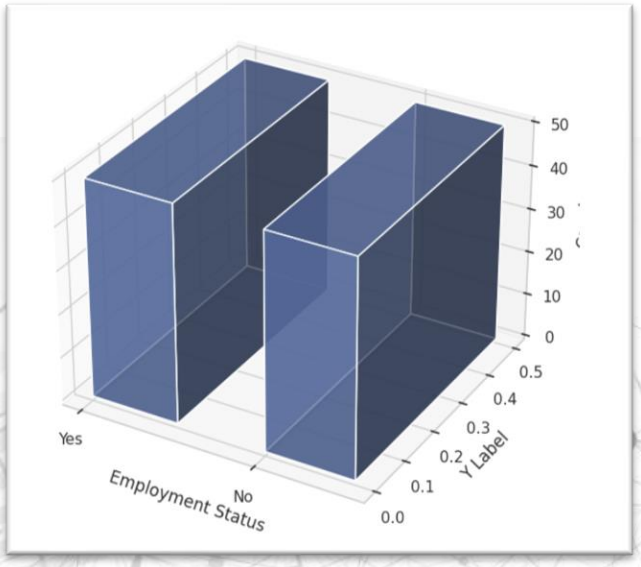
Printed Output:

The best classification model is **Gradient Boosting**, with an F1 Score of **0.72**. This means that among all the models tested, Gradient Boosting was the most effective at balancing precision and recall, making it the most suitable model for this classification task based on the given dataset.

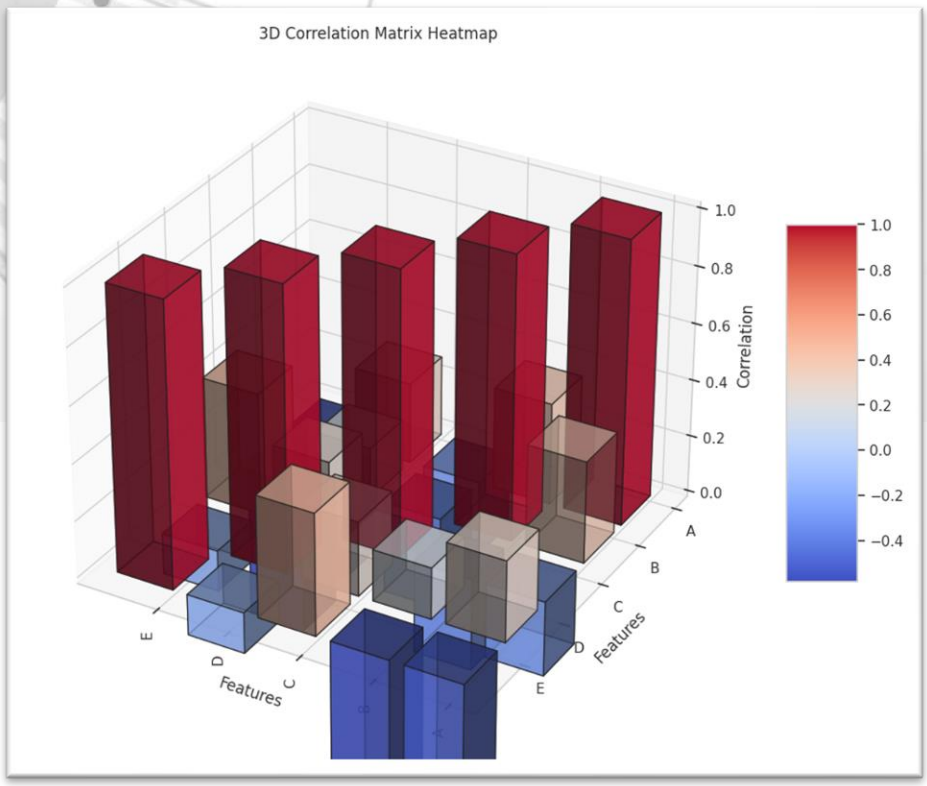
DATA VISUALIZATION





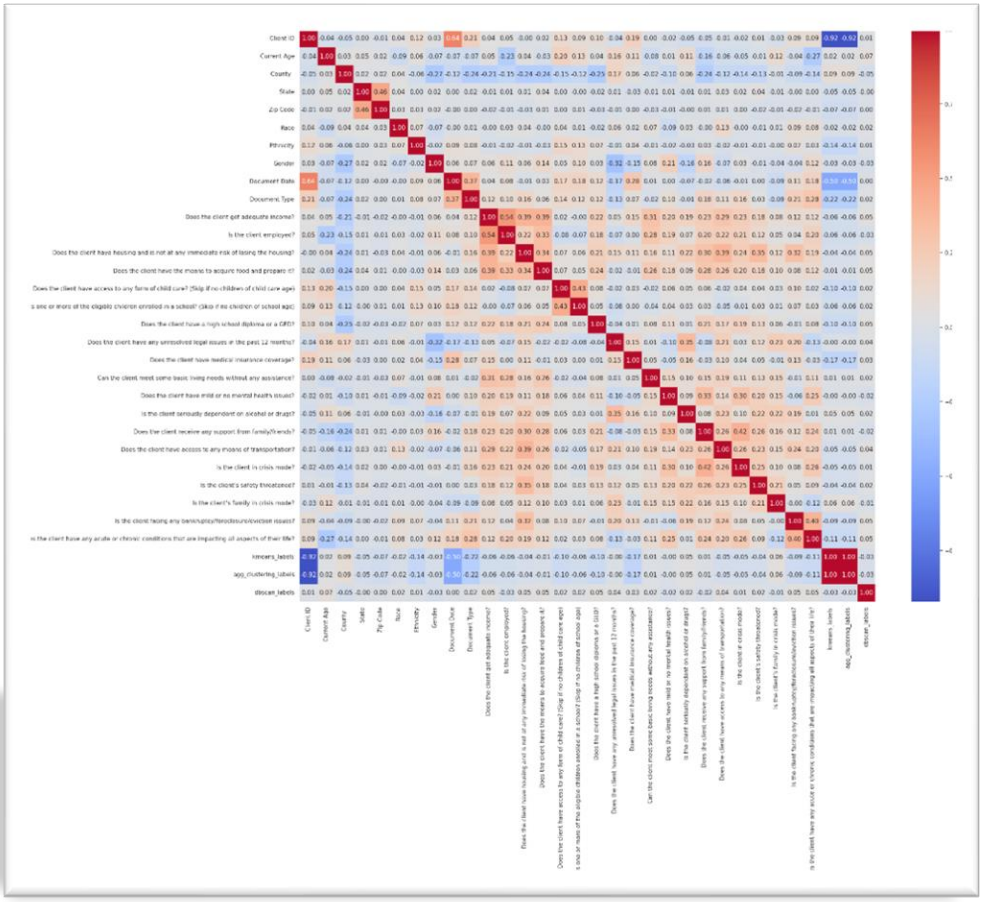


- The bar plot shows the distribution of the `Is the client employed?` variable across six categories.
- Most clients fall into category `0.0`, indicating many unemployed clients.
- Categories `1.0`, `2.0`, and `3.0` also have noticeable counts, representing various employment statuses.
- Categories `4.0` and `5.0` have very few clients, indicating these employment statuses are less common.
- This distribution highlights the employment landscape of the clients, with a significant proportion being unemployed.
- Understanding these categories can help in further analysis and targeted



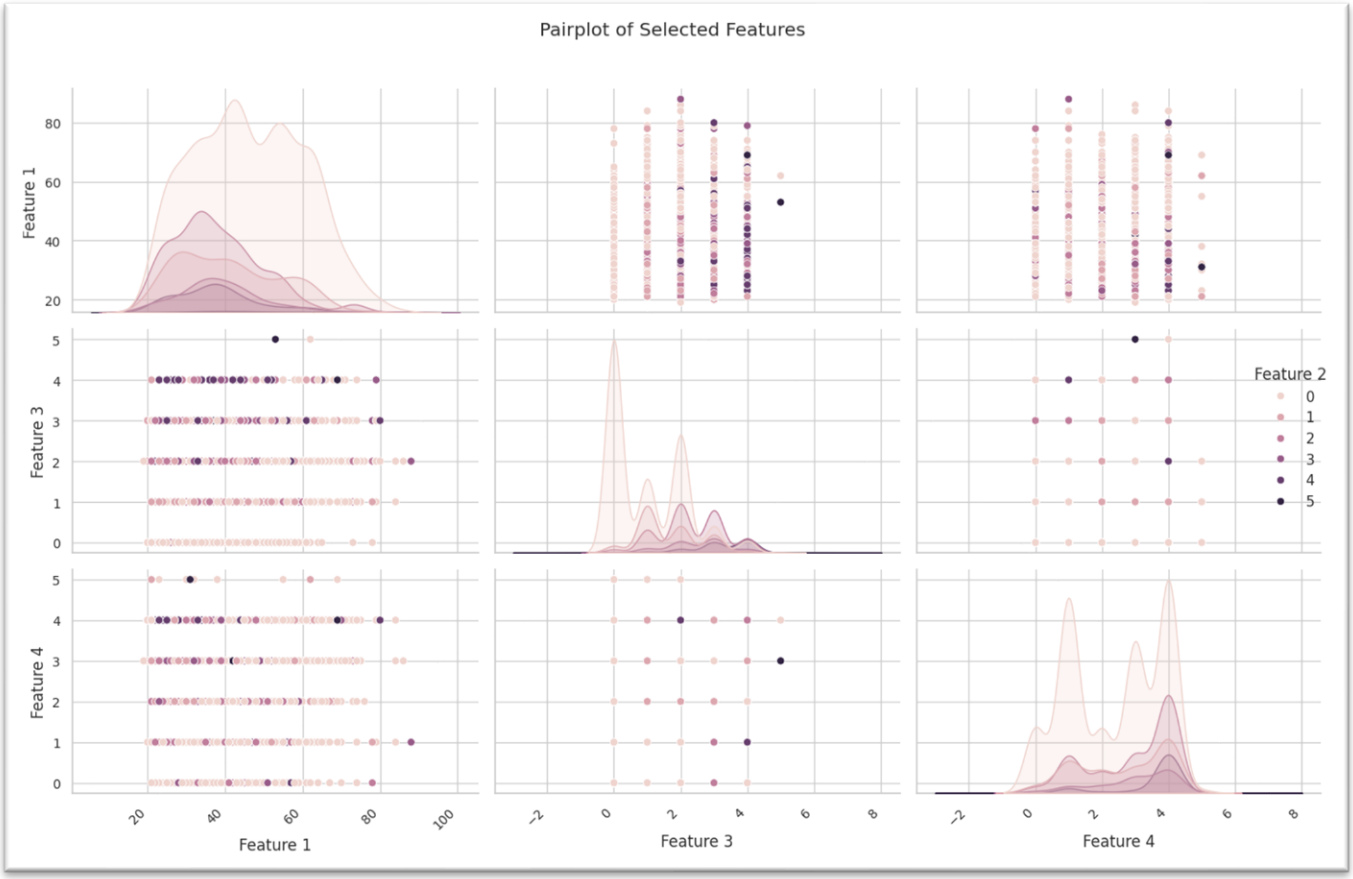
interventions.

HEAT MAP



- Clustering labels have a strong negative correlation with Client ID and Document Date.
- Current Age shows weak correlations with most variables but slightly negatively correlated with employment status (-0.23).
- Income adequacy and employment status are moderately positively correlated (0.54), indicating a link between job status and financial stability.
- Housing stability shows moderate positive correlations with income adequacy (0.39) and employment (0.32).
- Document Date positively correlates with Client ID (0.64) but negatively with clustering labels.
- Gender has a moderate negative correlation with clustering labels.
- Family/friend support moderately correlates with transportation access (0.42).
- Medical insurance coverage correlates with the ability to meet basic living needs without assistance (0.31)

PAIR PLOT

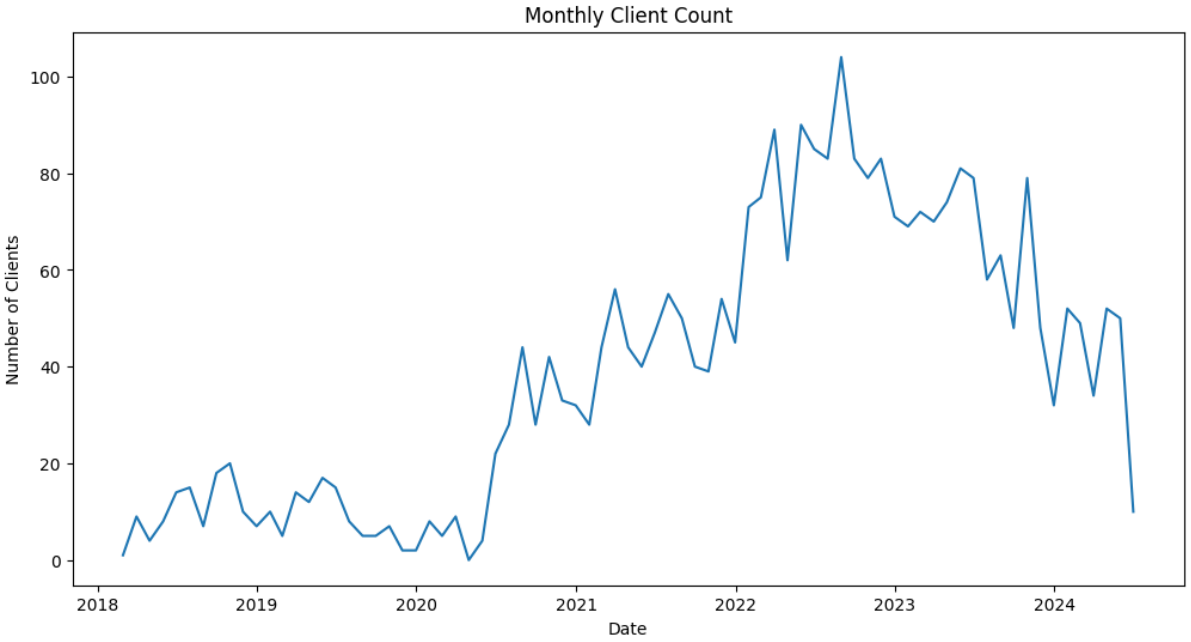


Key Observations

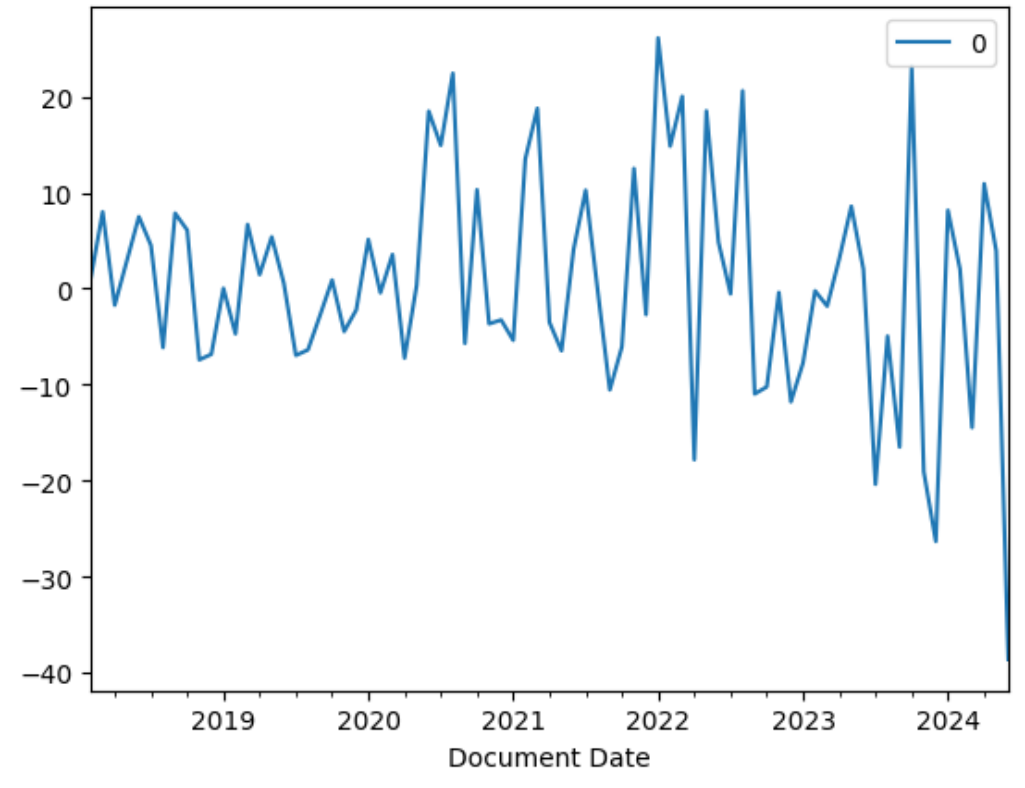
- **Age Distribution**: Most clients are between 20 and 70, with peaks indicating standard age ranges.
- **Discrete Features**: Features 2, 3, and 4 have discrete values, indicating categorical responses.
- **Pairwise Relationships**: The scatter plots show clusters indicating relationships between different features. For example, older clients might have different employment statuses than younger ones.
- **Multicollinearity Check**: The heatmap of the correlation matrix can further help identify if there are strong linear relationships between features.

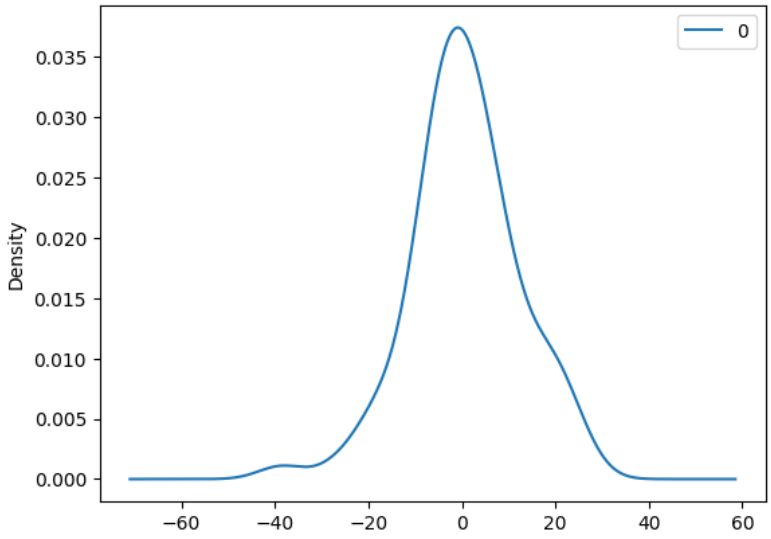
This plot helps to understand the distributions and relationships between different features in the dataset, providing insights into potential patterns and correlations.

TIME SERIES ANALYSIS

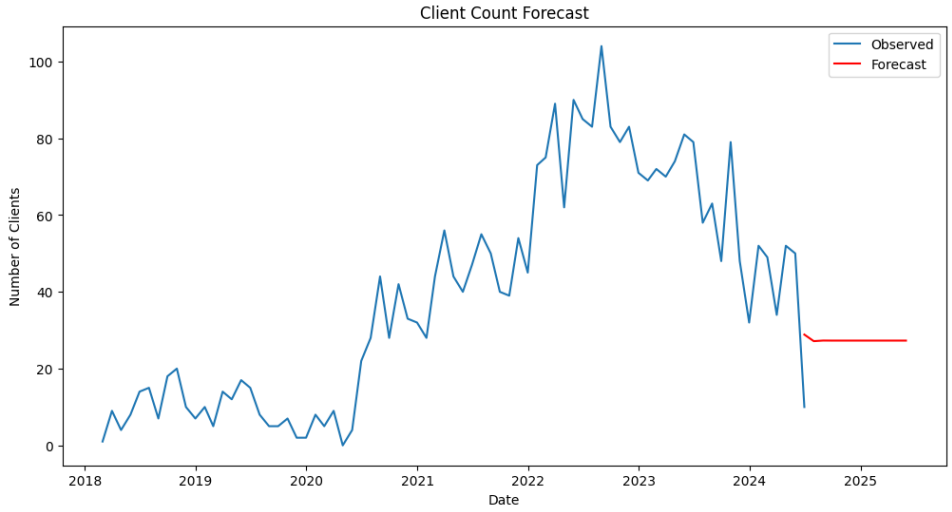


```
ADF Statistic: -1.3405926072937422
p-value: 0.6104027131442795
                               SARIMAX Results
==========================================================================================
Dep. Variable:                      Client ID   No. Observations:                   77
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -293.811
Date:                Sun, 04 Aug 2024   AIC                            593.621
Time:                        19:51:50   BIC                            600.613
Sample:                    02-28-2018   HQIC                           596.416
                         - 06-30-2024
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0907      0.218     -0.416      0.677      -0.518       0.336
ma.L1         -0.3944      0.202     -1.955      0.051      -0.790       0.001
sigma2       133.0491     17.958      7.409      0.000      97.852     168.246
==========================================================================================
=
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):
4.12
Prob(Q):                              0.90   Prob(JB):
0.13
Heteroskedasticity (H):               8.12   Skew:                             -
0.31
Prob(H) (two-sided):                  0.00   Kurtosis:
3.96
==========================================================================================
=

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-
step).
```





	0
count	77.000000
mean	0.615389
std	11.525736
min	-38.625365
25%	-6.066851
50%	0.033081
75%	7.496217
max	26.117625



Time Series Analysis

ARIMA

- **Purpose**: This model forecasts future values in a time series, such as the number of monthly clients needing services.
- **Use Case**: Forecasting Client Demand
    - **Benefits**:
        - Predict future demand for services.
        - Plan and allocate resources more effectively.
        - Anticipate periods of high demand and prepare accordingly.

Practical Benefits to DHIS2 or the Organization

1. **Improved Resource Allocation**:

   - Forecasting models can predict periods of high demand, enabling better planning and resource allocation.
2. **Targeted Interventions**:
   - Classification models help identify at-risk clients (e.g., unemployment), allowing for targeted support and interventions.
3. **Enhanced Understanding of Client Needs**:
   - Clustering models reveal patterns and segments within the client population, facilitating tailored services for different groups.
4. **Data-Driven Decision Making**:
   - Regression models identify vital factors impacting outcomes like income adequacy, guiding strategic decisions, and policy making.
5. **Monitoring and Evaluation**:
   - These models can track and measure the effectiveness of various programs and interventions over time.
6. **Automated and Scalable Solutions**:
   - Integrating these models into DHIS2 or similar systems automates the analysis, making it scalable and continuously updated with new data.

Example Integration into DHIS2

- **Risk Prediction Alerts**:
  - Classification models integrated into DHIS2 can trigger alerts for high-risk clients, prompting immediate action.
- **Resource Planning Dashboards**:
  - Forecasting models provide insights displayed on dashboards, aiding in resource planning and management.
- **Client Segmentation Reports**: Clustering results can be visualized in reports, helping staff understand the diverse needs of the client population.

WHY MACHINE LEARNING

The machine learning models applied to the dataset can offer significant insights and benefits to the organization, enhancing their ability to understand and respond to client needs. Here is a detailed explanation of how each type of model can be utilized:

Classification Models

1. Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, k-NN

- **Purpose**: These models predict categorical outcomes. For instance, they can predict whether a client will likely be employed based on their demographic and health information.
- **Use Case**: Predicting Employment Status
    - **Benefits**:
        - Identify clients at risk of unemployment.
        - Tailor employment support programs for those at higher risk.
        - Measure the impact of interventions aimed at improving employment rates.

Regression Models

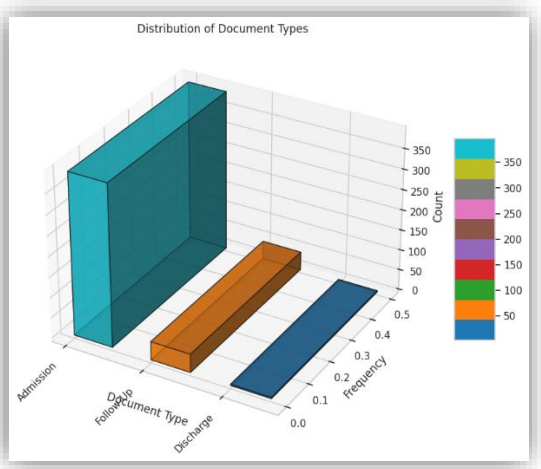
2. Linear Regression, Ridge, Lasso, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor

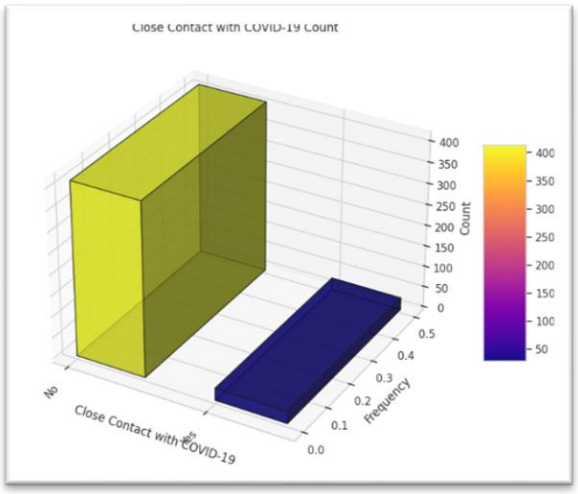
- **Purpose**: These models predict continuous outcomes. For example, they can predict clients' "Current Age" based on other variables, though a more meaningful use would be predicting continuous variables like income adequacy scores.
- **Use Case**: Predicting Adequate Income Scores
    - **Benefits**:
        - Assess factors contributing to income inadequacy.
        - Develop targeted financial support programs.
        - Evaluate the effectiveness of interventions on improving financial stability.

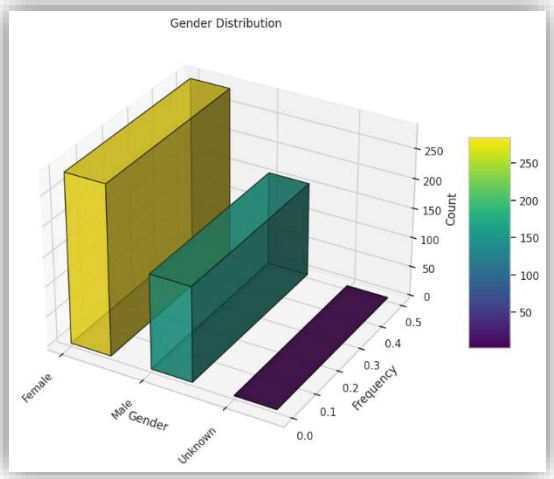
Clustering Models

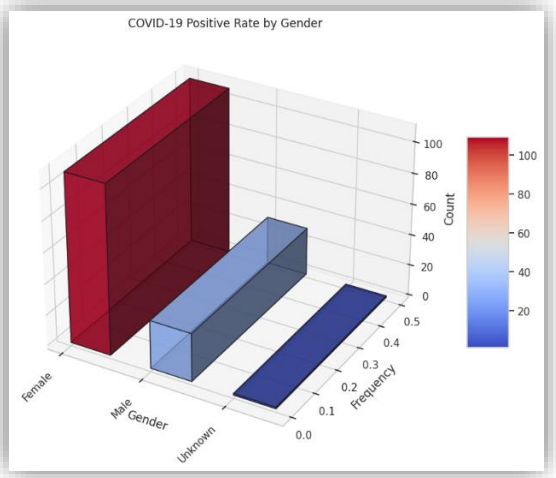
3. K-Means, Agglomerative Clustering, DBSCAN

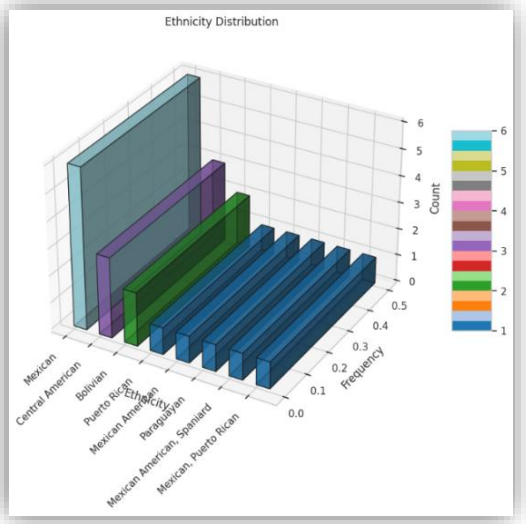
- **Purpose**: These models group clients into clusters based on similarities in their data. Clustering can reveal natural groupings within the client population.
- **Use Case**: Segmenting Clients
    - **Benefits**:
        - Identify distinct groups with similar needs or challenges.
        - Develop specialized programs for different client segments (e.g., those with severe mental health issues vs. those facing housing instability).
        - Understand the diversity of the client population to tailor services accordingly.

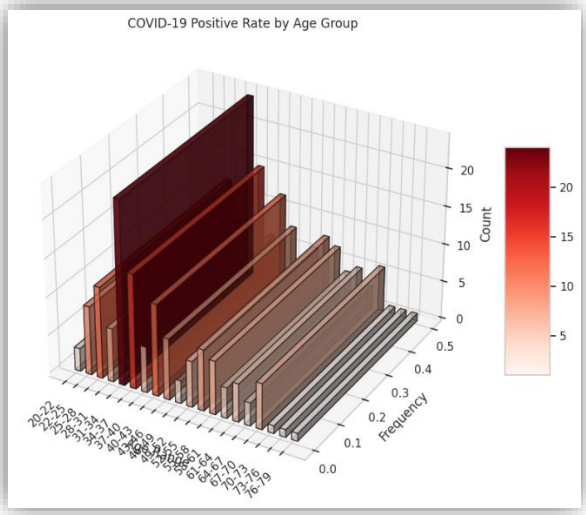

Distribution of Document Types
Admission
Follow-up
Discharge
Document Type
Frequency
Count


Close Contact with COVID-19 Count
No
Yes
Close Contact with COVID-19
Frequency
Count


Gender Distribution
Female
Male
Unknown
Gender
Frequency
Count

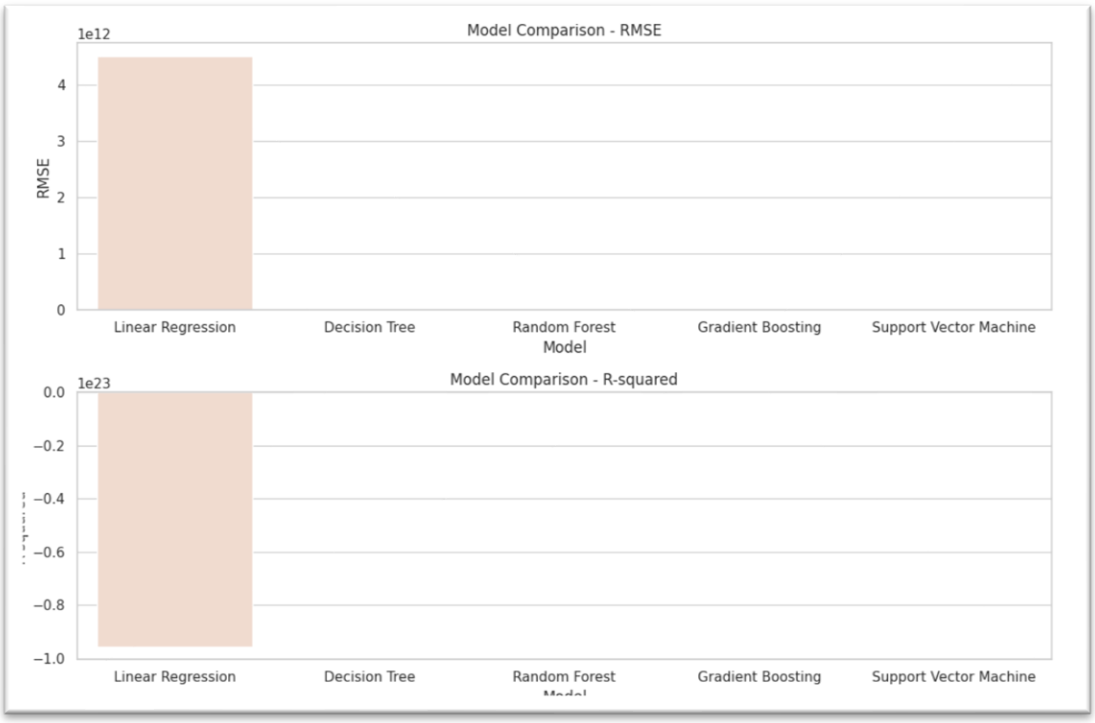

COVID-19 Positive Rate by Gender
Female
Male
Unknown
Gender
Frequency
Count


Ethnicity Distribution
Mexican
Central American
Bolivian
Puerto Rican
Mexican American
Ethnicity
Paraguayan
Mexican American, Spaniard
Mexican, Puerto Rican
Frequency
Count


COVID-19 Positive Rate by Age Group
Frequency
Count

DATASET 2 COVID 19 SCREENING

	Model	RMSE	R-squared
0	Linear Regression	4.518531e+12	-9.577600e+22
1	Decision Tree	1.589585e+01	-1.853037e-01
2	Random Forest	1.509615e+01	-6.904156e-02
3	Gradient Boosting	1.510430e+01	-7.019529e-02
4	Support Vector Machine	1.474326e+01	-1.964594e-02

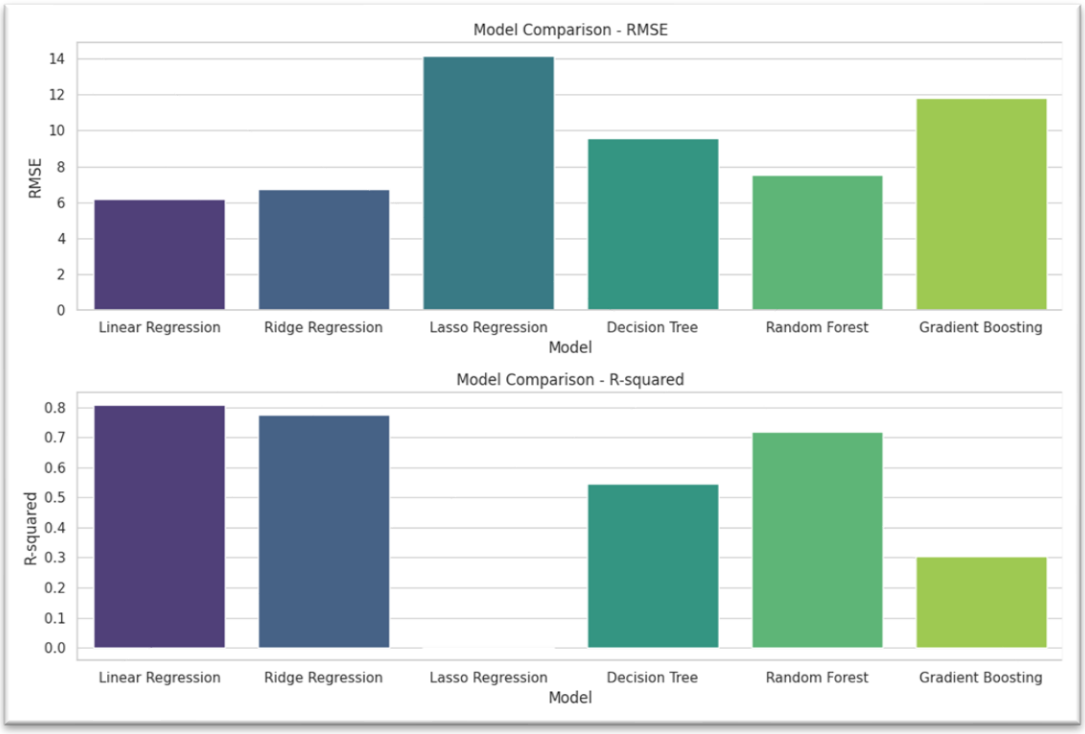


Linear Regression:
232.95737223762842 -0.13592901548851422
Ridge Regression:
226.61563361213385 -0.10500591207200016
Lasso Regression:
206.7992264777142 -0.008378655203050256
Decision Tree Regressor:
379.6179775280899 -0.8510643013064147
Random Forest Regressor:
240.96516629213477 -0.17497601163719811
Gradient Boosting Regressor:
214.92670144293578 -0.0480092303034354

Lasso Regression and Gradient Boosting Regressor are the best-performing models among those tested, as they have the lowest MSE and the least negative R^2 values.
Decision Tree Regressor performs the worst, likely due to **overfitting**.
Random Forest Regressor and Ridge Regression show intermediate performance but still have negative **R^2** values, indicating room for improvement.

DATASET 3 PHQ9 Depression Questionnaire

	Model	RMSE	R-squared
0	Linear Regression	6.195662	0.808513
1	Ridge Regression	6.724265	0.774445
2	Lasso Regression	14.170204	-0.001648
3	Decision Tree	9.554573	0.544608
4	Random Forest	7.504574	0.719059
5	Gradient Boosting	11.818029	0.303288



It compares the performance of various regression models on Dataset 3 using two evaluation metrics: RMSE (Root Mean Squared Error) and R-squared. Here are the key observations:

1. **Linear Regression**:
   - RMSE: 6.195662
   - R-squared: 0.808513
   - Shows a relatively low RMSE and high R-squared, indicating a good fit.
2. **Ridge Regression**:
   - RMSE: 6.724265
   - R-squared: 0.774445
   - Slightly higher RMSE than Linear Regression, with a slightly lower R-squared.
3. **Lasso Regression**:
   - RMSE: 14.170204
   - R-squared: -0.001648

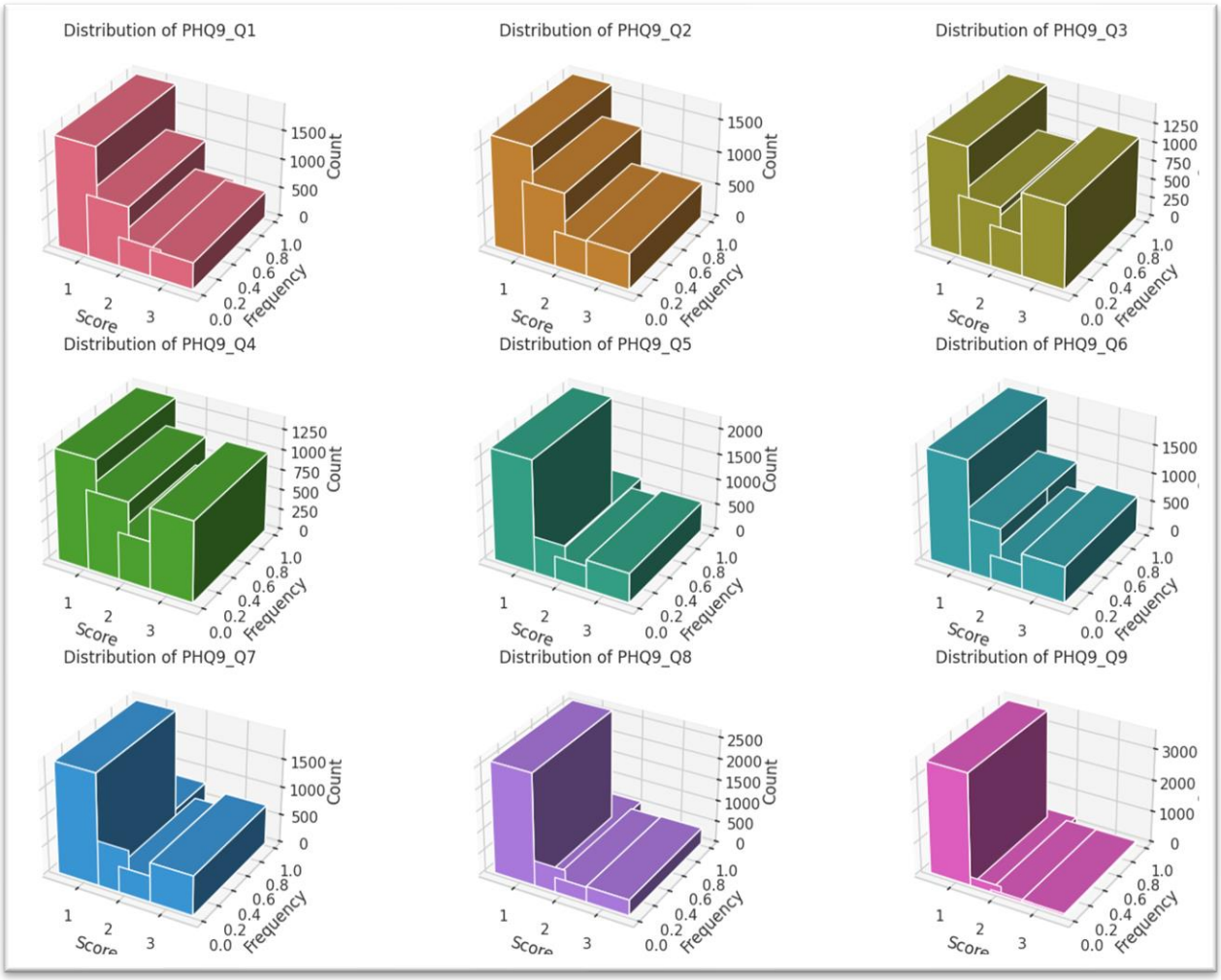
    - Significantly higher RMSE and negative R-squared, indicating poor performance and a bad fit.
4. **Decision Tree**:
    - RMSE: 9.554573
    - R-squared: 0.544608
    - Moderate RMSE and R-squared values, showing a decent fit but not as good as Linear or Ridge Regression.
5. **Random Forest**:
    - RMSE: 7.504574
    - R-squared: 0.719059
    - RMSE is slightly higher than Linear Regression but better than Decision Tree, with an excellent R-squared value.
6. **Gradient Boosting**:
    - RMSE: 11.818029
    - R-squared: 0.303288
    - Higher RMSE compared to Linear Regression and a lower R-squared, indicating it did not perform as well as expected.

Bar Plots:

- **Model Comparison - RMSE**:
    - Visual comparison shows **Lasso Regression** with the highest RMSE, indicating the **worst** performance.
    - **Linear Regression** has the **lowest RMSE**, suggesting the **best** performance among the models.
- **Model Comparison - R-squared**:
    - **Linear Regression and Ridge Regression** have the highest R-squared values, showing the **best** fit to the data.
    - **Lasso Regression** has a negative R-squared, indicating it did not capture the variance in the data well.
    - **Random Forest and Decision Tree** also perform well but not as high as Linear and Ridge Regression.

The comparison indicates that Linear Regression provides the best balance of low RMSE and high R-squared, making it the most suitable model for Dataset 3 in this context. Lasso Regression performs the worst among the evaluated models.

BAR PLOTS



The image contains a series of 3D bar plots representing the distributions of scores for different questions (Q1 to Q9) from the PHQ-9 (Patient Health Questionnaire-9), a tool commonly used to screen for depression.

Each plot shows:

- **Score** on the x-axis, which typically ranges from 0 to 3.
- **Count** on the y-axis, indicating the number of respondents who chose each score.
- **Frequency** on the z-axis represents the respondents' proportion for each score.

Observations:

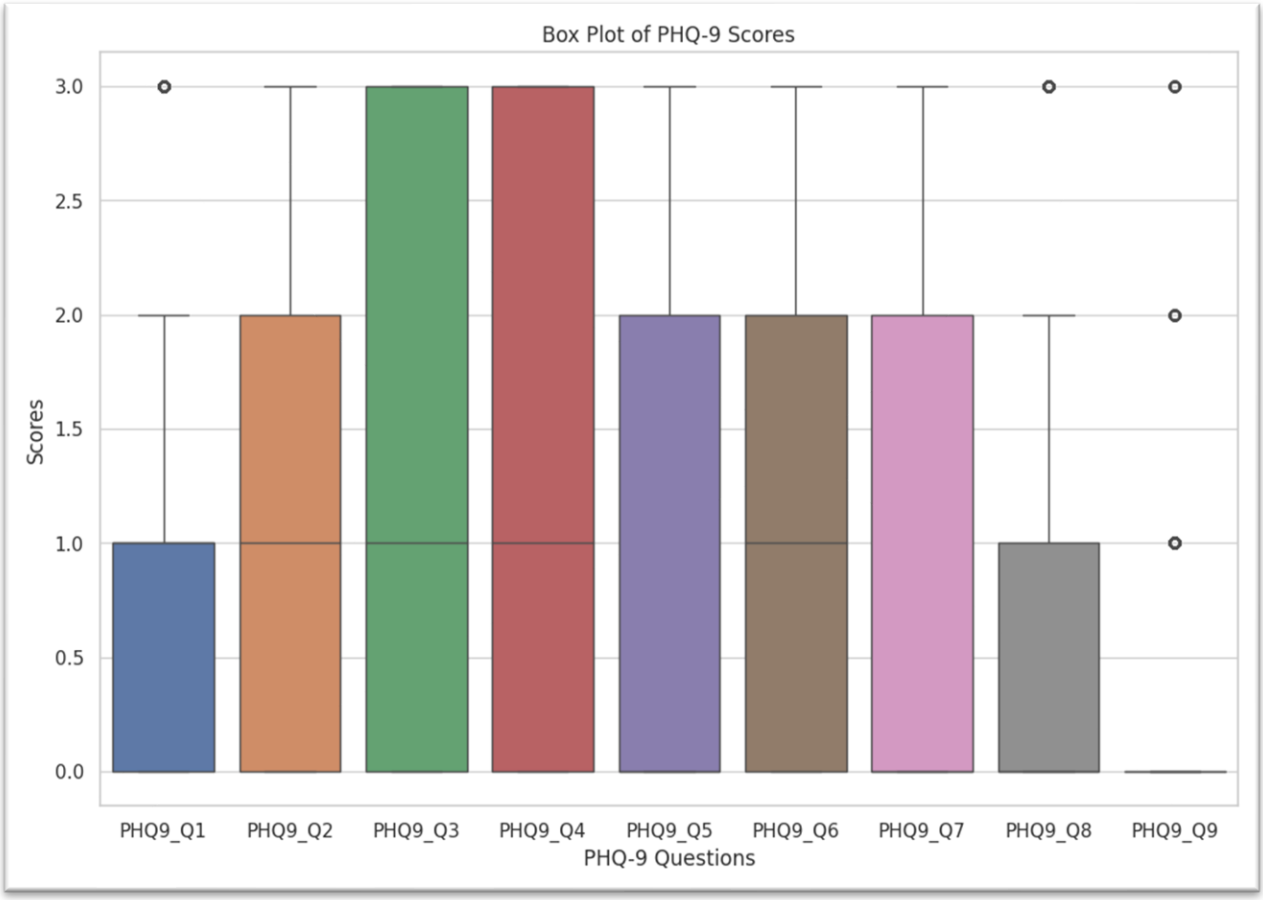
1. **Distribution of PHQ9_Q1 to PHQ9_Q9**:
   - **Q1** (top left) shows a higher count and frequency for lower scores (0 and 1), suggesting most respondents experienced little to no issues related to the question.
   - **Q2** and **Q3** (top center and top right) follow similar patterns, with the majority scoring lower, indicating fewer symptoms.

- **Q4** and **Q5** (middle left and middle center) also show higher frequencies for scores 0 and 1, with a slight increase in higher scores compared to the first three questions.
- **Q6** and **Q7** (middle right and bottom left) exhibit a more even distribution across all scores, indicating a more varied response among the respondents.
- **Q8** and **Q9** (bottom center and bottom right) demonstrate a skew towards lower scores, particularly Q9, with the highest count for a score of 0.

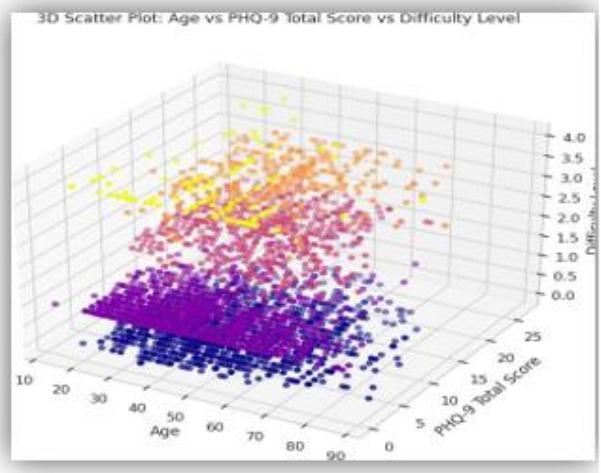
General Trends:

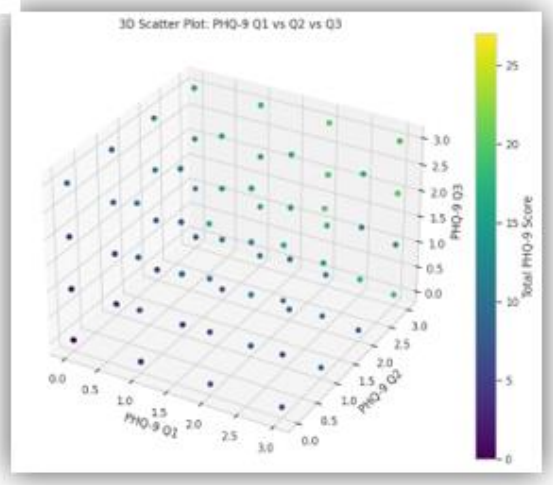
- **Lower Scores Dominance**: Most questions have higher counts and frequencies for scores 0 and 1, suggesting that most respondents report lower depression symptoms for these specific questions.
- **Varied Distributions**: Some questions (like Q6 and Q7) have a more balanced distribution, indicating varied responses and potentially highlighting areas where more respondents experience higher symptoms.
- **Count and Frequency Relationship**: The relationship between count and frequency helps to visualize how many respondents chose each score and the proportion of the total responses, providing a clearer picture of the overall distribution.

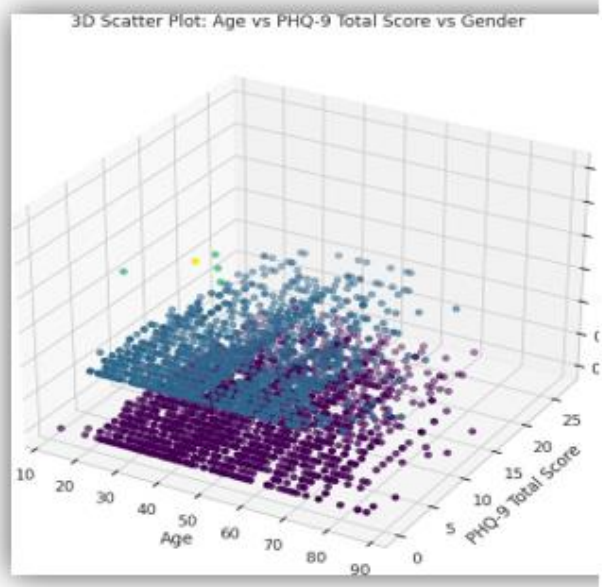
These visualizations help identify patterns in the responses to each PHQ-9 question, offering insights into which symptoms are more commonly reported and the overall distribution of depression severity among the surveyed population.

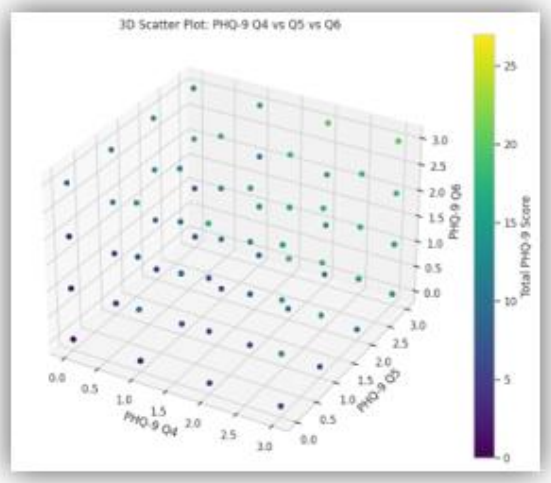


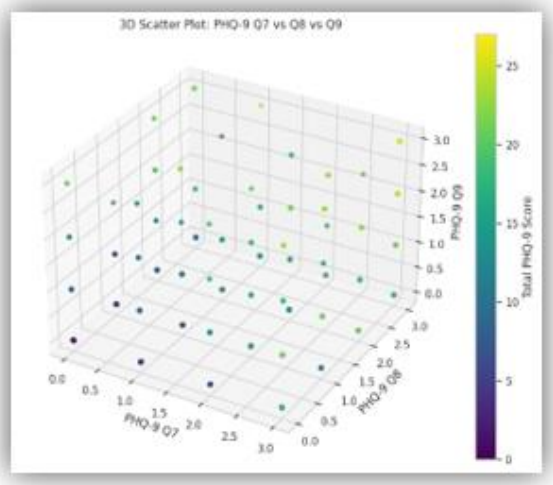
Overall Insights from PHQ-9 3D Scatter Plots

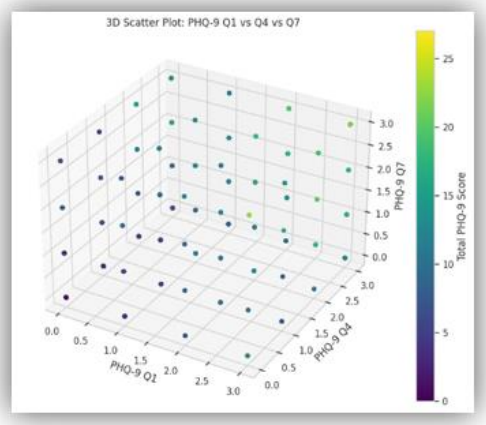


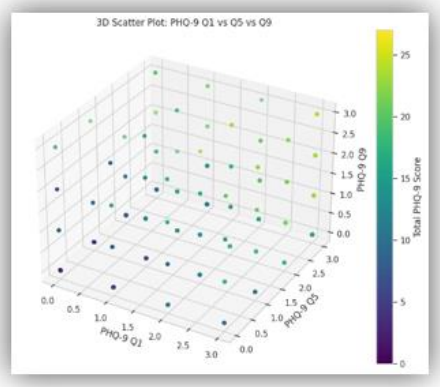


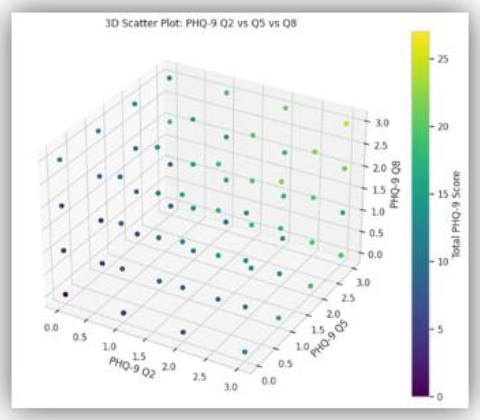


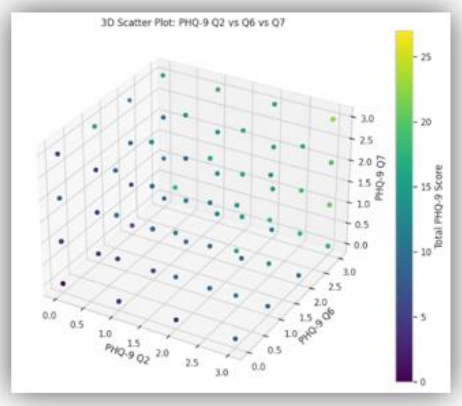


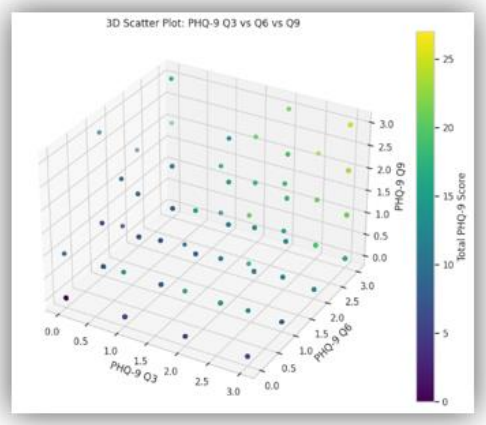


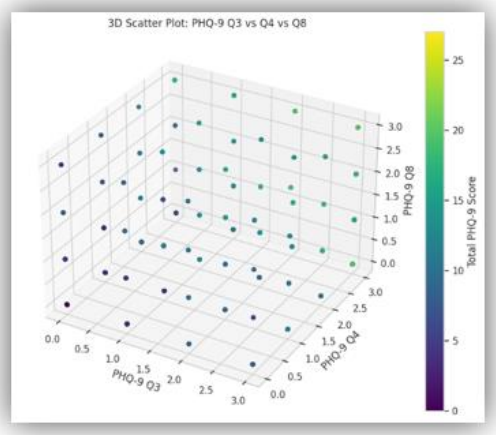










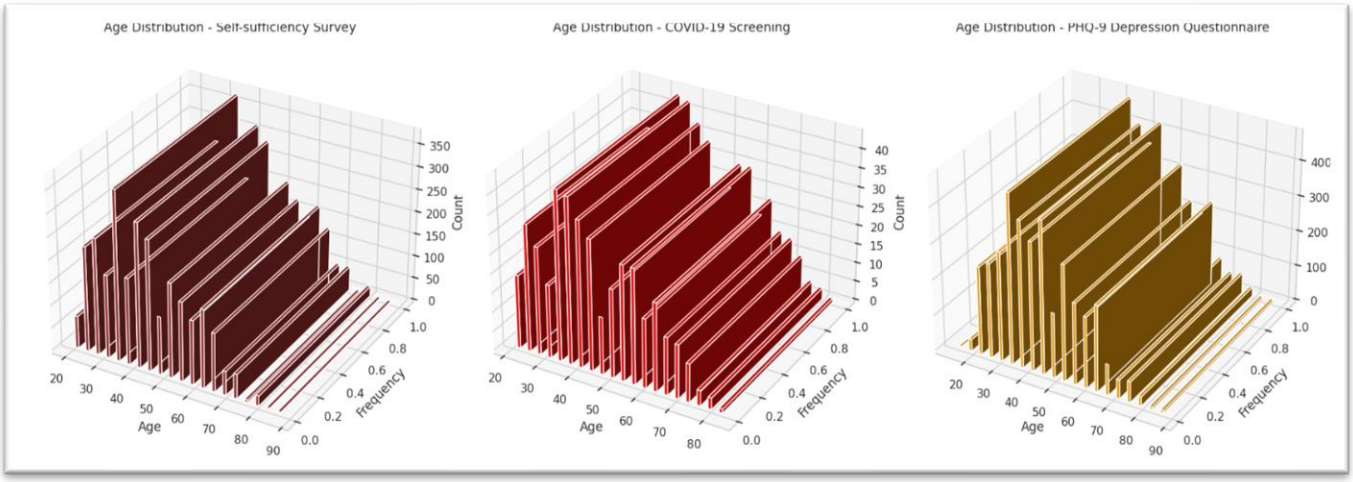


Overall Insights:

- **Age and PHQ-9 scores**: There is a broad distribution of PHQ-9 scores across different ages, with higher scores being less frequent among older individuals.
- **Gender and PHQ-9 scores**: The scatter plots show varying PHQ-9 scores for different genders, indicating no solid gender-specific trends in depression severity.
- **Difficulty Levels**: Higher PHQ-9 scores are often associated with higher reported difficulty levels.

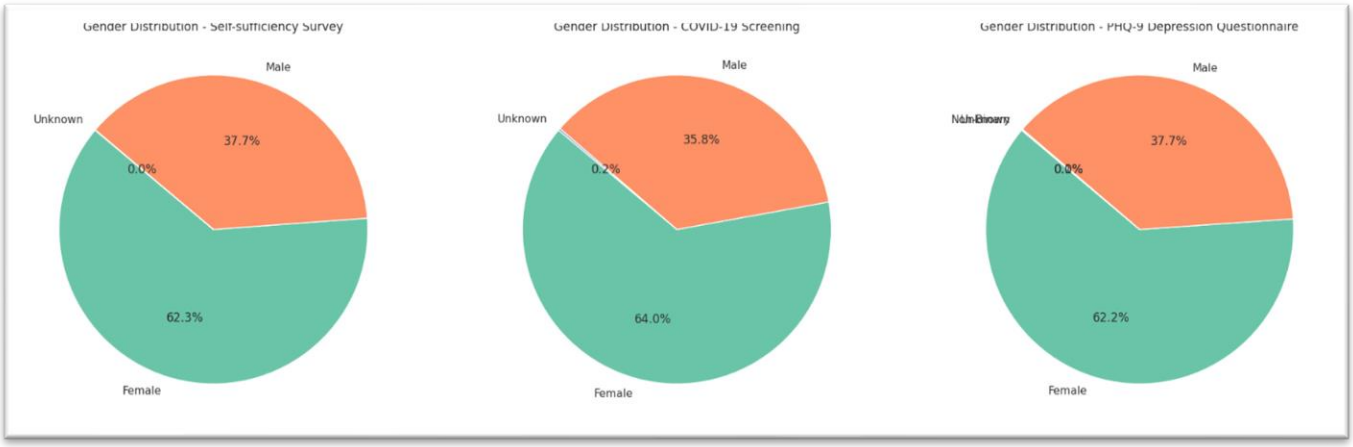
- **Question Interactions**: Individual responses to PHQ-9 questions show diverse patterns, with color gradients indicating how these responses correlate with the overall depression severity.
- **Clusters and Trends**: The visualizations help identify clusters or trends in the data, useful for further analysis or intervention planning based on PHQ-9 responses

COMPARISION OF 3 DATASETS



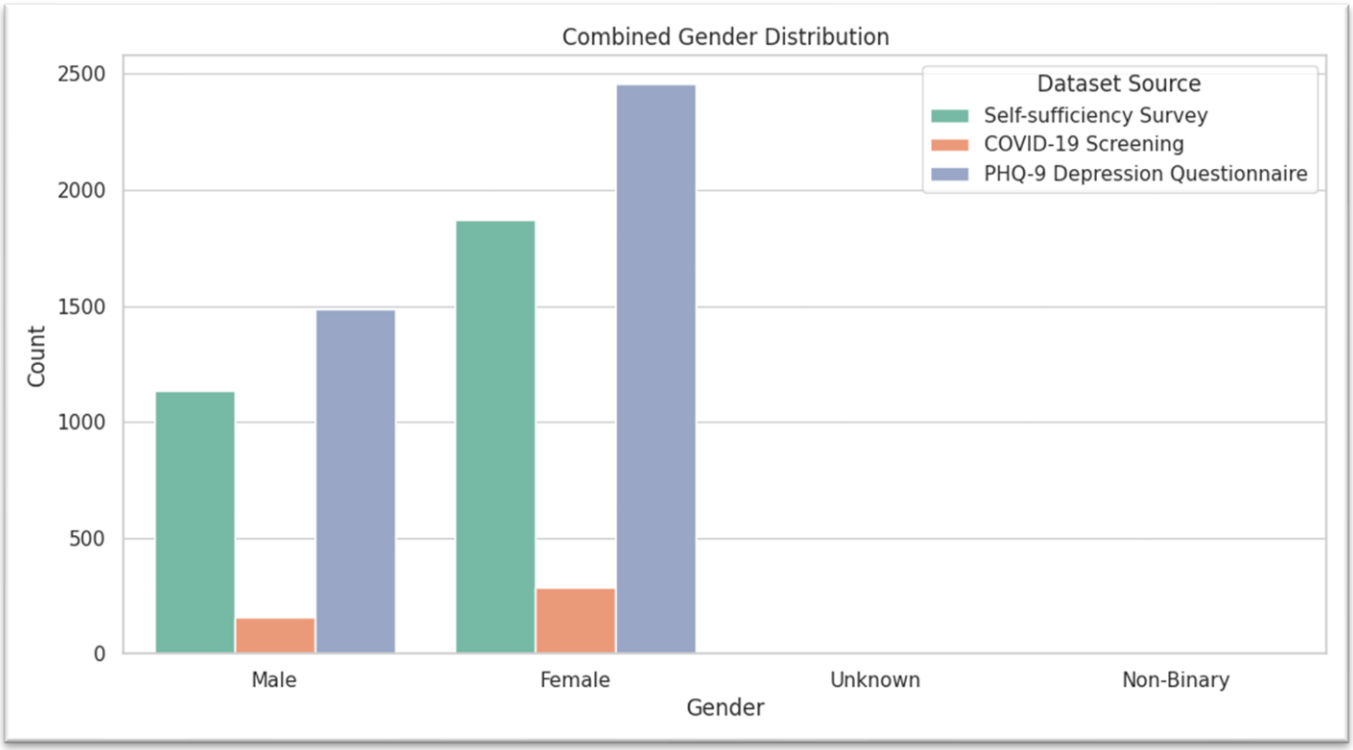
Overall Insights:

- **Consistent Trends**: All three surveys show more respondents in the middle age range (30-60 years), indicating that these age groups participate more frequently in these surveys.
- **Lower Participation at Extremes**: There is a noticeable decrease in participation among younger (under 30) and older (above 70) age groups across all three surveys.
- **Survey Participation**: The frequency and count data in these 3D plots help visualize the relative engagement of different age groups in various health-related surveys.



General Observations:

- **Higher Female Participation**: Across all three surveys, female respondents consistently outnumber male respondents.
- **Minimal Unknown/Non-Binary Data**: The unknown and non-binary categories are either absent or minimally represented, indicating a lack of data in these gender categories.
- **Consistent Trends**: The percentage of female participation remains relatively stable across the surveys, ranging from 62.2% to 64.0%, while male participation ranges from 35.8% to 37.7%.



Gender Participation: Females have a higher participation rate than males across all three surveys (Self-Sufficiency Survey, COVID-19 Screening, PHQ-9 Depression Questionnaire). This is evident from the higher counts in the Female category.

How the Analysis Helps with Capstone Project

Time Series Analysis

Relevance to Capstone Project:

- **Forecasting Health Trends**: Time series analysis can predict future health trends and client needs, allowing public health professionals to anticipate and respond proactively.
- **Resource Allocation**: The project can better allocate resources and staff to meet anticipated demands by forecasting client counts and other key metrics.
- **Integration with Dashboard**: The insights from time series analysis can be integrated into the DHIS 2 and Power BI dashboards, providing users with real-time predictive analytics.

Practical Application:

- **Example**: If the analysis predicts an increase in patients needing certain health services, RCCCP can prepare by allocating more resources or initiating preventive measures.

Regression Models

Relevance to Capstone Project:

- **Identifying Key Factors**: Regression models can identify key factors influencing public health outcomes, such as age, employment status, income adequacy, and housing stability.
- **Targeted Interventions**: Understanding these factors helps design targeted interventions to improve health outcomes.
- **Predictive Insights**: Regression models provide insights into the relationships between various health indicators and outcomes, which can be crucial for public health decision-making.

Practical Application:

- **Example**: If the model shows that housing stability significantly impacts health outcomes, RCCCP can focus on housing support programs to improve public health.

Classification Models

Relevance to Capstone Project:

- **Risk Identification**: Classification models help identify high-risk populations or areas, enabling RCCCP to prioritize interventions.
- **Real-time Decision Support**: Public health professionals can receive real-time alerts and recommendations by integrating these models into the DHIS 2 and Power BI dashboard.

Practical Application:

- **Example**: A classification model predicting the likelihood of clients needing emergency services can trigger proactive measures to prevent crises.

Clustering Analysis

Relevance to Capstone Project:

- **Segmenting Populations**: Clustering helps segment the population into groups with similar characteristics, making it easier to design and implement tailored public health programs.
- **Understanding Demographics**: By identifying natural groupings, RCCCP can better understand the diverse needs of different demographic groups.

Practical Application:

- **Example**: Clustering can reveal distinct groups within the population, such as those with high health risks due to socio-economic factors, allowing for targeted health interventions.

Detailed Example of Application:

Project Phases and Integration:

1. **Data Collection and Management**:
    - **Process**: Collect data using DHIS 2, ensuring it is clean and well-structured.
    - **Integration**: Use insights from regression models to validate the importance of collected data elements.
2. **Technical Setup and Configuration**:
    - **Process**: Set up the DHIS 2 server and necessary software components.
    - **Integration**: Ensure data collected is aligned with the needs identified by regression and classification models for future analysis.
3. **Development of Machine Learning Models**:
    - **Classification Models**: Predict high-risk areas and populations.
        - **Outcome**: Use models to identify clients at risk of severe health issues and provide early interventions.
    - **Regression Models**: Understand key factors influencing health outcomes.
        - **Outcome**: Design public health initiatives targeting identified factors (e.g., improving income adequacy).
    - **Clustering**: Group similar health outcomes or demographic profiles.
        - **Outcome**: Develop specialized programs for different clusters (e.g., focused health campaigns).
4. **Dashboard Creation**:
    - **Integration**: Incorporate time series forecasts to show predicted health trends.
    - **Visualization**: Use Power BI to visualize regression and clustering results, making it easy for users to interpret and act on the data.
    - **Interactive Features**: Allow users to explore scenarios based on predictive models.
5. **User Testing and Feedback Collection**:
    - **Process**: Conduct testing sessions to gather feedback on dashboard usability.
    - **Integration**: Refine models and visualizations based on user input to improve accuracy and relevance.
6. **Final Evaluation**:
    - **Process**: Validate the models and dashboard with new data.
    - **Integration**: Ensure the decision support system meets all project objectives and provides actionable insights.

Impact on Organization and Skills:

Organizational Impact:

- **Enhanced Decision-Making**: Integrating predictive models into the dashboard empowers RCCCP with real-time, data-driven decision support.
- **Improved Public Health Outcomes**: Targeted interventions based on model insights lead to better health outcomes and resource utilization.

Skills Development:

- **Technical Skills**: Enhanced understanding of machine learning, data management, and system integration.
- **Analytical Skills**: Improved ability to apply theoretical knowledge to practical challenges in public health.
- **Leadership and Collaboration**: Experience leading technical projects and working collaboratively with diverse teams.

Summary:

Created a comprehensive decision support system incorporating time series analysis, regression, classification, and clustering into DHIS 2 and Power BI integration. This system will provide valuable insights, predict trends, and identify critical health outcome factors. It will significantly enhance RCCCP's ability to make informed decisions, improve public health strategies, and allocate resources effectively.

CONCLUSION

Overall Assessment of the Thesis/Project Experience

The capstone project, **"Development of Machine Learning Enabled Decision Support Dashboard Using DHIS2 and Power BI for Public Health Informatics,"** has made substantial strides in leveraging technology to enhance public health decision-making. The project's successful integration of DHIS2 with Power BI and the implementation of machine learning models have created an innovative decision support dashboard, demonstrating the transformative potential of data-driven approaches in public health.

Key Achievements and Impact:

1. **Enhanced Data Analytics Capabilities:**
    - By combining DHIS2's robust data collection and management with Power BI's advanced visualization features, the project has produced a powerful tool that enables public health professionals to make more informed decisions based on real-time data and predictive insights.
2. **Successful Installation and Configuration:**
    - The meticulous installation and configuration of the DHIS2 server and robust data-cleaning protocols ensured high data quality and laid a strong foundation for subsequent analysis and machine learning phases.
3. **Development and Integration of Machine Learning Models:**
    - Utilizing Python and popular ML libraries, the project developed classification, regression, and clustering models that facilitate predictive analytics and the identification of key health trends. Integrating these models with Power BI has enabled real-time predictions and dynamic dashboard updates.
4. **Creation of an Interactive Dashboard:**

- The user-friendly DHIS2 dashboard includes geospatial maps, time series charts, and comparative visualizations, which streamline data management processes, reduce manual effort, and improve reporting efficiency.

5. **Challenges Addressed:**
    - The project tackled challenges such as version compatibility, performance issues, data migration, and user adoption through thorough testing, database indexing strategies, custom ETL scripts, and comprehensive training sessions.
6. **Comprehensive Evaluation:**
    - The final evaluation included thorough testing of dashboard features, validation of ML model predictions, performance testing, and security audits, ensuring the system met all project objectives and provided actionable insights.

Organizational Impact:

- The project has had a transformative impact on Randolph County Caring Community Partnership (RCCCP). The decision support dashboard has enhanced RCCCP's ability to visualize and interpret complex health data, leading to more informed decision-making and better resource allocation. This advancement underscores RCCCP's commitment to improving health outcomes through innovative solutions.

Personal and Professional Growth:

- The project experience has been highly rewarding and transformative, allowing for the application and expansion of technical and analytical skills. Integrating advanced models and developing a comprehensive decision-support system has significantly contributed to professional growth and reinforced the commitment to leveraging technology and data to address public health challenges.
- The hands-on experience underscored the significance of data-driven decision-making in public health. Overcoming challenges required resilience and adaptability while engaging with teams and incorporating feedback was crucial for refining the dashboard. This iterative process highlighted the importance of user-centered design and continuous improvement.
- The successful development of the decision support dashboard marked a significant milestone in the academic and professional journey, providing a solid foundation for future endeavors in the field.

Impact on RCCCP:

1. **Transformative Organizational Impact:**
    - The project provided RCCCP with a powerful tool for real-time data insights and predictive analytics, enhancing the organization's ability to visualize and interpret complex health data and leading to more informed decision-making.
2. **Streamlined Data Management:**

- The dashboard streamlined data management processes, reducing manual effort and improving data handling and reporting efficiency, allowing RCCCP to allocate more resources towards actionable insights and public health interventions.

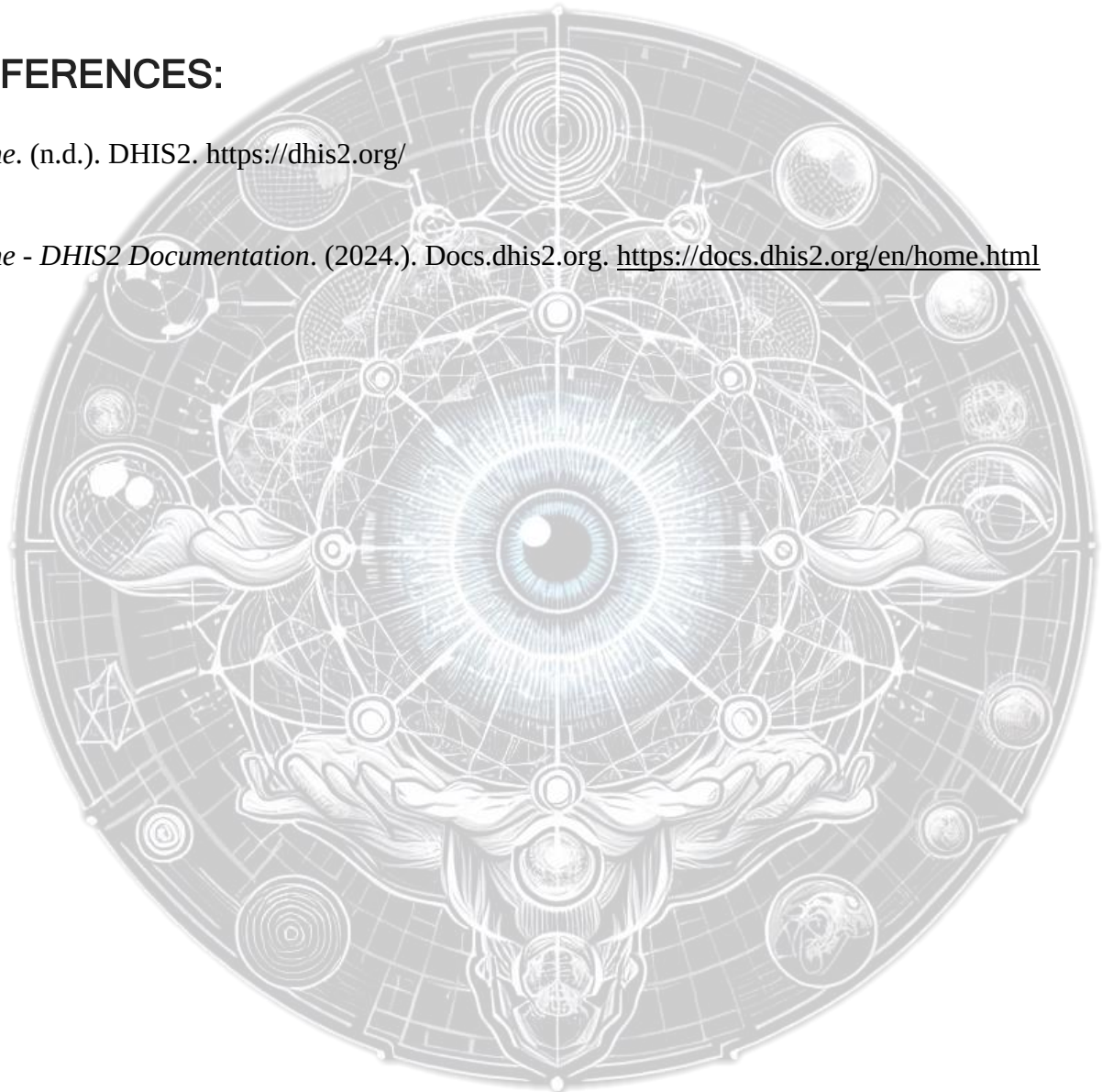
3. **Strengthened Data-Driven Approach:**
    - The project strengthened RCCCP's data-driven approach, supporting more proactive and effective public health strategies and significantly advancing the organization's technological capabilities.

In conclusion, this capstone project has demonstrated the immense potential of combining machine learning, data analytics, and powerful visualization tools to enhance public health informatics. The successful development and deployment of the decision support dashboard has provided valuable insights, predicted trends, and identified critical factors impacting health outcomes. This project has significantly enhanced RCCCP's ability to make informed decisions, improve public health strategies, and allocate resources effectively, marking a significant milestone in public health informatics.

REFERENCES:

Home. (n.d.). DHIS2. https://dhis2.org/

Home - DHIS2 Documentation. (2024.). Docs.dhis2.org. https://docs.dhis2.org/en/home.html



DEVELOPMENT OF A MACHINE LEARNING-ENABLED DECISION SUPPORT DASHBOARD USING DHIS 2 AND POWER BI FOR PUBLIC HEALTH INFORMATICS

INDIANA UNIVERSITY
INDIANAPOLIS

Sameer Mohammad,
Dr Zeyana Hamid, Ph.D.
Randolph County Caring Community Partnership(RCCCP), Date: 08/09/2024.
Indiana University, Luddy School of Informatics, Computing, and Engineering
Department of Health Informatics

dhis2
CARING COMMUNITY PARTNERSHIP

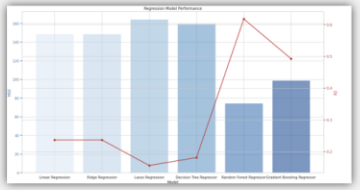
LUDDY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING
Indianapolis

SUMMARY/ANSTRACT

❑ This project aims to revolutionize public health decision-making by developing a machine learning-enabled decision-support dashboard. Integrating DHIS 2 with Power BI, the dashboard transforms raw health data into actionable insights for real-time and predictive analytics.

OUTCOMES

- ❑ Project and Configuration: Successfully installed and configured DHIS 2 server and required software
- ❑ ML Models: Developed models for classification, regression, and clustering
- ❑ Dashboard: Created an interactive dashboard with geospatial maps, time series charts, and comparative visualizations
- ❑ User Adoption: Conducted training sessions and collected feedback to refine the dashboard



PROJECT OBJECTIVES

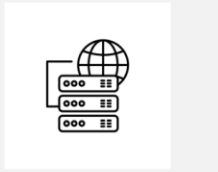
- ❑ **Organization:** Randolph County Caring Community Partnership (RCCCP)
- ❑ **Mission:** To use data analytics and informatics tools to improve public health outcomes.
- ❑ **Worksite:** Community Health Informatics project within RCCCP, focusing on enhancing data-driven decision-making through advanced analytics.
- ❑ **Administrative Structure:** Collaborative project management involving regular team meetings and feedback sessions.

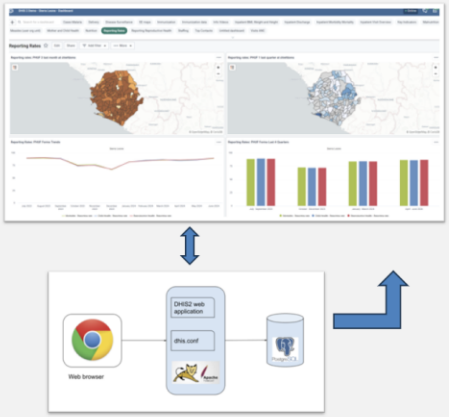
PROBLEM ADRESSED

- ❑ The project involved installing and configuring the DHIS 2 server, developing machine learning models, creating an interactive Power BI dashboard, and conducting user testing and feedback collection.

PROJECT EXECUTION

- ❑ **Task:** Development of a machine learning-enabled decision support dashboard.
- ❑ **Objectives**: Integrate DHIS 2 with Power BI, develop predictive ML models, and create a user-friendly dashboard.
- ❑ **Methodology**: Data collection from DHIS 2, Preprocessing and cleaning of data development and integration of ML model creation and refinement of an interactive dashboard
- ❑ **Specific Tasks:** Configuring Apache Tomcat, OpenJDK, PostgreSQL, and PostGIS Deploying the DHIS 2 WAR file; developing ML models using Sci-kit-learn and TensorFlow; Designing and implementing the Power BI dashboard
- ❑ **Tools Used**: DHIS 2, Power BI, Python, Sci-kit-learn, TensorFlow, Apache Tomcat, OpenJDK, PostgreSQL, PostGIS





TIMELINE

CONCLUSION

- ❑ The project significantly enhanced RCCCP's public health decision-making capabilities by developing a machine learning-enabled decision-support dashboard that integrated DHIS 2 with Power BI. This innovative solution transformed raw health data into actionable insights through predictive analytics and interactive visualizations.
- ❑ Key achievements included successfully installing and configuring the DHIS 2 server, developing various predictive models, and creating a user-friendly dashboard.
- ❑ The project faced and overcame data quality and tool integration challenges, ultimately improving data-driven decision-making for RCCCP.

Gratitude and Recognition

I am profoundly grateful to Professor Dr. Zeyana Hamid, who gave me the incredible opportunity to work on the DHIS2 capstone project. Her guidance, support, and encouragement have been instrumental in completing this project. Professor Hamid's expertise and dedication to fostering a deep understanding of public health informatics have greatly enriched my learning experience. Her commitment to excellence and ability to inspire her students have been remarkable.

"Education is the kindling of a flame, not the filling of a vessel." - Socrates. This quote resonates with my educational journey under Professor Hamid's mentorship. She has ignited a passion for learning and discovery, pushing me to explore new horizons and achieve my full potential. I am forever thankful for her unwavering support and for believing in my abilities.

-Sameer Mohammad