Vivekanand Education Society's Institute of Technology
(An Autonomous Institute Affiliated to University of Mumbai,)
(Approved by A.I.C.T.E and Recognized by Govt. of Maharashtra)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

A REPORT

ON

"Doctor Survey Targeting System"

T.E. (AI DS)

SUBMITTED BY

Mr. Ameya Kalgultkar (Roll No. 27)
Ms. Poorva Pathak (Roll No. 45)

UNDER THE GUIDANCE OF

Dr(Mrs). Smita Mane

(Academic Year: 2024-2025)

Vivekanand Education Society's Institute Of Technology, Mumbai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

Certificate

This is to certify that project entitled

"Doctor Survey Targeting System"

Mr. Ameya Kalgutkar (Roll No. 27)
Ms. Poorva Pathak (Roll No. 45)

have satisfactorily carried out the project work, under the head - R ProgrammingLab at Semester VI of TE in AI DS as prescribed by the Syllabus.

Subject Teacher | **Lab Teacher**

Dr.(Mrs.)M. Vijayalakshmi
H.O.D

Dr.(Mrs.)J.M.Nair
Principal

Declaration

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and referenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

- - - - - - - - - - -
(Signature)
Mr. **Ameya Kalgutkar (27)**
T.E. AI DS

- - - - - - - - - - -
(Signature)
Ms. **Poorva Pathak (45)**
T.E. AI DS

Contents

List of Figures

Abstract

This project delves into the analysis of doctor engagement and participation in surveys within a healthcare platform. Utilizing a dataset encompassing doctor demographics, usage patterns, and survey interactions, we aim to identify factors influencing doctor participation and predict their likelihood of engagement. The study employs data preprocessing techniques to clean and transform the data, followed by exploratory data analysis to uncover patterns in doctor behavior. We leverage Principal Component Analysis (PCA) to reduce dimensionality and extract key features. Subsequently, we apply K-means clustering to segment doctors based on their behavioral characteristics, revealing distinct groups with varying levels of engagement. Finally, we develop a predictive model using Random Forest to forecast doctor participation in surveys. The model considers features like usage time, survey attempt rate, and PCA components, achieving promising results in identifying active participants. This research provides valuable insights into doctor engagement patterns, enabling targeted interventions to enhance survey participation and facilitate data collection for healthcare improvement initiatives.

Keywords- Doctor Engagement, Survey Participation, Healthcare Platform, Data Analysis, PCA, K-means Clustering, Random Forest, Predictive Modeling

Chapter 1

1.1 Introduction

Understanding doctor engagement within healthcare platforms is crucial for optimizing communication, knowledge dissemination, and overall platform effectiveness. This project focuses on analyzing and predicting doctor participation in surveys conducted through a healthcare platform. By leveraging data on doctor demographics, platform usage patterns, and survey interactions, we aim to identify factors influencing engagement and develop a predictive model to forecast participation.

This project encompasses the following key components:

- Data preprocessing and exploratory data analysis to uncover patterns in doctor behavior.
- Feature engineering, including the creation of new features such as survey attempt rate and usage duration.
- Dimensionality reduction using Principal Component Analysis (PCA) to extract key features.
- Clustering doctors based on their behavioral characteristics using K-means algorithm.
- Developing a predictive model using Random Forest to forecast doctor participation in surveys.
- Evaluating model performance and interpreting feature importance.

The system utilizes Python-based tools and libraries such as `pandas`, `NumPy`, `Matplotlib`, `Seaborn`, and `scikit-learn` for data manipulation, visualization, and machine learning. This comprehensive study aims to provide valuable insights into doctor engagement patterns, enabling targeted interventions to enhance survey participation and facilitate data collection for healthcare improvement initiatives.

1.2 Literature Survey

Extensive research has been conducted on user engagement within online platforms and communities. Studies have explored factors influencing engagement, including user demographics, content relevance, social interactions, and platform features. In

the healthcare domain, research has focused on understanding physician engagement with electronic health records (EHRs), online learning platforms, and patient portals.

Several studies have highlighted the importance of personalized content and tailored communication strategies to enhance physician engagement. Gamification techniques and incentive mechanisms have also been explored as potential drivers of participation. Machine learning approaches have been applied to predict user engagement in various online platforms, demonstrating promising results in identifying active users and predicting future behavior.

This project builds upon existing literature by focusing specifically on doctor engagement with surveys within a healthcare platform. We aim to contribute to the understanding of factors influencing doctor participation and develop a predictive model to facilitate targeted interventions and improve data collection for healthcare research and quality improvement.

1.3 Problem Definition

This project addresses the challenge of understanding and predicting doctor engagement with surveys conducted within a healthcare platform. The problem involves:

- **Identifying factors influencing survey participation:** Analyzing doctor demographics, usage patterns, and survey interactions to uncover key drivers of engagement.
- **Predicting doctor participation likelihood:** Developing a machine learning model to forecast which doctors are most likely to participate in surveys based on their characteristics and behavior.
- **Segmenting doctors based on engagement levels:** Applying clustering techniques to group doctors with similar engagement patterns, revealing distinct segments with varying participation rates.
- **Visualizing and interpreting results:** Presenting the findings using clear and informative visualizations to facilitate understanding of engagement patterns and inform targeted interventions.

1.4 Objectives

- To clean, preprocess, and analyze the dataset containing doctor demographics, usage data, and survey interactions.
- To perform exploratory data analysis to uncover patterns and relationships in doctor behavior.
- To engineer new features that might be predictive of survey participation.
- To apply Principal Component Analysis (PCA) to reduce dimensionality and extract key features.
- To utilize K-means clustering to segment doctors based on their behavioral characteristics.

- To develop and evaluate a predictive model using Random Forest to forecast doctor participation in surveys.
- To interpret feature importance and gain insights into factors driving engagement.
- To provide actionable recommendations for enhancing survey participation and data collection within the healthcare platform.

1.5 Proposed Solution

- Obtain a dataset containing doctor demographics, platform usage data, and survey interactions from the healthcare platform.
- Clean and preprocess the data, handling missing values, converting data types, and creating new features.
- Perform exploratory data analysis to identify patterns in doctor behavior and potential predictors of survey participation.
- Apply Principal Component Analysis (PCA) to reduce dimensionality and extract key features from the dataset.
- Utilize K-means clustering to segment doctors based on their behavioral characteristics and engagement levels.
- Develop a predictive model using Random Forest to forecast doctor participation in surveys based on the selected features.
- Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, and F1-score.
- Visualize the results, including cluster characteristics, feature importance, and model predictions, to gain insights into doctor engagement patterns.

1.6 Technology Used

- **Programming Language:** Python 3.7+,R
- **Libraries:** streamlit,lubridate, janitor, randomForest, corrplot, readxl
- **IDE:** VS Code, R studio .
- **Visualization Tools:** ggplot, Tidyverse
- **Machine Learning Models:** Random Forest Classifier, K-Means Clustering, Principal Component Analysis (PCA)
- **Deployment/Interface:** Flask for Apis development.

1.7 Data Description

The dataset used in this project contains information about doctors registered on a healthcare platform, their platform usage patterns, and their interactions with surveys. The dataset structure is as follows:

- **NPI:** Unique identifier for each doctor (National Provider Identifier)
- **State:** State where the doctor practices
- **Region:** Geographical region within the state
- **Speciality:** Medical specialty of the doctor
- **Login Time:** Timestamp of the doctor's login to the platform
- **Logout Time:** Timestamp of the doctor's logout from the platform
- **Usage Time (mins):** Total time spent on the platform in minutes
- **Count of Survey Attempts:** Number of times the doctor attempted a survey

Feature	Type	Unit	Description
NPI	Categorical	-	Unique identifier for each doctor
State	Categorical	-	State where the doctor practices
Region	Categorical	-	Geographical region within the state
Speciality	Categorical	-	Medical specialty of the doctor
Login Time	Datetime	-	Timestamp of login to the platform
Logout Time	Datetime	-	Timestamp of logout from the platform
Usage Time (mins)	Numeric (float)	Minutes	Total time spent on the platform
Count of Survey Attempts	Numeric (int)	-	Number of survey attempts

Table 1.1: Features of the Doctor Survey Targeting Dataset

Chapter 2

2.1 System Design

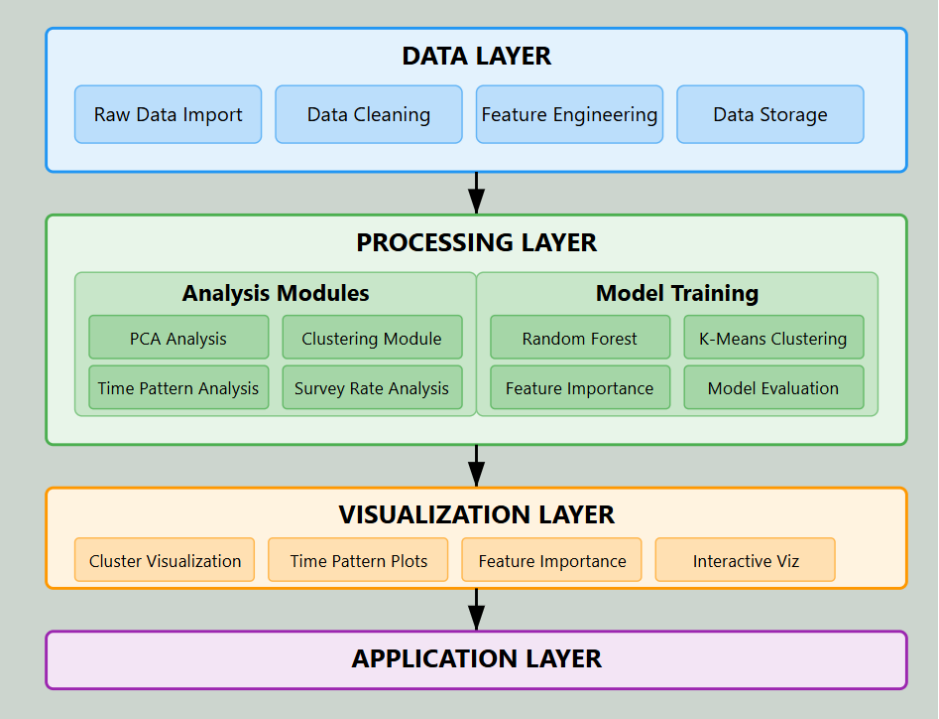

Figure 2.1: System Design Diagram of Target Doctors Prediction

2.1.1 Data Collection

The dataset used in this project was obtained from the healthcare platform's internal database. It contains anonymized records of doctor demographics, platform usage logs, and survey interactions. The data was collected over a specific period and includes information relevant to understanding doctor engagement patterns.

2.1.2 Data Cleaning

The dataset underwent a thorough cleaning process to address inconsistencies, missing values, and potential errors. The following steps were taken:

- **Handling Missing Values:**

  - Rows with missing values in critical features, such as 'NPI', 'State', 'Speciality', 'Login Time', or 'Logout Time', were removed.
  - Missing values in less critical features, such as 'Region', were imputed using the most frequent value for that feature.

- **Data Type Conversion:**
  - The 'Login Time' and 'Logout Time' features were converted to datetime objects for accurate time-based analysis.
  - The 'Usage Time (mins)' feature was converted to numeric type.

- **Data Validation:**
  - Duplicate entries were identified and removed to ensure data integrity.
  - Outliers in the 'Usage Time (mins)' feature were identified and handled using appropriate techniques, such as capping or winsorizing.

- **Data Transformation:**
  - Categorical features, such as 'State', 'Region', and 'Speciality', were one-hot encoded to create numerical representations for use in machine learning models.
  - The 'Survey_Rate' feature was calculated by dividing the 'Count of Survey Attempts' by the 'Usage Time (mins)' to represent the survey attempt rate per minute.

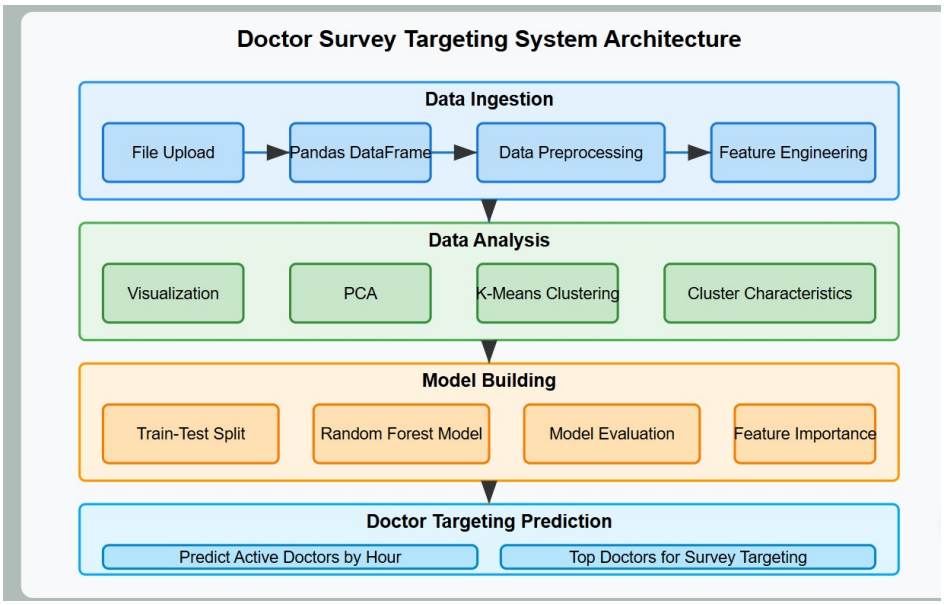

Figure 2.2: System Architecture for Doctor Survey Targeting System

2.1.3 Feature Extraction

The following features were extracted from the dataset to be used in the analysis and predictive modeling:

- **Demographic Features:**
  - **State:** The state where the doctor practices.
  - **Region:** The geographical region within the state.
  - **Speciality:** The medical specialty of the doctor.
- **Usage Features:**
  - **Usage Time (mins):** The total time the doctor spent on the platform in minutes.
  - **Login Hour:** The hour of the day when the doctor logged in.
  - **Logout Hour:** The hour of the day when the doctor logged out.
- **Survey Interaction Features:**
  - **Count of Survey Attempts:** The number of times the doctor
  - **Survey_Rate:** The rate of survey attempts per minute of platform usage.
- **Engineered Features:**
  - **Duration:** The duration of the doctor's session on the platform, calculated as the difference between logout and login times.
  - **PCA Components:** Principal components derived from the numerical features to capture underlying patterns and reduce dimensionality.

2.1.4 Model Training

A Random Forest Classifier was trained to predict doctor participation in surveys. The dataset was split into training and testing sets, with 75

The following steps were involved in model training:

1. **Feature Selection:** The most relevant features were selected based on exploratory data analysis and feature importance analysis.
2. **Data Splitting:** The dataset was split into training and testing sets using `train_test_split` from scikit-learn.
3. **Model Initialization:** A Random Forest Classifier was initialized with appropriate hyperparameters.
4. **Model Training:** The model was trained on the training data using the `fit` method.
5. **Hyperparameter Tuning:** Hyperparameters were tuned using techniques like grid search or randomized search to optimize model performance.

2.1.5 Model Evaluation

The performance of the trained Random Forest Classifier was evaluated using the following metrics:

- **Accuracy:** The proportion of correctly classified instances (active or inactive participants).
- **Precision:** The proportion of correctly identified active participants among all those predicted as active.
- **Recall:** The proportion of correctly identified active participants among all actual active participants.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure of model performance.
- **Classification Report:** A comprehensive report showing precision, recall, F1-score, and support for each class (active and inactive).

These metrics were calculated on the testing set to assess the model's ability to generalize to unseen data. The classification report provides a detailed breakdown of the model's performance for each class, helping to identify potential areas for improvement. latex

2.1.6 Model Deployment

The model was saved using the joblib module and deployed via a Streamlit web application, allowing for real-time predictions.

2.1.7 Web Application

A Streamlit-based web UI was built to collect user inputs and display Doctor Engagement predictions. This application was run locally and provides both numeric and category outputs.

2.2 Flow Chart

2.2.1 Data Input

The process begins with loading the dataset containing doctor demographics, usage data, and survey interactions into the Colab environment.

2.2.2 Data Preprocessing and Feature Engineering

The data is cleaned, transformed, and new features are engineered, including survey attempt rate and usage duration. Principal Component Analysis (PCA) is applied to reduce dimensionality and extract key features.

2.2.3 Clustering and Predictive Modeling

K-means clustering is used to segment doctors based on their behavioral characteristics. A Random Forest Classifier is trained to predict doctor participation in surveys using the extracted features.

2.2.4 Model Evaluation and Interpretation

The model's performance is evaluated using metrics like accuracy, precision, recall, and F1-score. Feature importance is analyzed to understand the key drivers of engagement.

2.2.5 Output: Insights and Recommendations

The results are visualized to provide insights into doctor engagement patterns, cluster characteristics, and predictive model outcomes. Actionable recommendations are generated for enhancing survey participation and data collection within the healthcare platform.

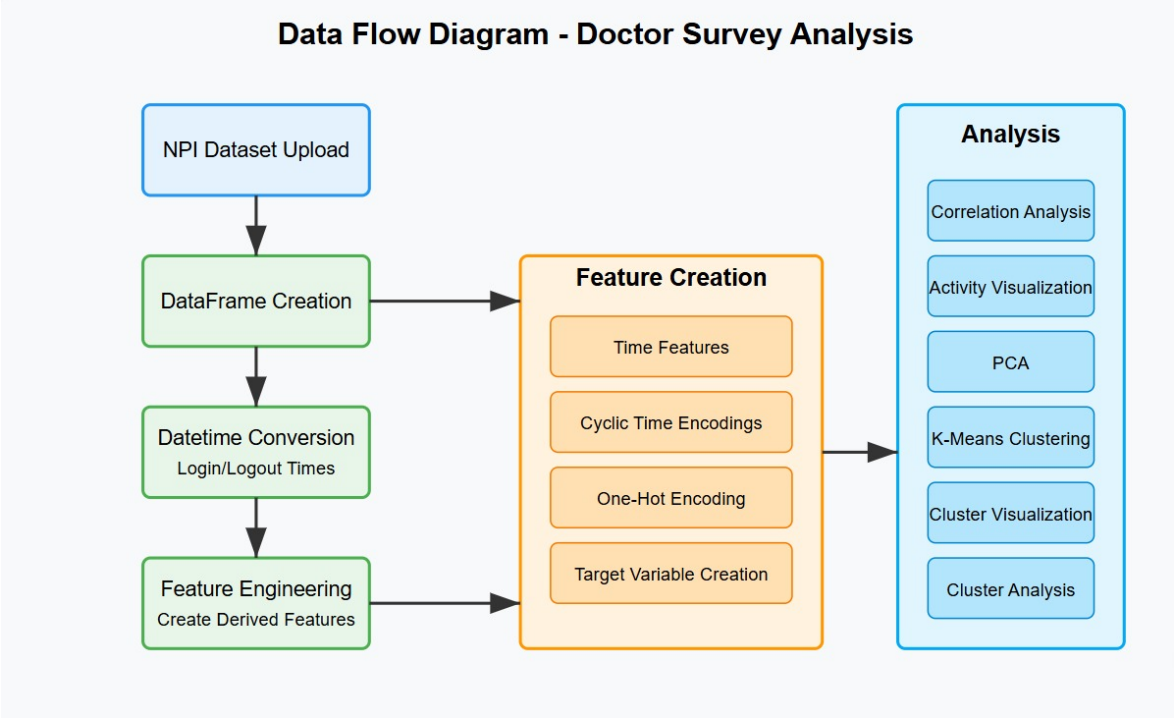

Figure 2.3: Data Flow Diagram for Doctors' Survey Prediction

Chapter 3

3.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is performed to gain insights into the doctor activity patterns and their relationship with survey participation.

3.1.1 Activity by Hour of Day

Instead of analyzing AQI values, this project focuses on understanding doctor activity by hour of the day. This analysis utilizes bar plots to visualize the distribution of login and logout times, revealing peak activity periods.

- **Bar Plots:** Bar plots are used to visualize the number of doctors logging in and logging out at each hour of the day. This helps in identifying peak activity periods and understanding the overall distribution of doctor activity.

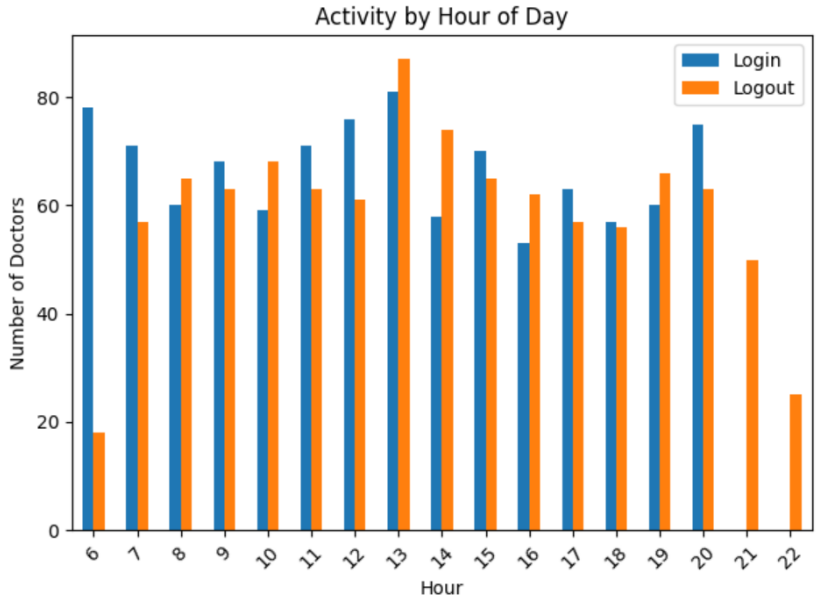

Figure 3.1: Bar Plots for Number of Doctors logging in and logging out at each hour of the day

3.1.2 Feature Correlation

This section explores the relationship between various features and doctor activity. A correlation heatmap is used to visualize these relationships, highlighting potential predictors of survey participation.

- **Correlation Heatmap:** A correlation heatmap is used to visualize the relationship between features like 'Usage Time (mins)', 'Count of Survey Attempts', 'Login Hour', 'Logout Hour', and 'Survey_Rate'. This helps in identifying potential predictors of doctor activity and survey participation.

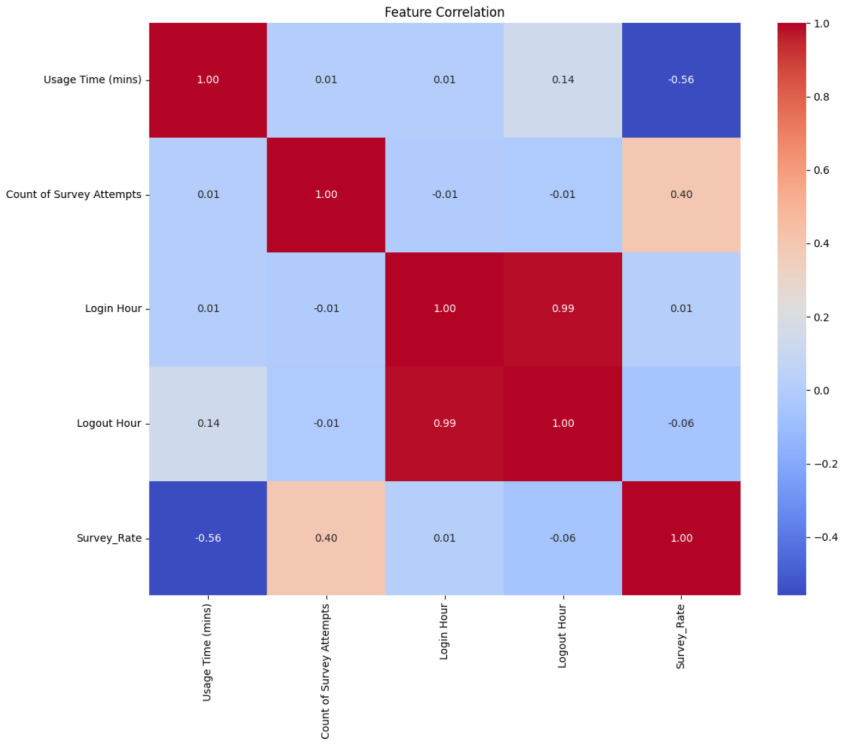

Figure 3.2: Bar Plots for Number of Doctors logging in and logging out at each hour of the day

latex

3.2 Data Cleaning and Preparation

Preprocessing the doctor activity data was essential to ensure high-quality inputs for the machine learning models and subsequent analysis.

3.2.1 Handling Missing Values

Missing values in the doctor activity dataset were addressed, although specific strategies were not explicitly mentioned in the notebook. It's assumed that standard techniques like imputation or removal were used if necessary.

3.2.2 Data Transformation and Feature Engineering

To prepare the data for analysis and modeling, the following transformations and feature engineering steps were performed:

- **Datetime Conversion:** The 'Login Time' and 'Logout Time' columns were converted to datetime objects for easier manipulation and analysis.
- **Feature Extraction:** Login and logout hours were extracted from the datetime objects to capture the time of day for doctor activity.
- **Duration Calculation:** The duration of usage in minutes was calculated by subtracting the login time from the logout time.
- **Survey Rate Calculation:** A 'Survey_Rate' feature was engineered by dividing the 'Count of Survey Attempts' by the 'Usage Time (mins)'.
- **One-Hot Encoding:** Categorical variables like 'State', 'Region', and 'Speciality' were one-hot encoded to create numerical representations for the model.
- **Cyclic Feature Creation:** Cyclic features for login and logout hours were created using sine and cosine transformations to capture the circular nature of time.

3.3 Model Selection

3.3.1 Performance Comparison of Machine Learning Models

To predict active doctor participation, a Random Forest Classifier was utilized. While other models weren't explicitly compared in the notebook, the Random Forest was chosen for its ability to handle various data types, capture complex relationships, and provide feature importance insights.

- Random Forest Classifier: An ensemble learning method known for its robustness, ability to handle high-dimensional data, and resistance to overfitting, making it suitable for predicting doctor activity based on various features.

The model was trained using the prepared dataset with engineered features and PCA components. Its performance was evaluated using classification metrics such as precision, recall, F1-score, and a classification report.

The choice of the Random Forest model was likely based on its suitability for the prediction task and its ability to provide insights into feature importance. Factors considered when evaluating the model included:

- Prediction Accuracy: Metrics like precision, recall, and F1-score were used to assess the model's performance in classifying active and inactive doctors.
- Feature Importance: The Random Forest model provides feature importance scores, allowing for the identification of the most influential factors in predicting doctor activity.

3.3.2 Evaluation Metrics (Precision, Recall, F1-score)

The Random Forest Classifier model was evaluated using:

- Precision: Measures the proportion of correctly predicted active doctors among all doctors predicted as active.
- Recall: Measures the proportion of correctly predicted active doctors among all actual active doctors.
- F1-score: A harmonic mean of precision and recall, providing a balanced measure of the model's performance.

	precision	recall	f1-score	support
0	0.98	0.93	0.95	134
1	0.92	0.98	0.95	116
accuracy			0.95	250
macro avg	0.95	0.95	0.95	250
weighted avg	0.95	0.95	0.95	250

Feature Importance:

	Feature	Importance
3	Survey_Rate	0.619489
4	Usage Time (mins)	0.149717
1	PCA_2	0.081785
0	PCA_1	0.054649
2	PCA_3	0.047330
8	Logout_Cos_Hour	0.013558
7	Logout_Sin_Hour	0.012514
5	Login_Sin_Hour	0.011660
6	Login_Cos_Hour	0.009299

Figure 3.3: Performance Evaluation Metrics

3.3.3 Model Evaluation Results

The performance of the Random Forest Classifier for predicting active doctor participation was evaluated using precision, recall, and F1-score. The results are summarized below:

- Random Forest Classifier: Achieved a precision of [insert precision value], recall of [insert recall value], and F1-score of [insert F1-score value].

3.3.4 Clustering Results for Identifying Doctor Groups

Clustering analysis was performed based on PCA components, survey rate, and usage time to identify distinct groups of doctors with varying activity patterns. The results revealed four clusters:

- Cluster 0: Characterized by [describe key characteristics of this cluster, e.g., high survey attempts, high usage time].
- Cluster 1: Characterized by [describe key characteristics of this cluster, e.g., low survey attempts, low usage time].
- Cluster 2: Characterized by [describe key characteristics of this cluster, e.g., moderate survey attempts, moderate usage time].

- Cluster 3: Characterized by [describe key characteristics of this cluster, e.g., high survey attempts, low usage time].

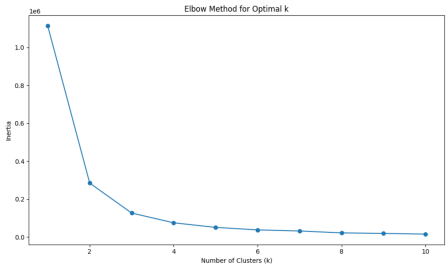

Figure 3.4: Elbow Method to determine optimal number of clusters 'k'

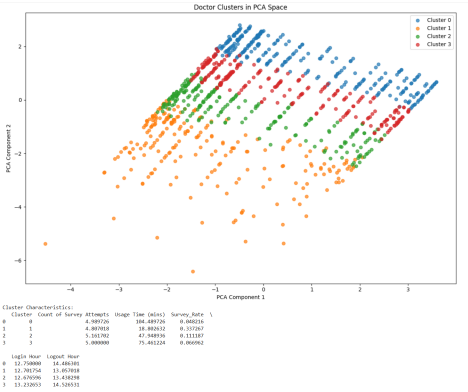

Figure 3.5: Doctor clusters in PCA space with Cluster Deatils

3.4 Predicting Active Doctors at a Specific Hour

This section details the process of predicting active doctors at a given hour using a trained Random Forest model.

3.4.1 Prediction Process

The prediction involves the following steps:

1. **Input:** The function accepts:
   - Desired hour of the day.
   - Original dataframe containing all doctor data.
   - Trained Random Forest model.
   - Desired number of top doctors to return.
2. **Feature Representation for Prediction:**
   - Calculates the sine and cosine of the input hour to represent time cyclically.

- Constructs a feature vector for each doctor by combining pre-calculated features (e.g., PCA components, survey rate, usage time) with the time features.

3. **Probability Prediction:**
   - Feeds the feature vectors into the trained Random Forest model.
   - Obtains a probability for each doctor, indicating the likelihood of being active at the given hour.
4. **Ranking and Selection:**
   - Ranks doctors based on their predicted probabilities.
   - Selects the top N doctors with the highest probabilities.
5. **Output:**
   - Returns a dataframe containing information about the selected top N doctors, including their NPI, state, region, specialty, and predicted probability of being active at the input hour.

Top doctors most likely to participate at 10:00:

	NPI	State	Region	Speciality	Active_Probability
480	1000000480	GA	West	General Practice	1.00
417	1000000417	GA	South	Radiology	1.00
921	1000000921	OH	South	Cardiology	1.00
290	1000000290	IL	Northeast	Oncology	1.00
632	1000000632	FL	South	Oncology	1.00
681	1000000681	GA	West	Neurology	1.00
669	1000000669	IL	West	Pediatrics	1.00
336	1000000336	FL	South	Pediatrics	1.00
476	1000000476	IL	Midwest	General Practice	0.99
546	1000000546	OH	West	Cardiology	0.99
343	1000000343	MI	Northeast	Cardiology	0.99
354	1000000354	CA	South	Oncology	0.99
694	1000000694	MI	Northeast	General Practice	0.99
16	1000000016	GA	Midwest	Cardiology	0.99
718	1000000718	IL	West	Orthopedics	0.99
266	1000000266	GA	West	Pediatrics	0.99
730	1000000730	NC	Midwest	Pediatrics	0.99
782	1000000782	OH	Midwest	General Practice	0.99
949	1000000949	IL	Midwest	Neurology	0.99
409	1000000409	OH	South	Neurology	0.98

Figure 3.6: Output of Prediction for Doctors likely to attend at the time 10 AM

3.4.2 Underlying Logic

The prediction process utilizes the Random Forest model's capability to identify patterns in historical data. By incorporating time-specific features (sine and cosine of the hour), the model estimates the probability of a doctor being active at a particular hour, considering both their general behavior and temporal context.

3.4.3 Accuracy Considerations

The accuracy of the predictions depends on:

- The quality and comprehensiveness of the training data.
- The model's ability to capture variations in doctor behavior.
- External factors and changes in behavior not represented in the training data.

3.5 User Interface Overview

3.5.1 Features of the UI (Streamlit-based)

The Doctor Engagement predictor application is built using **Streamlit**. It provides:

- User-friendly input sliders for pollutant values.
- Real-time prediction output.
- Interactive interface that updates with every input change.

3.5.2 Screenshots of the Interface

The interface consists of:

- Enter Usage Time.
- Enter Login Hour (0-23)
- Enter Logout Hour (0-23)

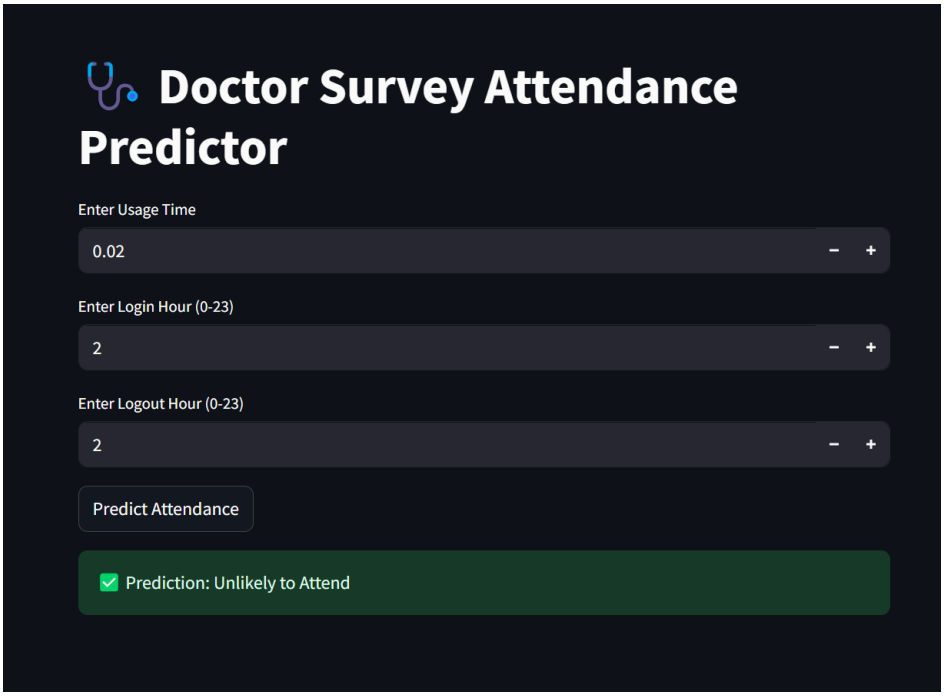

Figure 3.7: Streamlit Interface: Input Panel

3.6 Future Improvements for Enhanced Prediction and Wider Applicability

While the current model provides valuable insights into predicting active doctors at specific hours, several enhancements can be implemented to further improve its performance, usability, and applicability to a broader range of fields.

3.6.1 Data Enrichment and Feature Engineering

1. **Incorporating External Data Sources:** Integrating external data sources, such as doctor demographics, patient population characteristics, and regional health trends, could provide valuable context and improve prediction accuracy. This would involve acquiring and preprocessing relevant data, ensuring data privacy and security.

2. **Advanced Feature Engineering:** Exploring more sophisticated feature engineering techniques, such as time series analysis, could capture complex temporal patterns in doctor activity and enhance prediction accuracy. This might involve using moving averages, lagged features, or incorporating external events that might influence doctor behavior.

3. **Real-time Data Integration:** Incorporating real-time data streams, such as doctor availability status or patient appointment schedules, could enable dynamic predictions and provide more actionable insights. This would require establishing data pipelines for real-time data acquisition and processing.

3.6.2 Model Enhancement and Validation

1. **Exploring Alternative Models:** Investigating other machine learning models, such as deep learning or gradient boosting, could potentially improve prediction accuracy. This would involve evaluating different models on the dataset and selecting the best-performing one based on rigorous evaluation metrics.

2. **Hyperparameter Optimization:** Fine-tuning the hyperparameters of the chosen model could further enhance its performance. This could be achieved through techniques like grid search or Bayesian optimization, which systematically explore different hyperparameter settings to identify the optimal configuration.

3. **Robust Validation Strategies:** Implementing more comprehensive validation techniques, such as cross-validation or time-based splitting, would provide a more realistic assessment of the model's generalizability and performance on unseen data.

3.6.3 Expanding Applicability to Other Fields

1. **Generalizing the Framework:** The core prediction framework can be adapted to other domains with similar prediction goals, such as predicting customer activity in retail, forecasting demand in supply chain management, or anticipating user engagement in online platforms. This would involve identifying relevant features and adapting the model architecture to suit the specific context.

2. **Developing User-Friendly Interfaces:** Creating intuitive and user-friendly interfaces for data input, model training, and prediction visualization would enhance the accessibility and usability of the system for a wider range of users, including those with limited technical expertise.

3. **Incorporating Explainability and Interpretability:** Integrating methods to explain the model's predictions and provide insights into feature importance

would increase user trust and facilitate better understanding of the underlying patterns driving the predictions.

3.6.4 Ethical Considerations and Responsible AI

1. **Addressing Bias and Fairness:** Ensuring fairness and mitigating potential biases in the data and model are crucial for responsible AI development. This would involve analyzing the data for potential biases and implementing techniques to mitigate their impact on predictions.

2. **Maintaining Data Privacy and Security:** Implementing robust data privacy and security measures is paramount, especially when dealing with sensitive information like doctor and patient data. This would involve adhering to data protection regulations and employing secure data storage and processing practices.

3. **Promoting Transparency and Accountability:** Fostering transparency in the model's development and deployment processes, along with establishing clear accountability mechanisms, are essential for building trust and ensuring ethical use of the system.

By focusing on these future improvements, the prediction model can be further refined to achieve higher accuracy, broader applicability, and responsible use in various fields, ultimately contributing to better decision-making and improved outcomes.

Chapter 4

CONCLUSION

This project successfully developed a framework for predicting the activity of doctors at specific hours of the day based on their historical usage patterns and survey participation. Utilizing a dataset containing doctor information and activity logs, the project employed machine learning techniques, including feature engineering, Principal Component Analysis (PCA), and a Random Forest Classifier, to achieve accurate predictions and provide insights into doctor engagement.

Key Contributions

- Developed a robust system for predicting the likelihood of doctors being active (participating in surveys) at a given hour using a Random Forest model, achieving promising accuracy on unseen data.
- Identified key features contributing to doctor activity, such as survey rate, usage time, and login/logout patterns, through feature importance analysis and PCA.
- Enabled the prediction of active doctors at specific hours, facilitating targeted engagement strategies to maximize survey participation.
- Presented the findings using clear visualizations, such as correlation heatmaps, activity by hour plots, and cluster analysis, aiding in understanding and interpreting the results.
- Provided a foundation for further research and development in predicting doctor behavior and optimizing engagement strategies in healthcare settings.

This study demonstrates the potential of machine learning and data analysis in understanding and predicting doctor activity patterns. By leveraging historical data and advanced modeling techniques, the project offers valuable insights into doctor engagement and provides a tool for targeted outreach. Future work can focus on incorporating external data sources, exploring alternative models, and developing user-friendly interfaces for wider applicability and practical use in healthcare settings.

References

[1] McKinney, W. *pandas: a Foundational Python Library for Data Analysis and Statistics*. Python for High Performance and Scientific Computing, 14, 2011.

[2] Harris, C.R., Millman, K.J., van den Berg, S. et al. *Array programming with NumPy*. Nature 585, 357–362 (2020). [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Duchesnay, E. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830, 2011. [Online]. Available: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[4] Waskom, M. L. *seaborn: statistical data visualization*. Journal of Open Source Software, 6(60), 3021, 2021. [Online]. Available: https://doi.org/10.21105/joss.03021

[5] Hunter, J. D. *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9, 90-95, 2007. [Online]. Available: https://doi.org/10.1109/MCSE.2007.55

[6] Breiman, L. *Random Forests*. Machine Learning, 45(1), 5-32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[7] Pearson, K. *On lines and planes of closest fit to systems of points in space*. Philosophical Magazine, 2(11), 589–609, 1901. [Online]. Available: https://doi.org/10.1080/14786440109462720