

University of Westminster

# Project Report Monkey Pox Testing Analysis

Data Mining and Machine Learning (7BUIS008W)

Coursework 1

B Gayan Kalinga w2027429 20231057 [gayan.20231057@iit.ac.lk](mailto:gayan.20231057@iit.ac.lk)

**Table of Contents**

[Project Report Monkey Pox Testing Analysis 1](#_Toc163416563)

[Task 01: Domain Understanding Classification 1](#_Toc163416564)

[Task 02: Producing Your Experimental Design 1](#_Toc163416565)

[Task 03: Cleaning and Transforming the Data 2](#_Toc163416566)

[Task 04: Create Predictive Classification Models 3](#_Toc163416567)

[Task 05: How Good Is the Model 4](#_Toc163416568)

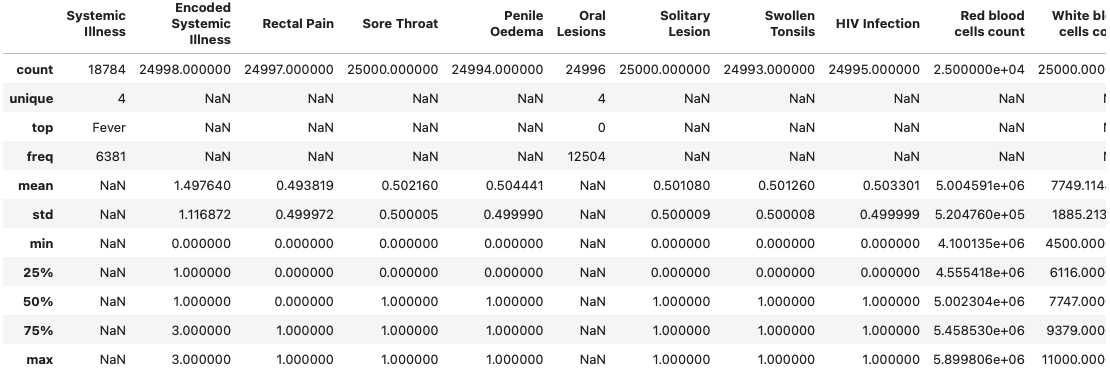
# Task 01: Domain Understanding Classification

Mark the variables logically applicable in the classification modeling of MPOX

|  |  |
| --- | --- |
| Attribute Name | Retain or drop |
| Test ID | Reain |
| Systemic Illness | Retain |
| Sore Throat | Retain |
| Rectal Pain | Retain |
| Penile Oedema | Retain |
| Oral Lesions | Retain |
| Solitary Lesion | Retain |
| Swollen Tonsils | Retain |
| HIV Infection | Retain |
| Red blood cells | Retain |
| White blood cells | Retain |
| Home Ownership | Drop |
| Age | Retain |
| Month of Birth | Drop |
| Health Insurance | Drop |
| Sexually Transmitted Infection | Retain |
| MPOX | Retain |

# Task 02: Producing Your Experimental Design

Basic Statistical Descriptions



Measurement Scale Types

A screenshot of a computer program

Description automatically generated

Distribution of the Result

A bar graph with blue and orange squares

Description automatically generated

# Task 03: Cleaning and Transforming the Data

A: Issues identified in the retained data set and possible variables

|  |  |  |
| --- | --- | --- |
| Dataset or Variable Issue | Name of the Variable | Issue Description |
| Variable issue | Systemic Illness | Instead of “Fever” user typed “fever” |
| Variable issue | Systemic Illness | 6216 records NaN values |
| Variable Issue | Encoded Systemic Illness | 2 records NaN values |
| Variable Issue | Rectal Pain | 3 Records NaN values |
| Variable Issue | Penile Oedema | 6 records NaN values |
| Variable Issue | Oral Lesions | 4 records NaN values |
| Variable Issue | Oral Lesions | 7 records with value No and 4 records with value Yes |
| Variable Issue | Swollen Tonsiles | 7 records NaN values |
| Variable Issue | HIV Infection | 5 records NaN values |
| Variable Issue | Age | 36 records NaN values |
| Variable Issue | Sexually transmitted infection | 4 records NaN values |
| Variable Issue | Oral Lesions | Multiple Data types with Str and Float |
| Variable Issue | MPOX PCR Result | Using Negative & Positive |
| Variable Issue | Age | Having string values |
| Variable Issue | Age | Having minus values |

B: Solution to mitigate issues found with justification for using that solution

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset or Variable Issue | Name of the Variable | The Issue | Solution | Justification |
| Variable issue | Systemic Illness | Instead of “Fever” user typed “fever” | Changed to Fever | Assume mistakenly enterd the lowercase fever |
| Variable issue | Systemic Illness | 6216 records NaN values | Drop the column | Dropped since the same data is available as Encoded Systemic Illness is with binary types |
| Variable Issue | Encoded Systemic Illness | 2 records NaN values | Filled with min value | Since we cannot get mean used the minimum |
| Variable Issue | Rectal Pain | 3 Records NaN values | Filled with the median | Mean does not provide accurate value and used median to replace |
| Variable Issue | Penile Oedema | 6 records NaN values | Filled with the median | Mean does not provide accurate value and used median to replace |
| Variable Issue | Oral Lesions | 7 records with value No and 4 records with value Yes | Filled YES with 1 and No with 0 | Since only 11 records assume 1 is true and 0 is false |
| Variable Issue | Oral Lesions | 4 records NaN values | Filled with the median | Mean does not provide accurate value and used median to replace |
| Variable Issue | Swollen Tonsils | 7 records NaN values | Filled with the median | Mean does not provide accurate value and used median to replace |
| Variable Issue | HIV Infection | 5 records NaN values | Filled with the median | Mean does not provide accurate value and used median to replace |
| Variable Issue | Age | 36 records NaN values | Remove the values | Since filling these data might impact in a bias and only 36 records dropping these |
| Variable Issue | Sexually transmitted infection | 4 records NaN values | Filled with the median | Mean does not provide accurate value and used median to replace |
| Variable Issue | Oral Lesions | Multiple Data types with Str and Float | Converted to Int | Since the values were integer base |
| Variable Issue | MPOX PCR Result | Using Negative & Positive | Converted to Int | To be easier to train machien learning models with the numeric values |
| Variable Issue | Age | Having string values | Converted to Int and removed due to the value is outlier | Value is 20 and it is one record showing as a outlier in the scatter diagram |
| Variable Issue | Age | Having minus values | Converted back to absolute value | Value has more than 500 records so converted plus value given the age is always plus vlaue |
| Variable Issue | Age | Having Mistakenly high values x 2 | Dropped the two values | Since only two values which is hard to assume the correct value is |
| Variable Issue | Age | Having 0 as value | Dropped the value | Since only two values which is hard to assume the correct value is |

C: Outputs (Before & After)

# Task 04: Create Predictive Classification Models

A: Classification algorithm details

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm Name | Type of Algorithm | Learnable Parameters | Possible Hyper – Parameters | Python package source code to call the algorithm |
| LR | Parametric | Weights for features | `C` inverse of regularization strength `penalty` regularization penlty | `sklearn.linear\_model`  `import LogisticRegression` |
| DT | Non-parametric | Splitting rules for nodes | `max\_depth` `min\_samples\_split` `min\_samples\_leaf` `max\_features` | `sklearn.tree`  `import DecisionTreeClassifier` |
| KNN | Non-parametric | Training Dataset | `n\_neighbors` `weights` `uniform` `distance` `metric`  `euclidean`  `manhattan` | `sklearn.neighbors`  `import KneighborsClassifier` |
| SVM (RBF) | Non-parametric | Support vectors, weights for features | `C` regularization  `gamma` kernel coefficient for rbf | `sklearn.svm`  `import SVC` |
| NB | Parametric | Prior probabilities for classes and Features | `alpha:1.0` to avoid zero posibilities  `fit\_prior: True` learn prior probabilities  `class\_prior: None` custom prior probabilities | `sklearn.naive\_bayes`  `import GaussianNB` |

B: Data shape function output

Justify the training-test split ratio and provide an in-text reference.

Code line from the source code.

# Task 05: How Good Is the Model

A: test confusion matrix for each trained model (output screenshots)

B: Five different classification evaluation metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric Name | “USE” or “DO NOT USE” | Justification in relation to the success criteria | Model Name | Metric Score |
| Accuracy |  |  | LR |  |
| DT |  |
| KNN |  |
| SVM (RBF) |  |
| NB |  |
| Recall |  |  | LR |  |
| DT |  |
| KNN |  |
| SVM (RBF) |  |
| NB |  |
| Precision |  |  | LR |  |
| DT |  |
| KNN |  |
| SVM (RBF) |  |
| NB |  |
| F-Measure |  |  | LR |  |
| DT |  |
| KNN |  |
| SVM (RBF) |  |
| NB |  |
| AUC-ROC |  |  | LR |  |
| DT |  |
| KNN |  |
| SVM (RBF) |  |
| NB |  |

C: Best classification model/s based on “USED” performance metrics score and describe how the model satisfies the needs of health care professionals

D:

E:

F: