MULUTU CATHOLIC DISPENSARY

 **  
MACHINE LEARNING MODEL FOR PREDICTING VACCINE COMPLETION AMONG CHILDREN**

AUTHORS

ABIGAEL SAMSON -ENGAGE/SEKU/T2-009/2024  
 BEATRICE WANGARI-ENGAGE/SEKU/T2-008/2024

**Table of Contents**

ACKNOWLEDGEMENT…………………………………………………………………………………………..2

ABSTRACT………………………………………………………………………………………………………………..3

[DECLARATION 4](#_Toc178687194)

[Chapter 1: Introduction 5](#_Toc178687195)

[Chapter 2: Methodology 7](#_Toc178687196)

[2.0 Method of data collection 7](#_Toc178687197)

[For this study on completion of vaccine among children secondary data was used to complement primary research. 7](#_Toc178687198)

[Secondary data refers to the information that has been collected by other researchers or organization and it was used here to analyse completion of vaccine among children. 7](#_Toc178687199)

[2.1 variables used 8](#_Toc178687200)

[2.20 machine learning algorithms 8](#_Toc178687214)

[logistic regression method 8](#_Toc178687215)

[2.3 coding and deployment 9](#_Toc178687220)

[Data cleaning 9](#_Toc178687221)

[1.Data Inspection 9](#_Toc178687222)

[2. Identify Missing Values 9](#_Toc178687225)

[3. Correcting Data Types 9](#_Toc178687231)

[4. Removing Duplicates 9](#_Toc178687233)

[5. Outlier Detection 9](#_Toc178687235)

[6. Standardizing Values 10](#_Toc178687238)

[7. Data Transformation 10](#_Toc178687241)

[8. Data Validation 10](#_Toc178687244)

[10. Final Review 10](#_Toc178687247)

[Chapter 3: Model Performance 13](#_Toc178687256)

[14](#_Toc178687257)

[Chapter 4: Conclusion and recommendation 18](#_Toc178687258)

[Challenges and limitations 18](#_Toc178687259)

AKNOWLEDGEMENT

We would like to thank the Almighty for the far he has taken us throughout the internship period at Mulutu Catholic Dispensary. We would like to extend our heartfelt gratitude to Mulutu Catholic Dispensary for providing us with the opportunity to complete our internship. The experience, mentorship and guidance we received have greatly enhanced our skills and understanding of healthcare practices.

We are especially thankful to the ENGAGE Project for facilitating and supporting our training throughout this internship. They have given us a good learning environment that has helped us to gain practical knowledge that will help us in future.

ABSTRACT

This report highlights the internship experience undertaken at the facility as part of a comprehensive healthcare training program facilitated by the ENGAGE Project. This internship provides experience in various healthcare facilities, we majored on immunization department. Through supervision and mentorship from the hospital staff, we gained a deep understanding of the healthcare services. This internship has greatly enhanced our competencies in clinical and non-clinical areas, preparing us for future roles in the healthcare sector.

DECLARATION

We hereby declare that the internship undertaken at Mulutu Catholic Dispensary was completed under the training and guidance of the ENGAGE Project. We declare that this is our original work and to the best of our knowledge as the authors of the project. We accept responsibility for any omission and error of any kind. This declaration is made in good faith and reflects the true nature of the internship and training program experienced.

**Chapter 1: Introduction**

**1.1 Introduction**

Vaccination is crucial for preventing infectious diseases in children, ensuring public health, and achieving herd immunity. On our project, these vaccines include polio, MR1, MR2, 1st dose, 2nd dose, 3rd dose and birth dose. These vaccines prevent diseases such as measles, polio, diphtheria and whooping cough. However, disparities in vaccination completion rates persist due to various factors, including socioeconomic status, geographic location, and parental awareness. Leveraging machine learning to predict vaccination completion can inform targeted interventions and policy-making. This issue of non completion of vaccines among children presents a significance public health concern. This leaves children vulnerable to preventable diseases.

This report focuses on the completion of childhood vaccinations at Mulutu Dispensary, examining the current vaccination rates and identifying the barriers to full immunization. There’s need to explore factors such as accessibility to healthcare facility, family status and challenges within the healthcare.

**1.2 Problem Statement**

The completion of childhood vaccination schedules is crucial for preventing the spread of infectious diseases and ensuring public health. Mulutu Dispensary faces challenges in achieving full vaccine completion among children. Despite the availability of vaccines, many children do not complete their vaccination schedules. Understanding the predictors of vaccination completion can aid health authorities in addressing these gaps. This report focuses on developing a machine learning model that accurately predicts vaccination completion among children. The report will assess the current vaccination rates, identify gaps in vaccine completion, and propose strategies to improve coverage and focusing on promoting healthcare assess.

**1.3 Justification**

Using machine learning techniques can enhance the understanding of the factors influencing vaccination rates. The ability to predict completion will allow for tailored outreach strategies, thereby improving public health outcomes. This model aims to support healthcare providers and policymakers in making informed decisions. This report is justified by the need to understand and address the factors contributing to low vaccine completion rates at the dispensary. Improved vaccination rates will not only protect individual children but also enhance community immunity, reducing the overall disease burden.

**1.4 Objectives**

The objectives of this report on the completion of vaccinations among children at Mulutu Dispensary are as follows:

* Assess the current vaccination completion rates.
* To develop a predictive machine learning model for vaccination completion.
* To analyse the key factors influencing vaccination rates among children.
* Evaluate the level of parental awareness and education regarding the importance of completing full vaccination schedule.
* To provide recommendations based on model insights for improving vaccination uptake.
* Analyse the operational challenges at the dispensary.

**Chapter 2: Methodology**

**2.0 Method of data collection**

For this study on completion of vaccine among children secondary data was used to complement primary research.

Secondary data refers to the information that has been collected by other researchers or organization and it was used here to analyse completion of vaccine among children.

**2.1 variables used**

* The dataset includes the following features:
* CHILD NO: Unique identifier for each child.
* SEX: Gender of the child (male/female).
* Y.O.B: Childs birth month and year.
* FAMILY STATUS: A categorical variable representing the family’s social economic status (represented by numeric codes like if the child has 1 parent is one and if she/he has both is 2)
* RESIDENCE: Where the child lives(village/location)
* BIRTH DOSE: Whether the child received the birth dose(0 or 1)
* 1st DOSE,2nd DOSE ,3rd DOSE: Vaccination doses
* In each dose the child is given 4 vaccines (1-only one vaccine given,2-only two vaccines given,3-only three vaccines given,4-four vaccines given, 0-no vaccine given)
* VIT A: whether the child received vitamin A supplements (0 or 1)
* 0-no vaccine given, 1- vaccine given
* MR1, MR2:doses for the measles –Rubella vaccine (0 or 1)
* 0-no vaccine given, 1- vaccine given

**2.20 machine learning algorithms**

**logistic regression method**

It’s used for binary classification of task.

(it helps in understanding the relationship between a set of independent variables a binary dependent variables)

The logistic function transforms linear combination of features into a probability value between 0 and 1.

This model is designed to calculate the probability that a child will complete vaccine combining the input feature and applying logistic regression model.

**2.3 coding and deployment**

**Data cleaning**

**1.Data Inspection**

* Load the Data: Use libraries like Pandas (Python)to load your dataset.
* Explore the Data: Check data types, summary statistics, and a few rows to understand the structure and content.

**2. Identify Missing Values**

* Detection: Use functions to find missing values (e.g., is null () in Pandas).
* Handling Strategies:
  + Remove Rows/Columns: If too many missing values are present.
  + Dropping missing values: After identifying the missing values, some were removed.
  + Imputation: Fill in missing values using mean, median, mode, or predictive models.

**3. Correcting Data Types**

* Ensure each column has the appropriate data type (e.g., dates should be in date format, numerical values should not be stored as strings).

**4. Removing Duplicates**

* Identify and remove duplicate rows using functions like drop duplicates () in Pandas.

**5. Outlier Detection**

* Use statistical methods (like Z-scores or IQR) to identify outliers.
* Decide whether to remove or adjust outliers based on their context.

**6. Standardizing Values**

* Normalization: Ensure consistent formatting (e.g., date formats, text case).
* Encoding Categorical Variables: Convert categorical data into numerical form (e.g., one-hot encoding).

**7. Data Transformation**

* Create new features or aggregate data if necessary.
* Normalize or scale numerical features if required by your algorithms.

**8. Data Validation**

* Cross-check cleaned data against original sources to ensure accuracy.
* Perform sanity checks (e.g., total should equal sum of parts).

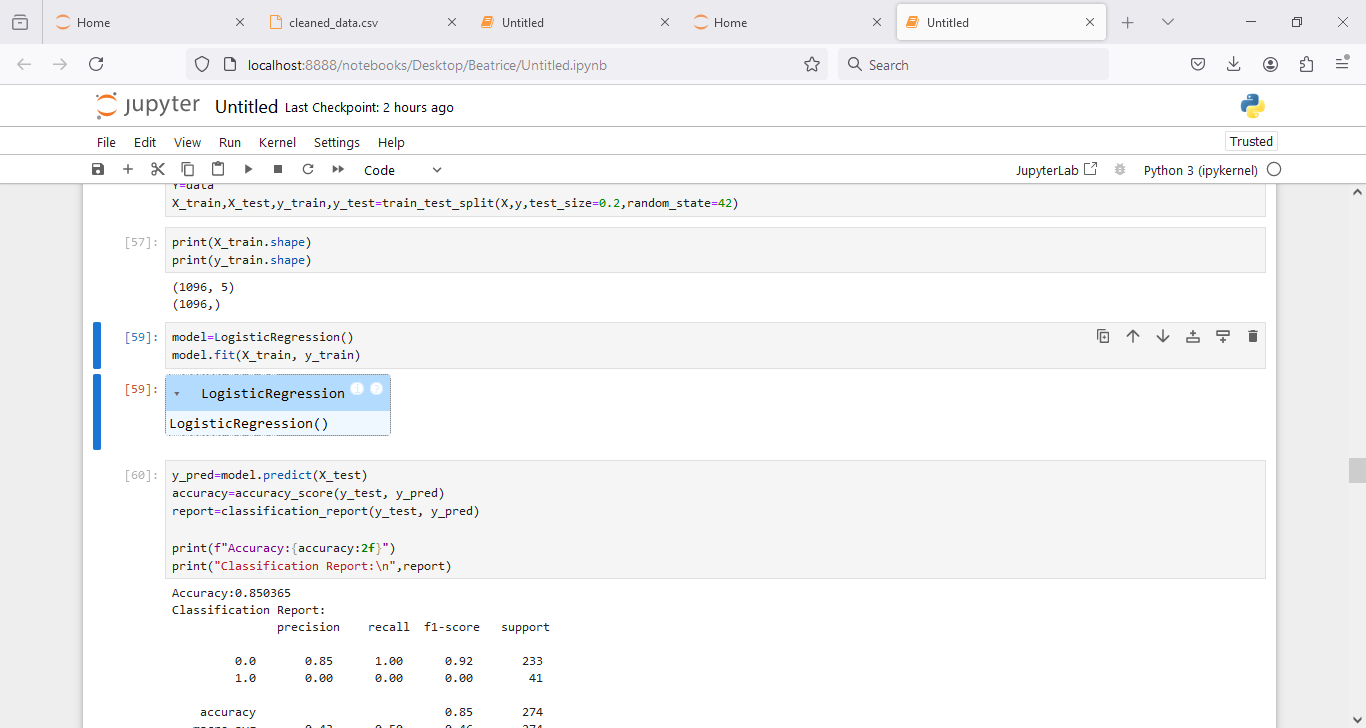
**10. Final Review**

* Review the cleaned dataset to ensure it meets your project requirements.
* Visualize the data if necessary to confirm that the cleaning process was successful.

**Tools and Libraries**

* Data Visualization: Matplotlib, Seaborn
* **Pandas** for data manipulation
* **Scikit-learn** for machine learning algorithms
* **Matplotlib** and **Seaborn** for data visualization
* **Import NumPy as np**

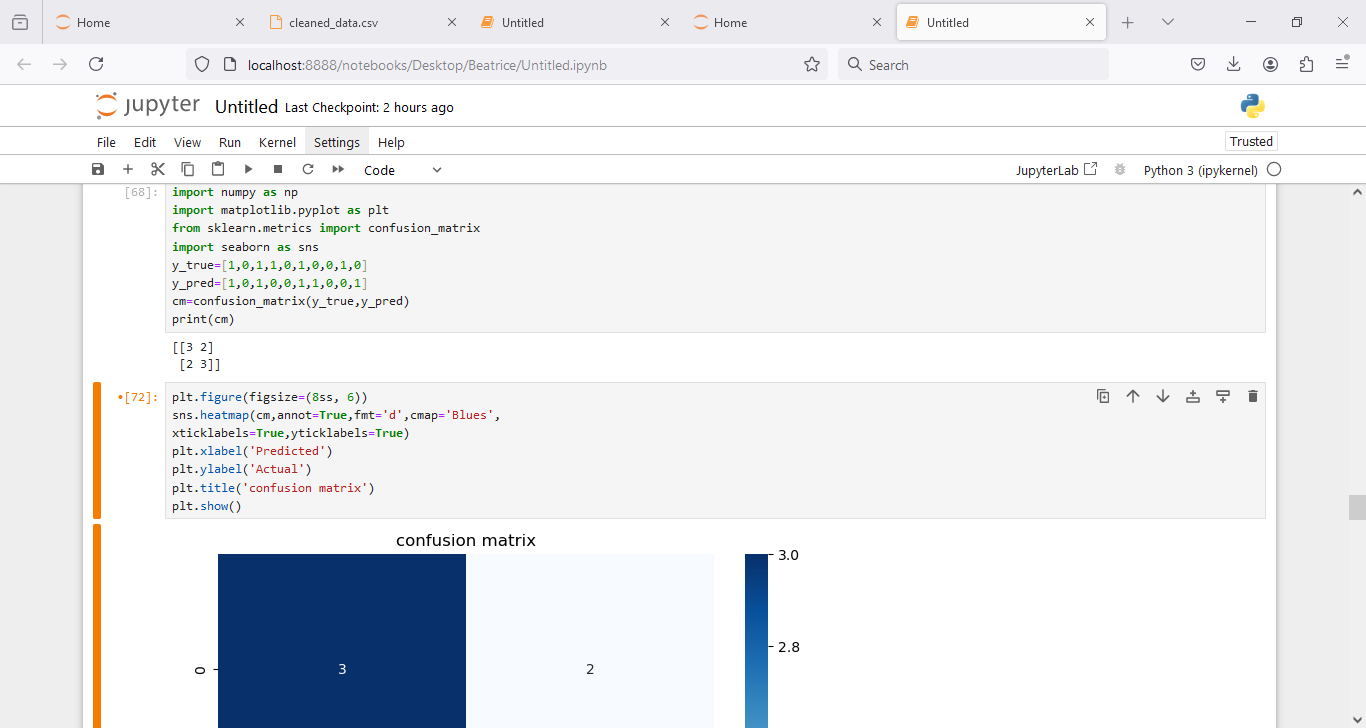
**2.5 Model Training**

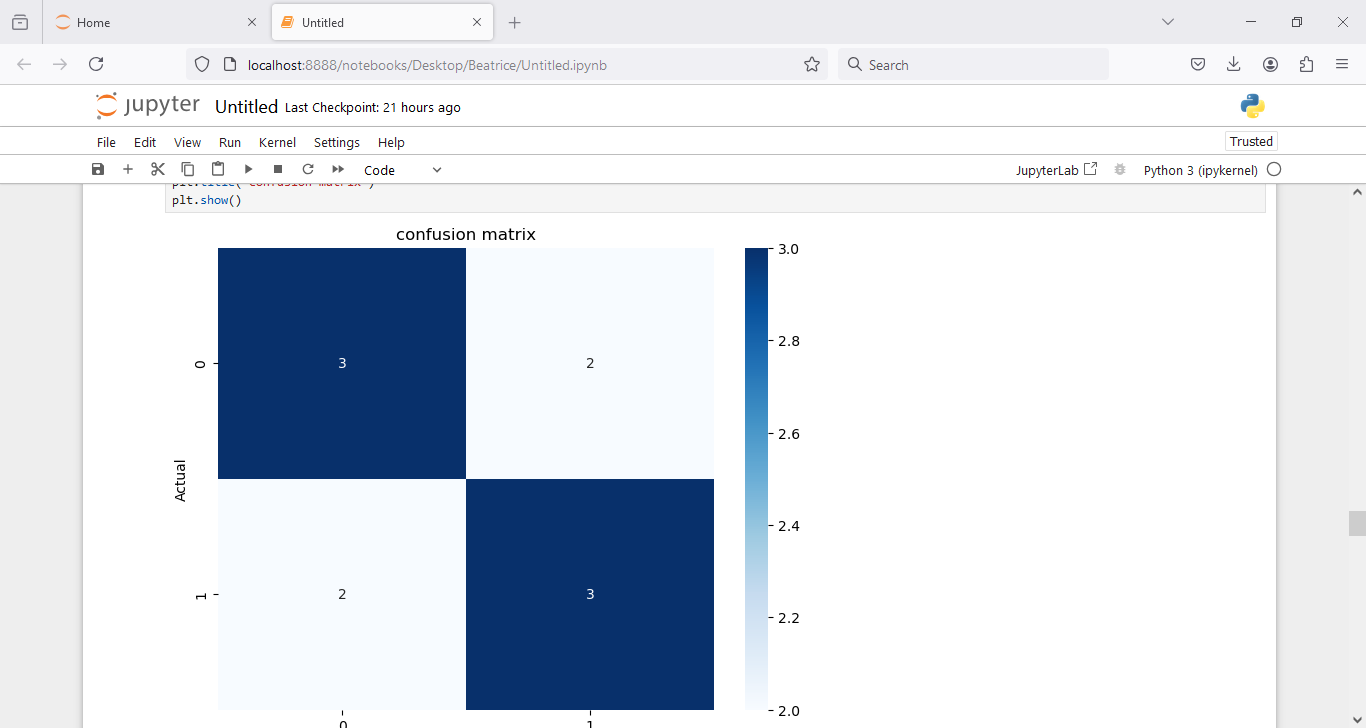
****

The dataset was split into training and testing sets. The model was trained using the training set, employing techniques such as cross-validation to ensure generalizability

**Confusion matrix**

It provides insight into model performance by showing true positive, false positive, true negative and false negative.





**2.6 Model Evaluation**

Model evaluation metrics included:

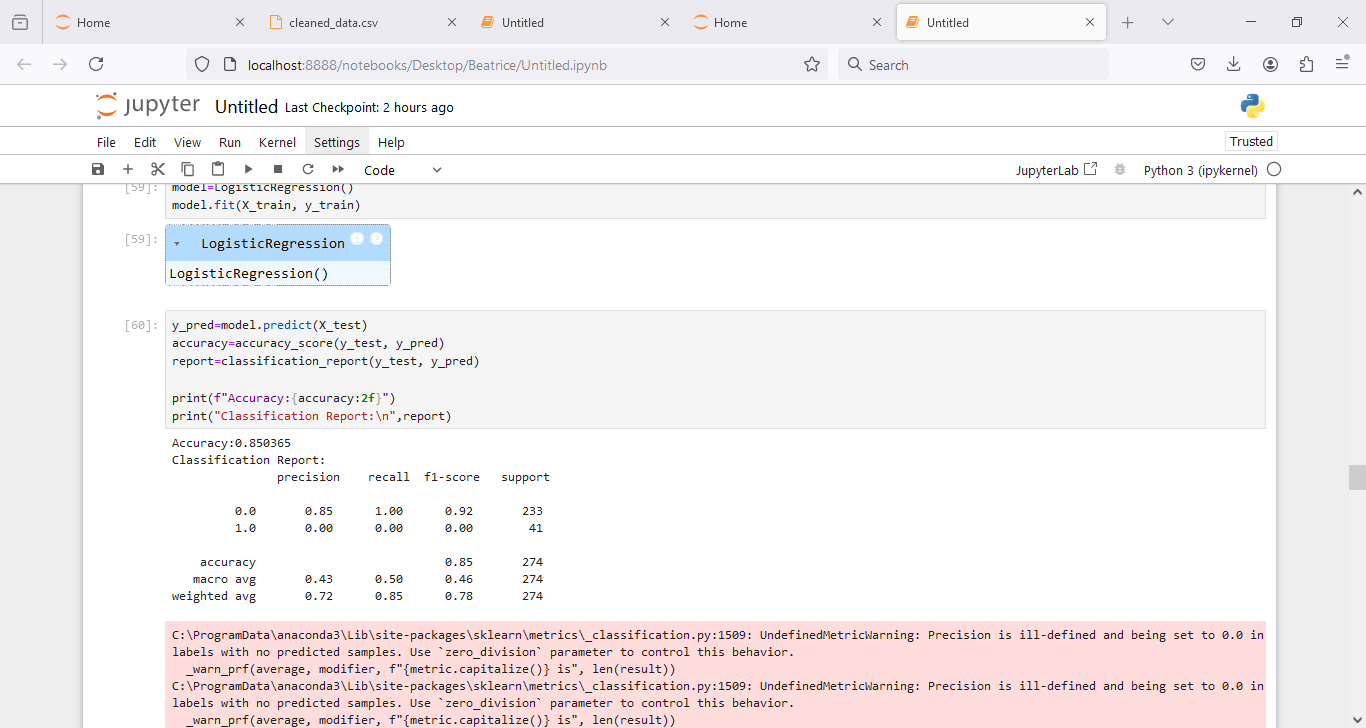
* **Accuracy:** 0.85
* **Precision:** 0.85
* **Recall:** 1.00
* **F1 Score:** 0.92

These metrics indicate a well-performing model, particularly in terms of recall, which suggests that the model is effective in identifying children who will complete vaccinations.

**Evaluation Metrix**

* Accuracy (0.85)-The proportion of correctly predicted outcomes, both true positive and true negatives, out of all predictions. This suggest that the model has a good performance and can be used to predict vaccine completion amongst children.
* Precision (0.85)-Of all the children predicted to complete the vaccine,85% actually did complete it.15% were false positive and 85% completed their vaccine.
* Recall (1.00)-The model correctly identified 100% of all children who actually completed the vaccine.
* F1 Score (0.92)-A balanced measure of precision and recall, showing the model is fairly good at both identifying and correctly predicting vaccine completion.

**Chapter 3: Model Performance**

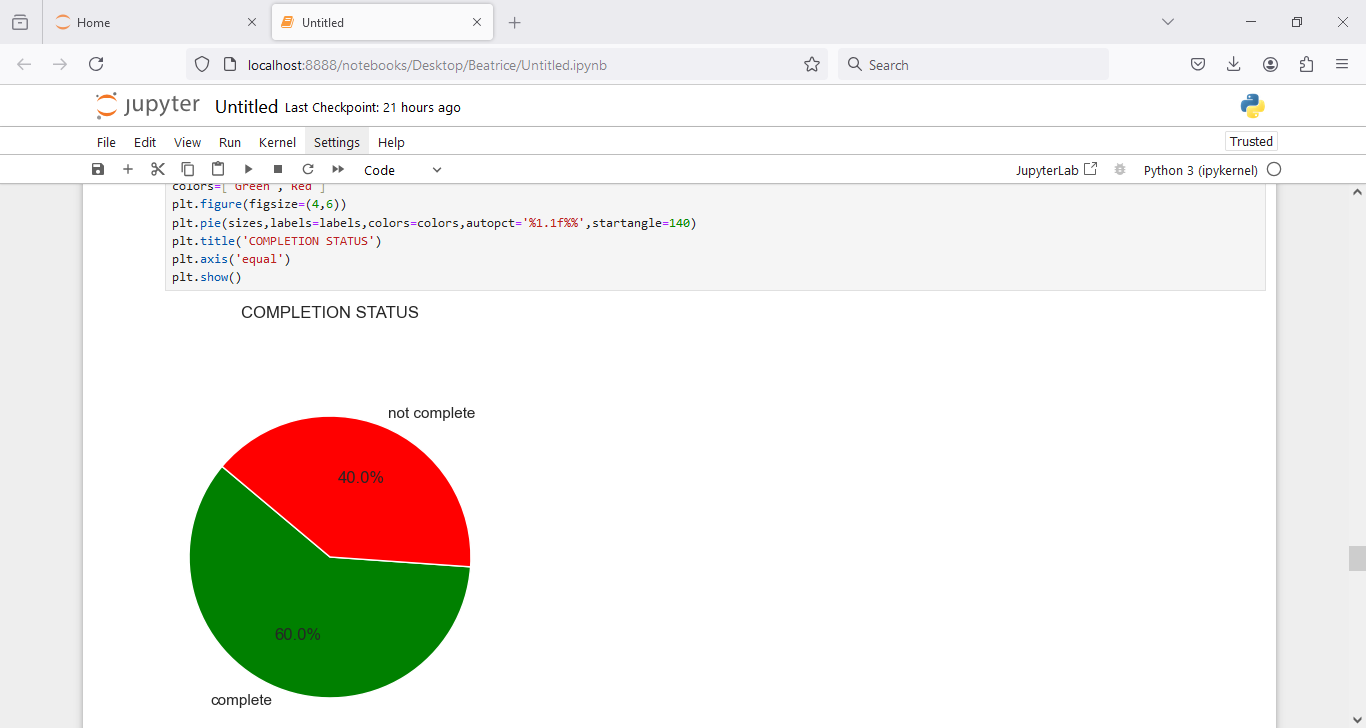
****

**Model performance metrics**

* **Accuracy:** 0.85
* **Precision:** 0.85
* **Recall:** 1.00
* **F1 Score:** 0.92

**3.1 Insights from Models**

The model's performance suggests that the most significant predictors of vaccination completion include family status, residence, and compleletion status. By analysing feature importance, we can identify which factors most influence completion rates.

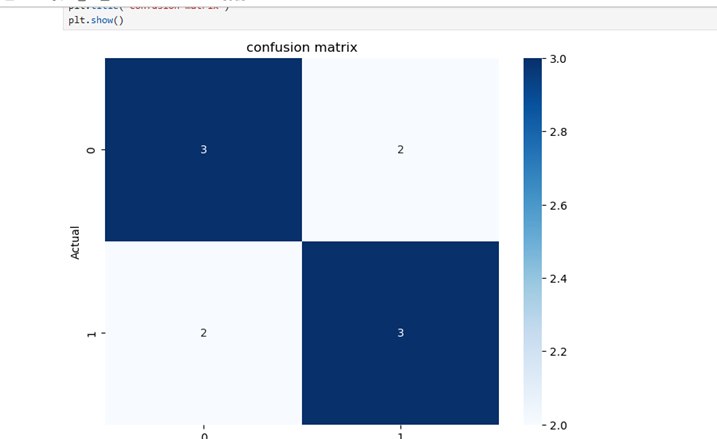


60% of the population has completed their vaccination the remaining 40% still represents a significant portion that requires attention.

Ongoing efforts to increase vaccine coverage are essential for ensuring the broader health and safety of the community.

**3.2 Interpretation of Results**

The high recall (1.00) indicates that the model successfully identifies all children who completed their vaccinations, while precision (0.85) shows a good balance in predicting true positives versus false positives. The F1 score of 0.92 demonstrates a solid balance between precision and recall.

****

**3.3 Confusion Matrix Analysis**

The confusion matrix provided the following insights:

Predicted positive predicted negative

Actual positive true positive (TP) false negative(FN)

Actual negative false positive(FP) true negative(TN)

Given the provided metrics:

* **Recall (1.00)-**There were no false negatives, every child who complected their vaccination was correctly predicted by the model.
* **Precision (0.85**)-Out of all the children the model identified as having completed their vaccine schedule, 85% actually did, while 15% were false positives.
* **Accuracy (0.85**)-indicates that 855 of the total predictions are correct.

**Chapter 4: Conclusion and recommendation**

**Challenges and limitations**

Incomplete data- the data was incomplete making it difficult to assess vaccination status accurately.

Access to healthcare- this can affect vaccination rates among different socioeconomic groups. Parents living in area like Katothya have difficulty in accessing healthcare facilities.

Vaccine hesitancy- fear or miscommunication about the vaccine.

Cultural beliefs- this can influence completion rates and may not be fully captured in data.

Understanding of vaccine schedules- parents may lack awareness of the importance of following the vaccination schedule.

Distribution and accessibility of the vaccines in the health facility.

Lack of follow ups and reminders to parents.

Funding limitations leading o reduced outreach in interior areas.

These challenges and limitations should be carefully considered. Acknowledging the can help in framing the report accurately and responsibly and potentially guide recommendations for addressing identified issues.

**Conclusion**

The analysis of the machine learning model's performance in predicting vaccination outcomes for children indicates an overall accuracy of 85%. However, the classification report reveals critical issues, particularly regarding the model's performance on the minority class (1.0), which represents children who are not vaccinated. The precision, recall, and F1-score for this class are all zero, indicating that the model fails to identify any unvaccinated children effectively. This imbalance suggests a need for improved strategies to ensure that machine learning models not only achieve high overall accuracy but also perform adequately across all classes, particularly for vulnerable populations.

**Recommendations**

1. **Address Class Imbalance**: Implement techniques such as oversampling the minority class or under sampling the majority class to balance the dataset. Consider using advanced techniques like Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic examples.
2. **Model Optimization**: Explore different machine learning algorithms and hyperparameter tuning to enhance model performance on both classes. Ensemble methods, like Random Forest or Gradient Boosting, could provide better discrimination between classes.
3. **Feature Engineering**: Investigate and include additional relevant features that may impact vaccination status, such as socio-economic factors, family status, and geographical data. This may improve the model's ability to predict vaccination outcomes.
4. **Evaluation Metrics**: Use more comprehensive evaluation metrics, such as ROC-AUC and precision-recall curves, to gain better insights into model performance, especially for the minority class.
5. **Continuous Learning**: Set up a feedback loop to continually update the model with new data as it becomes available. This will help adapt to changing vaccination patterns and improve long-term effectiveness.
6. **Collaboration with Public Health Officials**: Work closely with public health organizations to ensure the model's predictions can be used effectively in outreach and education campaigns targeting unvaccinated children.
7. **Community Engagement**: Implement community-based initiatives informed by the model’s insights to address barriers to vaccination, ensuring that interventions are culturally appropriate and effectively communicated.