**MACHINE LEARNING MODEL FOR PREDICTING TEENAGE PREGNANCY.**

Presented by:

Favor Ndanu

ENGAGE/SEKU/T2-002/2024

Favormutanu254@gmail.com

Faith Cheptoo

ENGAGE/SEKU/T2-013/2024

+254718774716

Faithcheptoo720@gmail.com

Tabitha Kariuki

ENGAGE/SEKU/T2-004/2024

Tabithakariuki0002@gmail.com

ACKNOWLEDGEMENT.

We would like to express our sincere gratitude to the Engage stakeholders who collaborated with us in this project.

Again we would like to thank our coordinator Dr. Fransisca Mwangangi and all our trainers who taught us during the training process. Special thanks to our parents for giving their consent and releasing us to participate in this training. Thanking also our fellow participants who made this experience a wonderful one and the learning experience amazing.

ABSTRACT

Teenage pregnancy has been a huge problem with building physical, mental and economic implications for teenage mother’s future and that of their babies as well.

The attributes collaborated for this study include age, residence, access to reproductive health care, education and teenage pregnancy. The findings for this study suggested that higher rates of pregnancy are related to rural area residence lower income level lesser education and issues related to the region. To solve this problem machine learning model i.e. Logistic regression algorithm was employed on the collected data. The performance of this model was calculated using three metrics which included accuracy, precision and F1-score values which indicated that this developed regression model is more accurate it holds the value of 63%.

**CHAPTER ONE: INTRODUCTION**

* 1. Background

Teenage pregnancy remains a pressing global health issue. In this case the teenage females are of age 14-19 years who have not completed their core education and has few or no skills and are dependent on their families for survival or cannot survive on their own hence becoming burdens to their parents. Teenage pregnancy has negative health consequences.

Children born are likely to be premature with low birth weights while their mothers likely to suffer postpartum depression. Due to this teenage mothers have high rates of depression and suicidal cases.

* 1. Problem statement.

Keeping this into consideration we have performed an analytical based research study in order to find correlation between various factors so as to draw conclusions to solve this problem of teenage pregnancy.

* 1. Justification.

We made use of machine learning model using logistic regression. Machine learning is a subset of artificial intelligence that enables a set of algorithms and statistical models that enable computers to perform tasks without explicit instructions. Algorithms were used to identify patterns with data which are then used to create a data model that can make predictions. Machine learning adaptability makes it a greater choice in scenarios where a data is constantly changing in this case teenage pregnancy and teenage motherhood rates making it a great fit for this project. Under this machine learning a logistic regression model has been used.

* 1. Objectives.

The primary objective of this study is to develop a model to identify teenagers at risk of being pregnant. The model will help in ensuring that the girls can access reproductive health care and contraceptives to reduce the rate of teenage pregnancies.

**CHAPTER 2: METHODOLOGY.**

In our project we used machine learning model to analyze secondary data whereby, the data we used on our model had been collected from various databases and files. To predict and identify significant determinants of teenage pregnancy using python software and machine learning algorithms were applied. Our teenage predictors were chosen using feature engineering technique.

We considered the variables like age, education level, poverty level, access to reproductive health care, regions, residence and teenage pregnancy. This were the attributes that contributed to prediction of teenage pregnancy. The effectiveness of our model was evaluated using accuracy, precision and F1\_score for the evaluation performance of our prediction. In this case our predictor variable was teenage pregnancy which made independent variables appropriate for our analysis.

**2.1 Data Source**

Th**e** origin of our dataset was from existing databases and files (e.g. public health database, school surveys).

Our data was collected through surveys and administrative records.

**2.2. Dataset Description**

The dataset contains 6,136 observations and hadKey variables which include; Age, Poverty Level, Education Level, Access to Reproductive Health Care, and Teenage Pregnancy.

2.1. Data Pre-processing.

We issued a high quality dataset for machine learning to make predictions. As a result, managing the missing data and cleaning our data to remove the duplicates. Dataset preprocessing was an essential step and removed the outliers.

2.2. Data Analysis

To describe the socio-demographic characteristics using frequency and percentages. Our data analysis stages included pre-processing the data, feature selection, data splitting, addressing imbalanced data, model building and model performance testing. We used python tool in this our study.

2.3. Feature Selection

The goal of feature selection is to rank and prioritize the most important predictors in the dataset. This is determined by computing the information gain values for each of the selected variables. To find the major factors that significantly resulted in teenage pregnancy, we used logistic regression model in this work. It is a relatively effective method for reducing model complexity and accelerating the process of machine learning algorithms.

2.4. Data Split

For machine learning approaches, our datasets are randomly divided into two parts that is training dataset is used to train the model and then test dataset is used to predict the response variable to see the predicted outcome is similar to the actual outcomes.

2.5. Choice of Model

We chose logistic regression for this problem because of its suitability for binary classification, interpretability. The logistic regression algorithm, included its assumptions (e.g., linear relationship between features and log odds, independence of observations).

2.6. Method of building a predictive model.

We picked the most effective model to do training after we arranged and splitted into training and testing samples. To produce production in our model it was necessary to select the appropriate classification for the result variables. select the appropriate classification for the result variables categorical in nature as well as the numerical variables. In our working model we used supervised classification method were employed i.e. logistic regression. This algorithm was chosen for accuracy, training time ability to handle missing data and ease of understanding the learning.

2.7. Performance evaluation for predictive model.

Following our model training each model performance was assessed and contrasted with one another. The prediction models performance was assed where by precision, accuracy and f1- score were utilized in this model to evaluate the models performance.

**Accuracy**: Measures the overall correctness of the model.

**Precision**: Indicates the proportion of true positives among all positive predictions.

**Recall**: Measures the proportion of true positives identified among all actual positives.

**F1-Score**: Provides a balanced measure of precision and recall.

CHAPTER 3.

### 1. ****Accuracy****

* **Value:** 0.63

The accuracy is the ratio of correctly predicted instances (both True Positives and True Negatives) to the total instances. In this case, the model correctly predicted 63% of the instances.

### 2. ****Precision, Recall, and F1-Score****

These metrics are reported for each class (0 and 1) and averaged across both classes.

Class O:

* **Precision:** 0.64

Of all the instances predicted as class 0, 64% are actually class 0. Precision indicates how many of the predicted positives are true positives.

* **Recall:** 0.78

Of all the actual class 0 instances, 78% were correctly identified by the model. Recall measures the ability of the model to find all relevant instances of the positive class.

* **F1-Score:** 0.70

The F1-Score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall. A score of 0.70 indicates a decent balance between precision and recall for class 0.

* **Support:** 3476

The number of actual instances of class 0 in the dataset.

Class 1:

* **Precision:** 0.60

Of all the instances predicted as class 1, 60% are actually class 1.

* **Recall:** 0.44

Of all the actual class 1 instances, 44% were correctly identified by the model.

* **F1-Score:** 0.51

The F1-Score for class 1 is 0.51, indicating a lower balance between precision and recall compared to class 0.

* **Support:** 2660

The number of actual instances of class 1 in the dataset.

### ****Averages****

### Macro averages

* **Precision:** 0.62

The average precision across both classes without considering the number of instances in each class. It gives equal weight to precision for both classes.

* **Recall:** 0.61

The average recall across both classes. It also gives equal weight to recall for both classes.

* **F1-Score:** 0.60

The average F1-Score across both classes. Like precision and recall, it gives equal weight to the F1-Score for each class.

Weighted average:

* **Precision:** 0.62

The average precision weighted by the number of instances in each class. It accounts for the fact that there are more instances of class 0 than class 1.

* **Recall:** 0.63

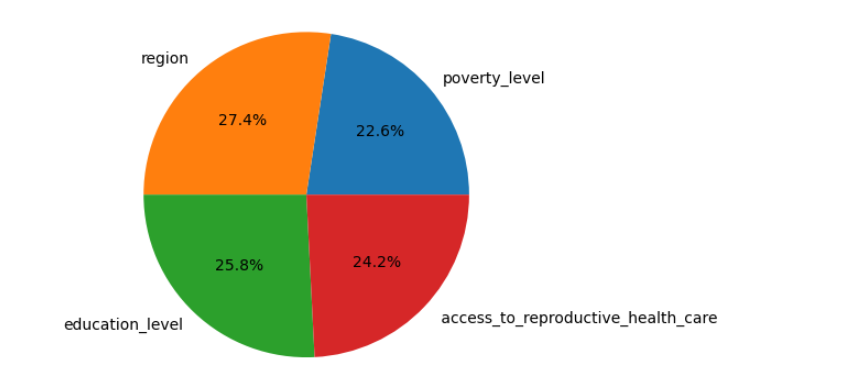
The average recall weighted by the number of instances in each class.

* **F1-Score:** 0.62

The average F1-Score weighted by the number of instances in each class. It provides a balanced measure of performance considering the class imbalance.

### Summary

* **Class 0** has higher recall and precision compared to **Class 1**, meaning the model is better at identifying instances of class 0.
* **Class 1** has lower recall and precision, suggesting the model is less effective at identifying this class.
* The **macro average** provides an overall view of the model's performance, treating all classes equally.



A pie chart is a circular graph divided into slices to illustrate numerical proportions. Each slice represents a category’s share of the total. For predicting teenage pregnancy, a pie chart with the following categories poverty level, access to reproductive health care, region, and region can offer a visual representation of how different factors contribute to the risk.

To interpret a pie chart depicting factors predicting teenage pregnancy, each slice represents a different contributing factor and its relative importance in predicting teenage pregnancy rates.

1. **Poverty Level (22.6%)**:

This slice represents the proportion of teenage pregnancy risk attributed to living in poverty.

A percentage of 22.6% indicates that poverty is a significant factor. Teenagers from low-income backgrounds may face more challenges, such as limited access to education and health resources, which can increase their risk of teenage pregnancy.

1. **Access to Reproductive Health Care (24.2%)**:

This slice shows the impact of access to reproductive health services, including contraception, sexual education, and healthcare facilities.

At 24.2%, this slice underscores the importance of healthcare access. Limited access to these services can lead to higher rates of unintended pregnancies among teenagers, as they may lack information and resources to prevent pregnancy.

1. **Education Level (25.8%)**:

This slice indicates the influence of education level on teenage pregnancy. It includes factors such as the level of formal education and educational attainment of the teenager and their parents.

With 25.8%, education level is a substantial factor. Lower educational.

attainment is often linked with higher teenage pregnancy rates, as education can.

provide both knowledge and opportunities that help in making informed life

choices

4. **Region (27.4%)**:

This slice represents the effect of geographic region on teenage pregnancy rates. It may reflect regional differences in socioeconomic conditions, cultural norms, and access to services.

The largest slice at 27.4% highlights the significant impact of the region where the teenager lives. Different regions may have varying levels of support, healthcare services, and educational resources, which can greatly influence teenage pregnancy rates.

**Overall Interpretation**: The pie chart helps visualize the relative importance of these factors in predicting teenage pregnancy:

* **Region** (27.4%) has the largest share, indicating that geographic location plays the most significant role in teenage pregnancy rates. This could be due to regional differences in cultural attitudes, economic conditions, and availability of resources.
* **Education Level** (25.8%) is the next largest factor, showing that higher educational attainment is associated with lower teenage pregnancy rates. Education can affect life choices and access to information.
* **Access to Reproductive Health Care** (24.2%) follows closely, suggesting that availability and quality of reproductive health services are critical in preventing teenage pregnancies.
* **Poverty Level** (22.6%) is also significant but slightly less influential compared to the other factors. However, it still plays a crucial role as financial constraints can limit access to education and healthcare.

In summary, the pie chart illustrates that while all these factors are important, the region has the highest impact on teenage pregnancy rates, followed by education, healthcare access, and poverty level. Addressing these factors through targeted interventions can help reduce teenage pregnancy rate

CHAPTER 4:

**Conclusions**

1. **Educational Attainment**: Teen pregnancies are often linked to lower educational attainment and reduced future economic opportunities for young mothers. Completing education and accessing vocational training can significantly improve outcomes for both the mother and the child.
2. **Access to Healthcare**: Limited access to sexual and reproductive health services, including contraception and prenatal care, contributes to higher rates of teenage pregnancies. Adequate healthcare services are crucial for prevention and support.
3. **Socioeconomic Factors**: Poverty, lack of parental support, and unstable living conditions increase the likelihood of teenage pregnancies. Addressing these underlying socioeconomic issues is essential for effective prevention.
4. **Family and Community Influence**: Family dynamics and community norms play a significant role in teenage pregnancies. Supportive families and communities can provide crucial guidance and resources to prevent teenage pregnancies.
5. **Sexual Education**: Comprehensive sexual education that covers contraception, consent, and healthy relationships is vital. Education should be inclusive, culturally sensitive, and age-appropriate to be effective.

**Recommendations**

1. **Enhanced Sexual Education**:
   * Implement and expand comprehensive sexual education programs in schools, covering topics like contraception, sexual health, and relationship dynamics.
   * Include information on consent and healthy relationships to foster informed decision-making.
2. **Improved Access to Healthcare**:
   * Increase availability and accessibility of reproductive health services for teenagers, including contraception, pregnancy testing, and prenatal care.
   * Provide confidential services to ensure that teens feel safe seeking help without fear of judgment or breach of privacy.
3. **Support Systems for Teen Parents**:
   * Develop and support programs that offer parenting classes, childcare, and financial assistance to young parents.
   * Establish mentorship and peer support groups to help teenage parents navigate the challenges of parenthood.
4. **Community and Family Engagement**:
   * Engage families and communities in discussions about teenage pregnancy prevention, emphasizing the importance of supportive relationships and open communication.
   * Foster community programs that provide positive role models and mentorship for teens.
5. **Policy and Advocacy**:
   * Advocate for policies that support teenage mothers and fathers, such as improved access to education, healthcare, and social services.
   * Support initiatives that address socioeconomic disparities and provide resources for at-risk youth.
6. **Research and Data Collection**:
   * Invest in research to better understand the causes and consequences of teenage pregnancies, and to evaluate the effectiveness of prevention and support programs.
   * Use data-driven approaches to tailor interventions and policies to the specific needs of different communities.

By implementing these recommendations, it is possible to reduce the incidence of teenage pregnancies and improve outcomes for young parents and their children.