Healthcare Data Analytics Platform

A PROJECT REPORT

Submitted for the partial fulfillment

of

Capstone Project requirement of B. Tech CSE

Submitted by

1.SHAURYA POTDUKHE, 22070521011
2. RUTHVEK KANAN, 22070521031
3. SHREYAS KASTURE, 22070521032

B. Tech Computer Science and Engineering

Under the Guidance of

Dr. PIYUSH CHAUHAN

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Wathoda, Nagpur
2025

CERTIFICATE

This is to certify that the Capstone Project work titled **“Healthcare Data Analytics Platform”** that is being submitted b**y SHAURYA POTDUKHE [22070521011], RUTHVEK KANAN [22070521031], SHREYAS KASTURE [22070521032]** is in partial fulfillment of the requirements for the Capstone Project is a record of bonafide work done under my guidance. The contents of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for award of any degree or diploma, and the same is certified.

Dr. Piyush Chauhan
Capstone Project Guide

Verified by:

Dr. Parul Dubey
Capstone Project Coordinator

The Report is satisfactory/unsatisfactory

Approved by

Prof. (Dr.) Nitin Rakesh
Director, SIT Nagpur

ABSTRACT

Healthcare data requires efficient interactive systems because data-driven decision-making has become essential throughout the healthcare sector. The "Healthcare Data Analytics Platform" demonstrates a complete solution which analyzes and visualizes patient survey data acquired from Cancer Awareness and Rehabilitation Foundation (CARF). The platform uses contemporary data science tools along with Linear Regression, Logistic Regression, and Multiple Regression for machine learning which identifies behavioral and satisfaction patterns in patients and standard users undergoing cancer treatment. The tool implements interactive analytics and predictive modeling through its implementation of Python-library combination with Pandas and Scikit-learn together with Matplotlib and Plotly and Seaborn visualization options. The project delivers real-time analytics through Power BI dashboards together with Streamlit-based web applications for visualizations that serve technical and non-technical personnel. The study reveals that platform engagement and user satisfaction depend on three specific features which include user-friendliness and symptom checkers and newsletter subscriptions. Research results function to detect improvement priorities and determine future feature enhancement plans. Through this platform CARF joins other NGOs in improving their understanding of patient health needs and shows how data intelligence drives healthcare service development. The proposed work involves growing the available data collection and adding sophisticated machine learning algorithms while implementing up-to-date info from mobile and internet applications with the goal of building an adaptive healthcare information system.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The healthcare industry experienced major changes throughout recent times because of increased interest in data-driven decisions. The implementation of digital technology created possibilities to acquire and analyze and use data for better patient healthcare alongside improved operational effectiveness and preventative health services. Healthcare institutions are facing difficulties transforming untreated data sources into meaningful clinical information even though digital healthcare solutions have expanded.

We created a Healthcare Data Analytics Platform which uses patient data analysis from interactions along with survey results combined with behavioral information for data comprehension. The project brought together the Cancer Awareness and Rehabilitation Foundation (CARF) which contributed over 200 survey results from cancer patients alongside their family members.

The platform uses both statistical analysis together with machine learning algorithms and visual presentation elements to detect behavioral patient patterns and assess website ease of use and multiple healthcare system components. Real-time monitoring of patient engagement and satisfaction is possible through custom dashboards offered by the solution to stakeholders.

Streamlit technology enables an accessible implementation of this platform thus expanding deployment possibilities to healthcare facilities researchers along with NGOs through a scalable platform.

1.1 Objectives

A fundamental purpose of the Healthcare Data Analytics Platform project focuses on the following goals:

- Building a reliable data analytics system with intuitive design will allow organizations to make decision support through patient and caregiver feedback analysis.

- A comprehensive exploratory data analysis (EDA) should be performed on survey responses from cancer patients as well as their families to determine platform usage engagement rates and accessibility problems and healthcare platform satisfaction levels.

- The construction of predictive machine learning models based on Linear Regression and Logistic Regression and Multiple Regression will analyze factors affecting user experience and feature adoption and healthcare-related participation elements.

- Real-time interactive visual patient data intelligence dashboards can be designed using Power BI with Streamlit for deployment purposes.

- The team will help CARF along with other NGOs achieve enhanced usability in their digital healthcare solutions through evidence-based recommendations.

- The platform should receive cloud deployment for scalability purposes to enable secure and convenient access of insights to healthcare administrators along with decision-makers.

1.2 Literature Survey

- The work by Raghupathi & Raghupathi (2014) shows big data healthcare usage along with analytics benefits that include enhanced clinical results and improved patient experiences as well as cost reduction.
- Obermeyer et al. (2016) established how predictive models enabled healthcare facilities to discover at-risk patients for effective resource distribution.
- Zhang & Walji (2011) examined healthcare website usability through user-oriented design to achieve better medical service satisfaction for patients.
- The article by Amato et al. (2013) examined machine learning applications in biomedical data analysis which formed the base for regression modeling in health data mining.
- According to Murdoch & Detsky (2013) artificial intelligence presents enormous diagnostic capabilities which shrink human mistakes and platform decisions.
- The paper by Chen et al. (2017) focuses on discussing cloud-based healthcare analytics alongside its benefits for both accessibility and collaborative data science applications and scalability features.
- The healthcare data analysis using real-time dashboards supported by Streamlit receives backing from Dinh et al. (2019).
- The adoption of health IT systems depends on trust and user interface and perceived usefulness according to Kankanhalli et al. (2016).
- Rajkomar et al. (2018) demonstrated through research that deep learning with electronic health records helps medical staff predict upcoming health occurrences.
- Wang & Hajli (2017) conducted research about social media platforms used for digital health engagement to describe how patients interact with healthcare technology.

- The data pre-processing technique for healthcare which handles missing values and its validation process was presented by Nguyen et al. (2014) in your research project.
- Dash et al. (2019) examined how predictive analytics trends affect public health while focusing on its functionality in both early intervention activities and patient participation initiatives.
- According to Topol (2019) the author demonstrates in Deep Medicine how AI together with data analytics helps healthcare professionals deliver personalized treatments and enhance medical results through digital infrastructure.
- The paper written by Jensen et al. in 2012 addressed electronic health record mining for patient-centered research together with ethical considerations about health data analytics.
- The work by Bates et al. (2014) studied clinical decision support systems together with analytics tools which boost patient care security and service quality.

CHAPTER 2

EXISTING AND PROPOSED SYSTEMS

2.1 Existing System

Healthcare data management systems which operate traditionally use Electronic Health Records (EHRs) together with manual reporting and static data storage options. Widespread usage of these systems causes multiple restrictions to appear during analytic procedures as well as when engaging users and making real-time decisions.

Limited Analytical Capabilities

Medical facilities primarily collect patient data using existing storage platforms yet they do not have advanced analysis tools. Analytics work requires manual derivation due to time-consumption and efficiency problems that impact large datasets processing.

Poor User Engagement Metrics

User satisfaction ratings together with accessibility statistics and interaction data about platform use do not appear in the current digital healthcare tools. It becomes hard to determine effective system components from aspects that require improvement.

Lack of Predictive Modeling

Healthcare institutions lack the adoption of machine learning models for patient requirement determination as well as symptom pattern recognition and platform user behavior analysis. The lack of proactive healthcare services emerges due to this approach.

Fragmented Systems

Multiple healthcare systems with distinct functions maintain separate databases that prevent the creation of cohesive information relationships and understanding between elements.

Limited Visualization Tools

Basic and non-interactive display capabilities form the standard offering of most legacy system platforms. Decision-making platforms alongside real-time filtering and segmentation are non-existent for stakeholders to access through interactive dashboards.

No Cloud Integration

Due to their lack of cloud-based design older systems create problems related to scalability and maintain data silos. Through cloud platforms users attain secure system entry and they receive features to scale their operations while integrating their application with contemporary tools.

Security and Compliance Challenges

Healthcare organizations maintain ample exposure to security risks because they lack centralized control and continuous compliance assessments their current systems do not meet modern healthcare requirements such as HIPAA and GDPR.

2.2 Proposed System

The platform delivers numerous patient benefits through its special design for improved access and enhanced healthcare professional analytics while improving healthcare data access.

Comprehensive Data Analytics

Users can perform deep exploratory data analysis (EDA) through the platform which analyzes trends involving user satisfaction with the system and user performance statistics. The system delivers data analytics through both statistical reports and graphical breakdowns of important metrics which includes symptom checker analytics alongside language selection analytics along with feature popularity analytics.

Predictive Modeling

The platform employs Linear Regression in combination with Logistic Regression and Multiple Regression algorithms to accomplish its functions through machine learning.

- Predict user satisfaction scores.
- Estimate navigation efficiency.
- The research examines external variables that impact both patient health-seeking actions and attendance in events.

The implemented models provide organizations with predictive abilities to take proactive decisions while planning improvements to the platform.

Interactive Visual Dashboards

- Real-time interactive dashboards are generated through the combination of Power BI alongside Streamlit. These visualizations allow stakeholders to:
- Users can sort information according to age group or placement data or linguistic elements.
- Track trends in patient feedback.

Businesses assess how well their new product elements and graphical user interface changes perform.

Cloud-Based Deployment

- Deployment of the solution on Streamlit Cloud facilitates the following functions:
- Accessibility from any device.
- The system adjusts its capacity automatically based on how much user demand exists at present.
- Additionally, the system provides continuous data updating with instant refresh capabilities.

User-Centered Design

Developers built the system using genuine survey information which CARF NGO obtained from both cancer patients and caregivers during their research. Real user needs drive the design of the analytics platform so it becomes usable and focused on patients.

Security and Scalability

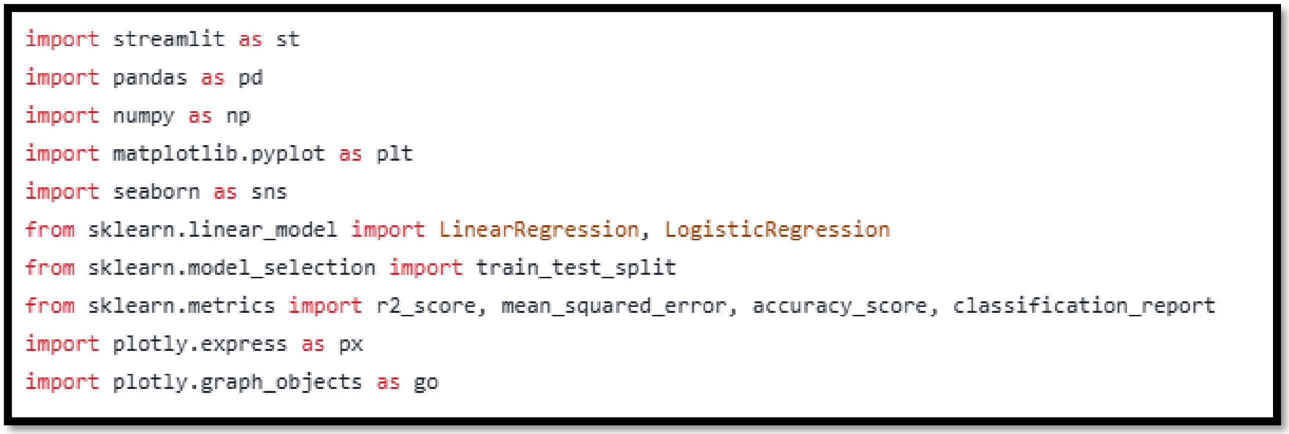
The platform utilizes Firebase together with MongoDB Atlas and FastAPI to enable safe data storage and quick API command execution and seamless growth capacity for additional users and platform features.

Multi-Device Responsiveness

The platform presents an enhanced user experience on all devices and screen sizes through its optimization process for mobile and desktop platforms.

CHAPTER 3
PROJECT IMPLEMENTATION

3.1 Importing Required Libraries

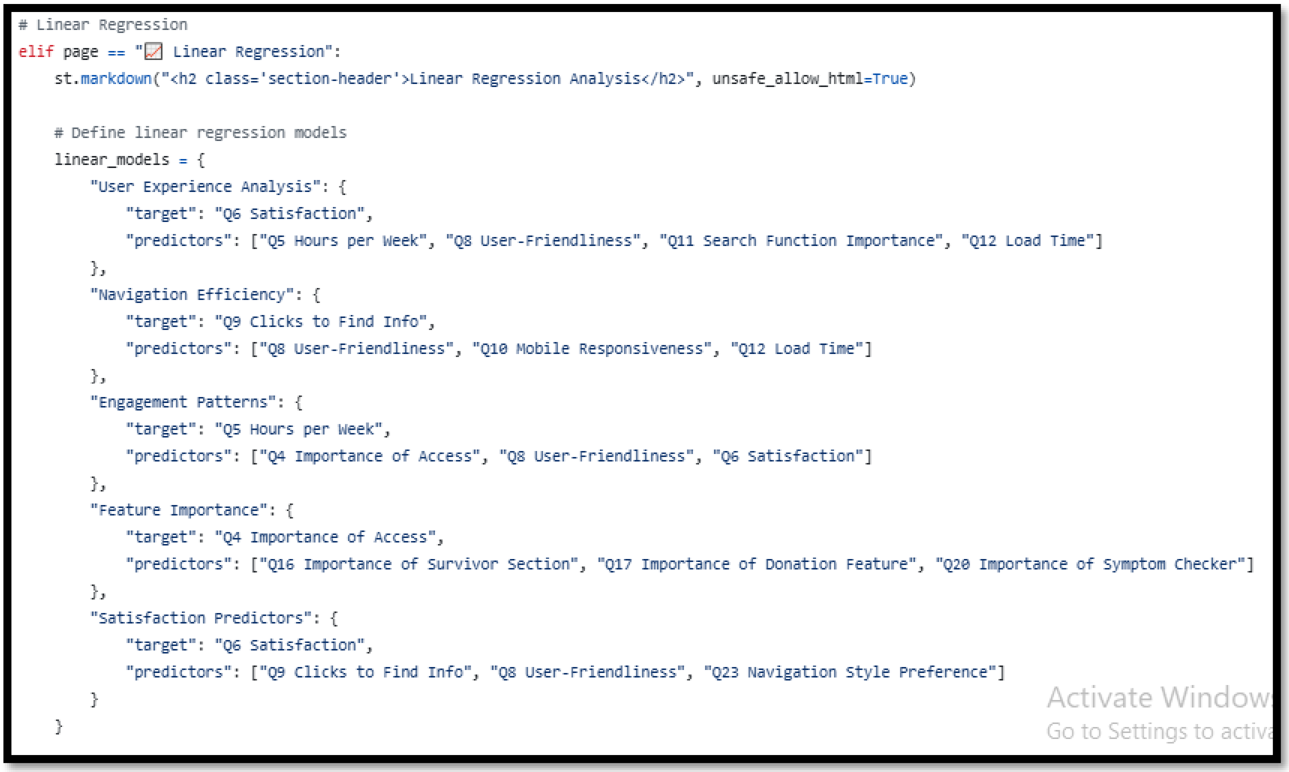


```python
import streamlit as st
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_squared_error, accuracy_score, classification_report
import plotly.express as px
import plotly.graph_objects as go
```

Our project implementation starts with importing Python libraries that provide data processing capabilities and visualization and machine learning functions. The data handling and manipulation through pandas and numpy become efficient while visualization tools come from matplotlib and seaborn and plotly. The scikit-learn library executes linear and logistic regression models as part of the application. The creation of an interactive web-based user interface uses streamlit which enables users to access data and view results in real time through its interface. The data analytics platform gets its foundation from these various libraries that work together.

3.2 Linear Regression

This supervised machine learning algorithm applies Linear Regression to develop models that connect between one or multiple independent variables and a target dependent variable. The regression line being the best-data-driven straight projection serves as a model to forecast target values from input features.

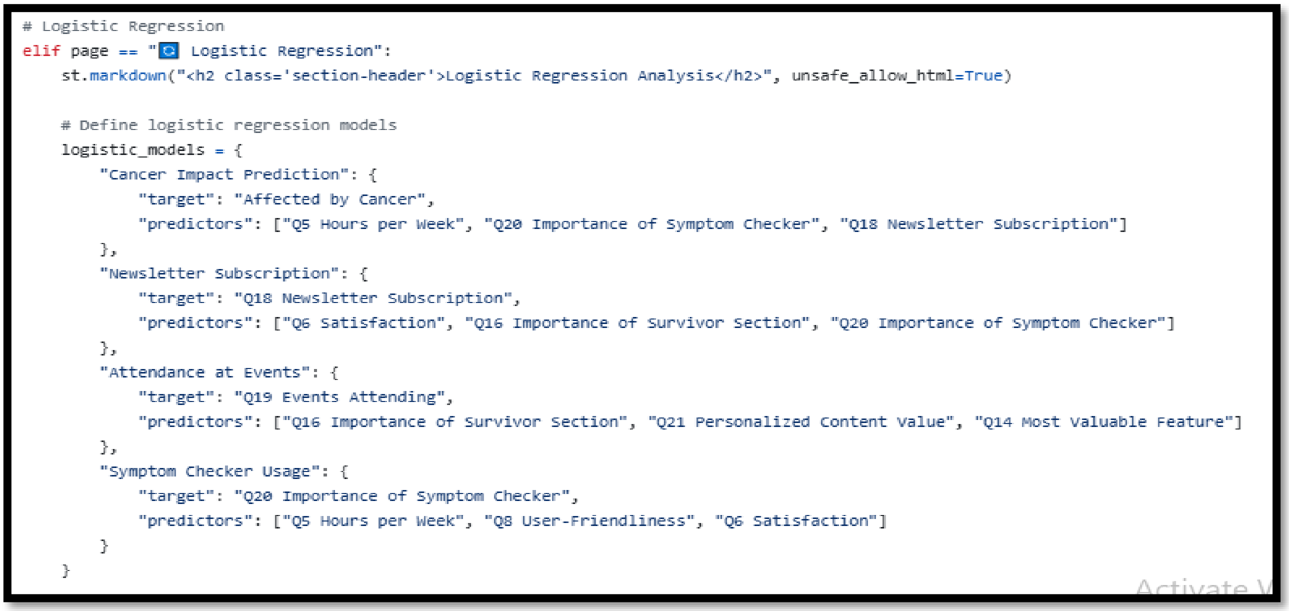


```
# Linear Regression
elif page == "📈 Linear Regression":
    st.markdown("<h2 class='section-header'>Linear Regression Analysis</h2>", unsafe_allow_html=True)

    # Define linear regression models
    linear_models = {
        "User Experience Analysis": {
            "target": "Q6 Satisfaction",
            "predictors": ["Q5 Hours per Week", "Q8 User-Friendliness", "Q11 Search Function Importance", "Q12 Load Time"]
        },
        "Navigation Efficiency": {
            "target": "Q9 Clicks to Find Info",
            "predictors": ["Q8 User-Friendliness", "Q10 Mobile Responsiveness", "Q12 Load Time"]
        },
        "Engagement Patterns": {
            "target": "Q5 Hours per Week",
            "predictors": ["Q4 Importance of Access", "Q8 User-Friendliness", "Q6 Satisfaction"]
        },
        "Feature Importance": {
            "target": "Q4 Importance of Access",
            "predictors": ["Q16 Importance of Survivor Section", "Q17 Importance of Donation Feature", "Q20 Importance of Symptom Checker"]
        },
        "Satisfaction Predictors": {
            "target": "Q6 Satisfaction",
            "predictors": ["Q9 Clicks to Find Info", "Q8 User-Friendliness", "Q23 Navigation Style Preference"]
        }
    }
```

3.3 Logistic Regression

Logistic Regression works as a supervised machine learning model to detect categories either in binary or multiple classifications. The prediction within logistic regression produces probabilities for categorical outcomes while linear regression generates continuous predictions.

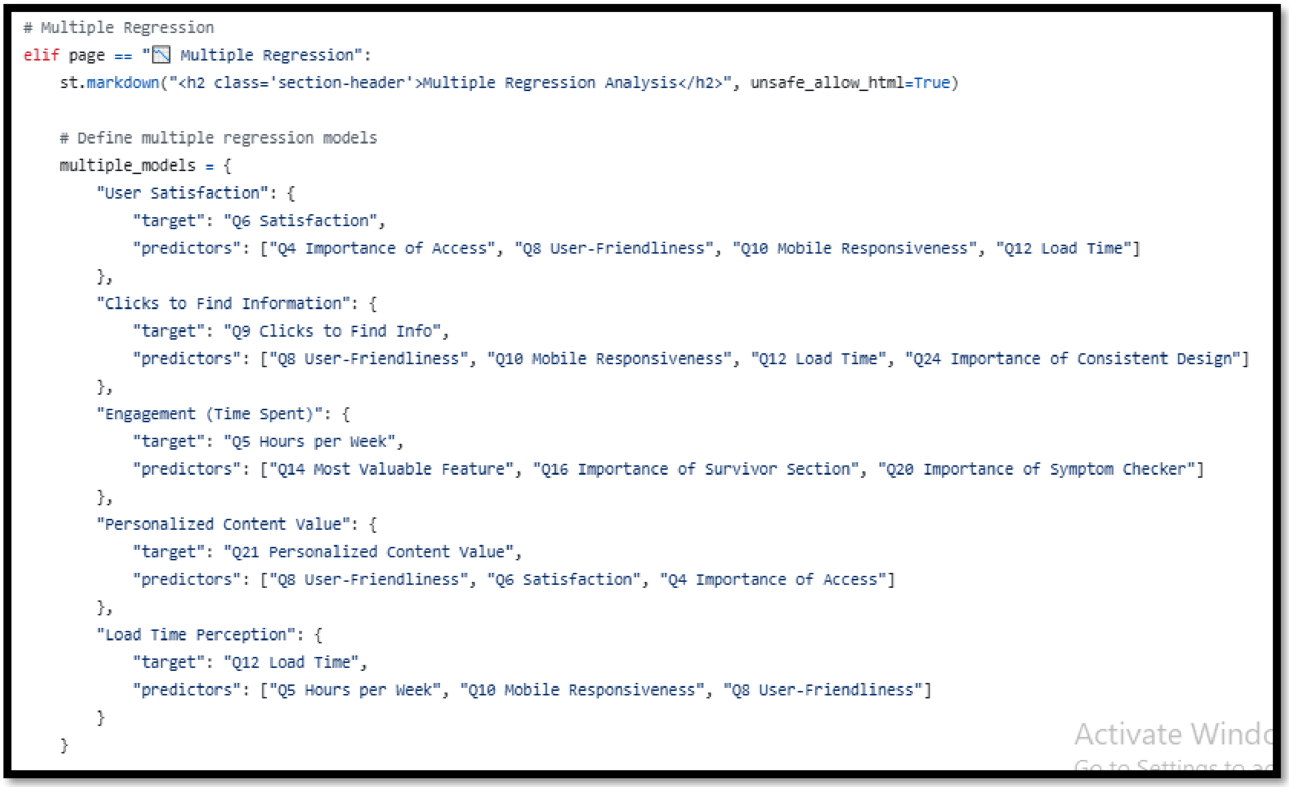


```
# Logistic Regression
elif page == "🔄 Logistic Regression":
    st.markdown("<h2 class='section-header'>Logistic Regression Analysis</h2>", unsafe_allow_html=True)

    # Define logistic regression models
    logistic_models = {
        "Cancer Impact Prediction": {
            "target": "Affected by Cancer",
            "predictors": ["Q5 Hours per Week", "Q20 Importance of Symptom Checker", "Q18 Newsletter Subscription"]
        },
        "Newsletter Subscription": {
            "target": "Q18 Newsletter Subscription",
            "predictors": ["Q6 Satisfaction", "Q16 Importance of Survivor Section", "Q20 Importance of Symptom Checker"]
        },
        "Attendance at Events": {
            "target": "Q19 Events Attending",
            "predictors": ["Q16 Importance of Survivor Section", "Q21 Personalized Content Value", "Q14 Most Valuable Feature"]
        },
        "Symptom Checker Usage": {
            "target": "Q20 Importance of Symptom Checker",
            "predictors": ["Q5 Hours per Week", "Q8 User-Friendliness", "Q6 Satisfaction"]
        }
    }
```

3.4 Multiple Regression

Linear regression extends into Multiple Regression as a method to forecast dependent variable values through multiple independent variables. The technique provides clarity regarding the connection between several predictors and one target variable.

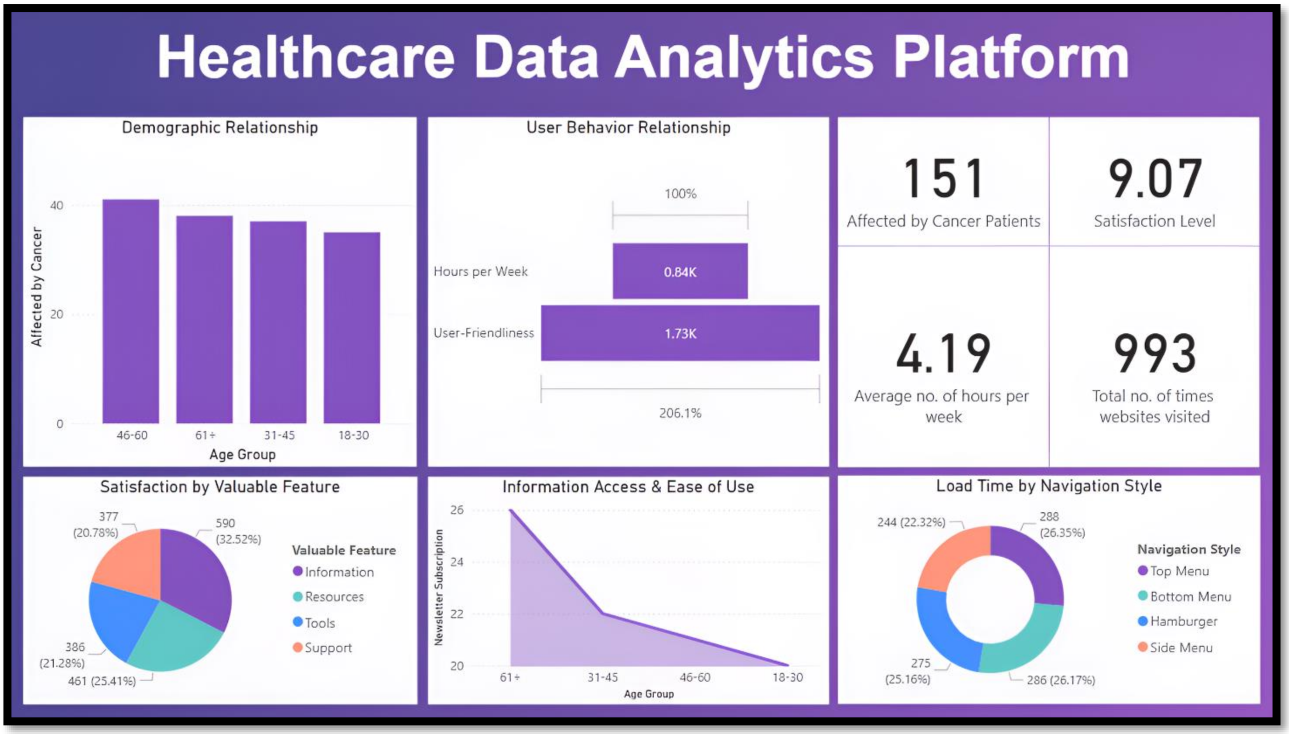


```
# Multiple Regression
elif page == "📉 Multiple Regression":
    st.markdown("<h2 class='section-header'>Multiple Regression Analysis</h2>", unsafe_allow_html=True)

    # Define multiple regression models
    multiple_models = {
        "User Satisfaction": {
            "target": "Q6 Satisfaction",
            "predictors": ["Q4 Importance of Access", "Q8 User-Friendliness", "Q10 Mobile Responsiveness", "Q12 Load Time"]
        },
        "Clicks to Find Information": {
            "target": "Q9 Clicks to Find Info",
            "predictors": ["Q8 User-Friendliness", "Q10 Mobile Responsiveness", "Q12 Load Time", "Q24 Importance of Consistent Design"]
        },
        "Engagement (Time Spent)": {
            "target": "Q5 Hours per Week",
            "predictors": ["Q14 Most Valuable Feature", "Q16 Importance of Survivor Section", "Q20 Importance of Symptom Checker"]
        },
        "Personalized Content Value": {
            "target": "Q21 Personalized Content Value",
            "predictors": ["Q8 User-Friendliness", "Q6 Satisfaction", "Q4 Importance of Access"]
        },
        "Load Time Perception": {
            "target": "Q12 Load Time",
            "predictors": ["Q5 Hours per Week", "Q10 Mobile Responsiveness", "Q8 User-Friendliness"]
        }
    }
```

CHAPTER 4

RESULTS AND DISCUSSIONS



User behavior data and platform metrics appeared on the newly developed dynamic Power BI dashboard that also recorded user satisfaction levels. The analysis yielded several essential insights that patrons can utilize for their business strategy.

Demographic Relationship

Most users who experience cancer belong to age segments between 46–60 years and 61+ years with a similar number of individuals in the 31–45 age bracket. Support features together with content delivery should be optimized specifically for middle-aged and senior audiences who face the biggest challenges from cancer.

User Behavior Relationship

- The user-friendliness feature demonstrated distinct value for user engagement since users interacted with the system 1.73K times which enhances user retention.
- The meaningful relationship between Hours per Week of usage demonstrates that users who find the site easy to use will explore the platform more extensively.

Key Metrics

- 151 respondents were affected by cancer.
- A total of 163 participants answered with 9.07 out of 10 being the typical satisfaction rate.
- Platinum users spend about 4.19 hours weekly on the system.
- The website shows strong usage metrics as users have accessed it 993 times.

Satisfaction by Valuable Feature

Information takes the lead as the most appreciated feature at 32.52% since users value this aspect highest among Support (25.41%), Resources (21.29%), and Tools (20.78%). The research results highlight why user training materials must focus on delivering helpful information in addition to offering support features.

Information Access & Ease of Use

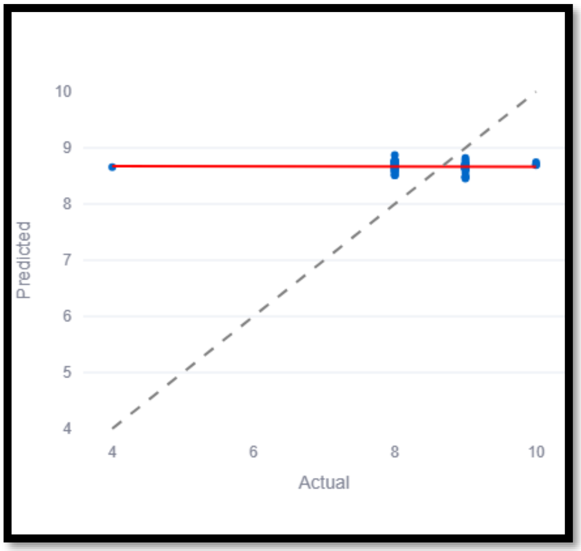
- Newsletter subscriptions decrease with age:
- Among all groups the members in the 61+ demographic show the most active response to newsletter content.
- Followed by 31–45, 46–60, and the 18–30 group. The survey results indicate that elderly users show preference for regular update services but younger people prefer direct access through alternative options.

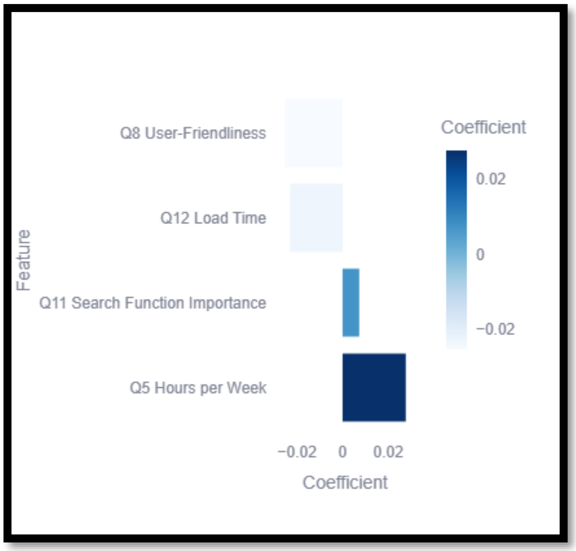
Load Time by Navigation Style

Load times reached their peak with Top Menu (26.35%) and Side Menu (26.17%).

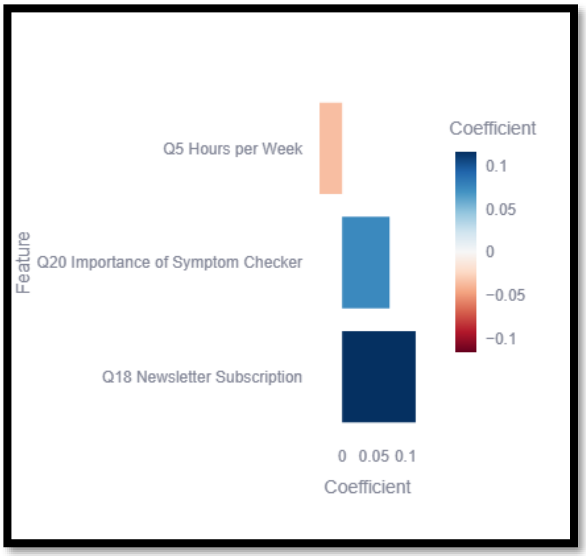
Among the explored navigation styles the Bottom Menu matched the competence of Hamburger Menu (22.32%) while Top Menu (26.35%) and Side Menu (26.17%) occupied the top performance slots. The evaluation supports designers in maximizing the UI/UX design through both user feedback and system efficiency measurement.

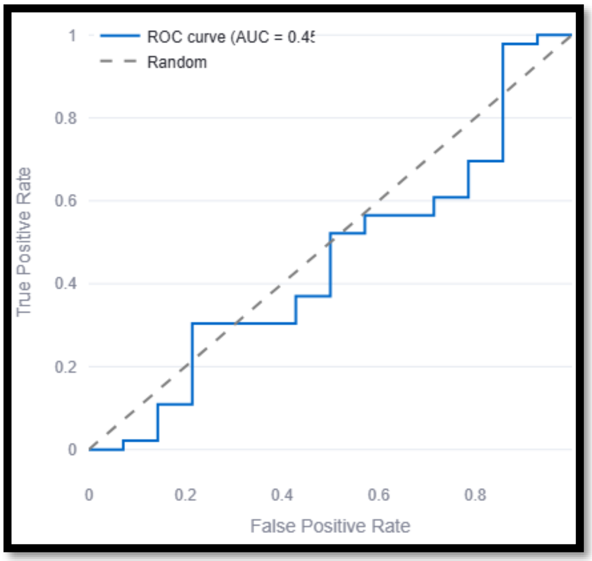
Linear Regression Analysis

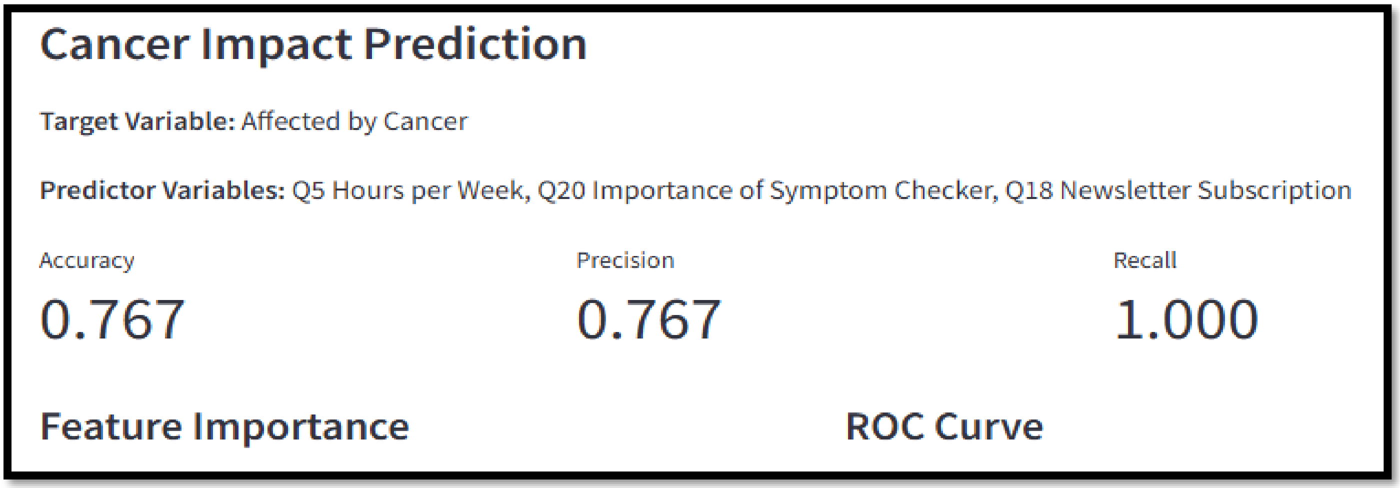




Logistic Regression







Cancer Impact Prediction

Target Variable: Affected by Cancer

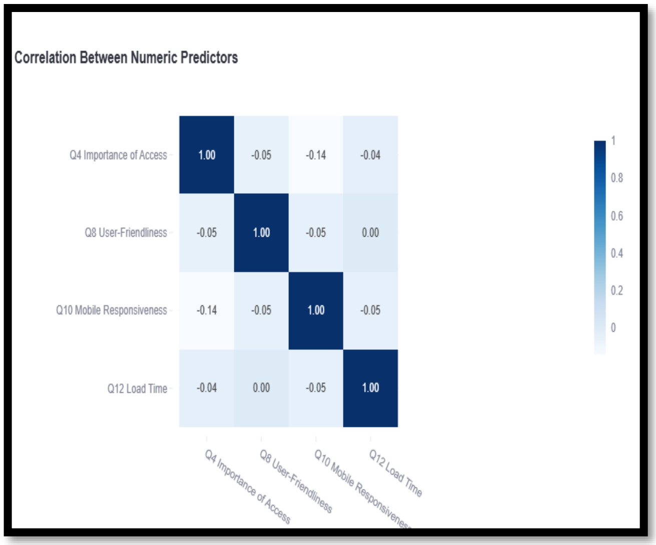
Predictor Variables: Q5 Hours per Week, Q20 Importance of Symptom Checker, Q18 Newsletter Subscription

Accuracy	Precision	Recall
0.767	0.767	1.000

Feature Importance

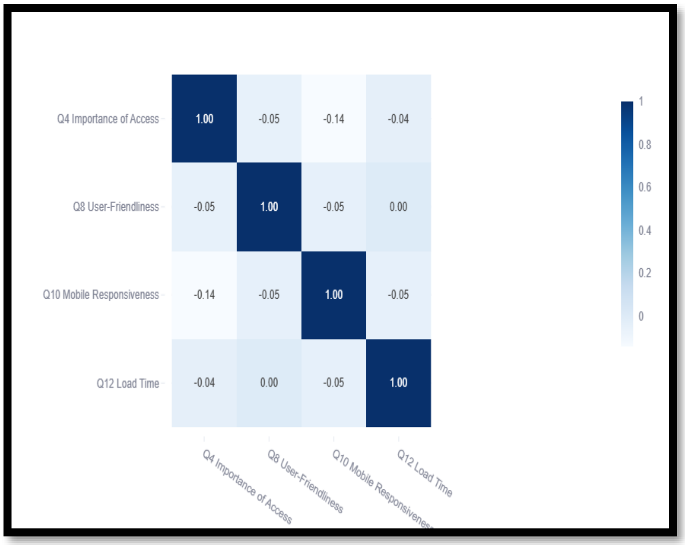
ROC Curve

Multiple Regression



Correlation Between Numeric Predictors

	Q4 Importance of Access	Q8 User-Friendliness	Q10 Mobile Responsiveness	Q12 Load Time
Q4 Importance of Access	1.00	-0.05	-0.14	-0.04
Q8 User-Friendliness	-0.05	1.00	-0.05	0.00
Q10 Mobile Responsiveness	-0.14	-0.05	1.00	-0.05
Q12 Load Time	-0.04	0.00	-0.05	1.00



	Q4 Importance of Access	Q8 User-Friendliness	Q10 Mobile Responsiveness	Q12 Load Time
Q4 Importance of Access	1.00	-0.05	-0.14	-0.04
Q8 User-Friendliness	-0.05	1.00	-0.05	0.00
Q10 Mobile Responsiveness	-0.14	-0.05	1.00	-0.05
Q12 Load Time	-0.04	0.00	-0.05	1.00

CHAPTER 5

CONCLUSION AND FUTURE WORK

The platform reveals important trends about user behavior and demographic characteristics and satisfaction patterns when it uses in-depth data exploration in combination with linear and logistic as well as multiple regression analysis and interactive visualization tools.

The project revealed significant findings that consisted of:

- The outreach and support programs for cancer patients should prioritize **middle-aged and senior citizens** because they face higher cancer risks.
- **Desirable user experiences** together with available information play strong roles when users rate their satisfaction levels. They also determine users' decisions to stay involved with a platform.
- The **user experience and behavior patterns** of the platform's visitors are strongly affected by the Symptom Checker feature together with Survivor Stories and Personalized Content options.
- Users placed the highest value on information content due to its **precise medical data accuracy and ease of access.**
- **Interaction through Power BI dashboards** allowed stakeholders to view insights in clear intuitive ways thus supporting their decision-making process.
- The platform functions as both an **analytical tool** and a lead service for digital health improvement which guarantess user need alignment.

Future Work

The project has established a solid base yet various upgrade possibilities exist for further growth:

- The system should automatically import **real-time data** obtained from hospital records and health monitoring apps and online consultations for maintaining accurate insights that provide stronger value.

- The number of patient satisfaction insights increases when **Natural Language Processing (NLP)** extracts emotional sentiment from feedback through Sentiment Analysis.

- The platform should get a **mobile-friendly application** to expand accessibility since users prefer mobile use.

- The project should expand its survey base to various **geographic locations and population demographics** for generating more comprehensive discovery about the situation.

- A/B tests will evaluate different user interface navigation styles and theme variations and layout modifications for ongoing **user interface enhancement**.

These future implementations will enable the platform development toward becoming an intelligent healthcare system to deliver customized scalable digital care solutions for healthcare organizations.