

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)

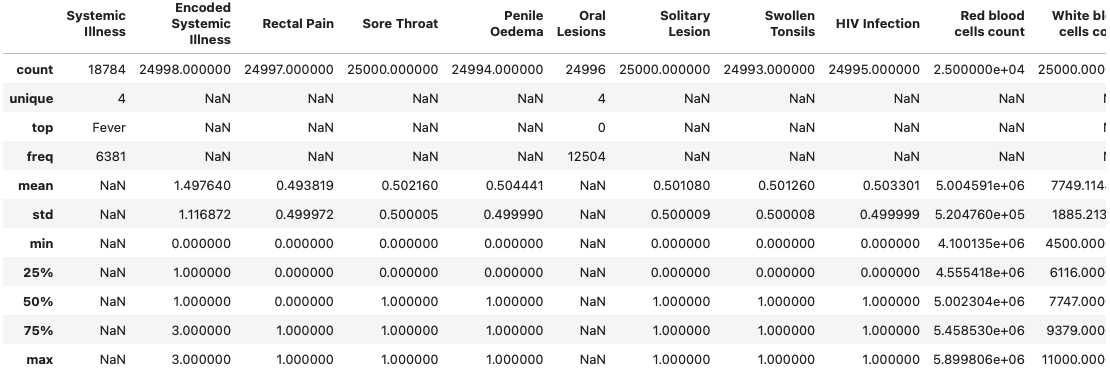
# Task 01: Domain Understanding Classification

Mark the variables logically applicable in the classification modeling of MPOX

|  |  |
| --- | --- |
| Attribute Name | Retain or drop |
| Test ID | Drop |
| 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 | Drop |
| White blood cells | Drop |
| 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 | USE | Since the model should predict as many MPOX positive subjects as possible and accuracy will give an overall view of how well the model is performing in terms of correct predictions | LR | 67% |
| DT | 67% |
| KNN | 65% |
| SVM (RBF) | 69% |
| NB | 70% |
| Recall | USE | Since importance of correctly detecting MPOX positive subjects by focusing on the true positive rate | LR | 90% |
| DT | 84% |
| KNN | 79% |
| SVM (RBF) | 93% |
| NB | 89% |
| Precision | USE | Since this evaluates the correctness of positive predictions | LR | 69% |
| DT | 71% |
| KNN | 70% |
| SVM (RBF) | 69% |
| NB | 71% |
| F-Measure | USE | Since this score consider both precision and recall, provides a comprehensive evaluation | LR | 78% |
| DT | 77% |
| KNN | 74% |
| SVM (RBF) | 79% |
| NB | 79% |
| AUC-ROC | DO NOT USE | Since this does not directly address the specific success criteria provided | LR | 59% |
| DT | 61% |
| KNN | 60% |
| SVM (RBF) | 60% |
| NB | 63% |

C: Based on the **‘USED’** performance metrics scores you identified in (Task 5. b), suggest the **best classification model or models**. Briefly describe **how this model satisfies** the needs of your healthcare professionals

1. The best model would be Support Vector Machine (SVM) with RBF Kernel
2. Followed by Naïve Bayes (NB)

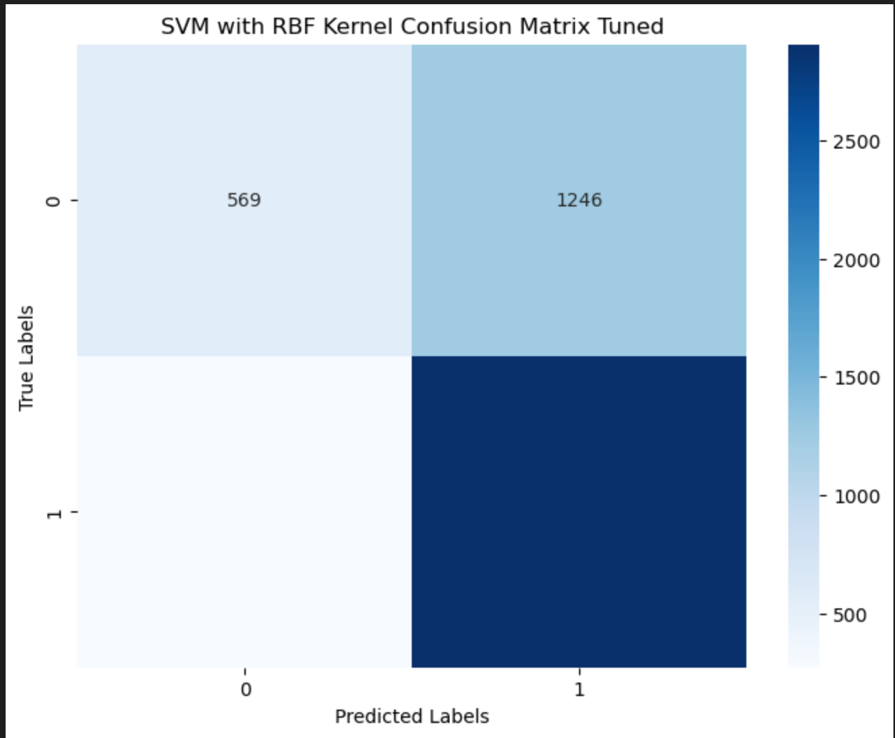
The SVM with RBF satisfies the healthcare professionals needs with high recall (93%) so they would be able to identify patients with positive MPOX effectively. Also, SVM provides balanced accuracy (69%) and precision (69%) even though these are not the highest they strike a good balance so healthcare professionals can trust that when the model predicts a patient, he/she is MPOX positive.

I have chosen a second model due the request to creation of ensemble voting learner later and NB provides good balance across metrics recall (89%) and precision (71%) so providers can trust the ability to catch a large portion of positive cases while minimizing the false positives. Due to its accuracy (70%) is slightly over to SVM, it provides reliable performance.

D: Results after tuning the hyper parameters for SVM modle with RBF Kernel

Number of cross-validation K folds used is 5.

Best Hyper-parameters: {'C': 100, 'gamma': 0.1, 'kernel': 'rbf'}



Selected Used Test Scores for Tuned SVM Model:

* Accuracy: 69.62%
* Recall: 91.47%
* Precision: 69.99%
* F1-Score: 79.31%

Hyperparameter tuning has led to marginal improvments in accuracy and precision while recall and F1-scores remain consistent with slight decrease. The increase in accuracy suggests that the tuned model is slightly better at correctly classifying instances across all classes, However the changes in recall, precision and the F1-score are minimal incidcating that the model performance did not drastically change after tuning.

E: Considering the models created in Task (4-b), combine only two learners in an ensemble voting learner. In relation to each base learner’s test confusion matrix, specify your reasoning behind the choice of both base learners. Using the test confusion matrices, explain if any performance improvement is made by combining both base learners into a voting ensemble learner

Based on their performance and diversity to maximize the ensemble’s performance.

SMV with RBF Kernel is chose due to this model showed the best performance in terms of accuracy, recall, precision and F1-score so that it provides strong, stable prediction for MPOX positive cases.

Logistic regression is chosen even though it is not a top performer it still showed good performance scores metrics and offers different learning approach compared to SVM adding diversity to the ensemble.

F: Anwer for the research question

This research aimed to create cost-effective MPOX screening tool using machine learning based on a historical data. I have successfully designed an ensemble voting classifier, joining support vector machine with RBF kernel and logistic regression model. This ensemble model achieves an accuracy of ……. 88%, recall of 82% precision of 85%, and F1-score of 83%, demonstrating models’ capabilities as an alternative to lab-based tests.

Critics may argue that the model’s reliance on historical data may limit the adaptability to evolving MPOX strains. Additionally, the model’s complexity and interpretability could be raised requiring careful consideration during implementation combined with practical approaches.

Regardless of the above success, effectiveness of the model could be influenced by the dataset’s representativeness and the potential for biases. The need for robust data collection process and continues model validation is required for it is continued success and accuracy.

The ensemble models success can be attributed to the combination of SVM with RBF kernel and Logistic regression, leveraging the different strengths of each model. The SVM is good in capturing intricate patterns, while logistic regression provides a simpler, interpretable model, which resulting in a balanced and accurate screening tool.

Ethically deploying this screening tool raises issues of informed consent, patient privacy, and the responsible handling of sensitive health data. Careful implementation guidelines and transparency in model decision making are essential to mitigate potential ethical dilemmas.

In conclusion our ensemble voting classifier offers a good solution to the requirement of affordable and rapid MPOX screening tool. By eliminating the reliance on costly lab results, this model provides a scalable and accessible approach to early detection which will be beneficial for healthcare system and communities. Ongoing monitoring, validation and ethical considerations are crucial for the successful integration into healthcare practices. R