

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)

# Task 01: Domain Understanding Classification

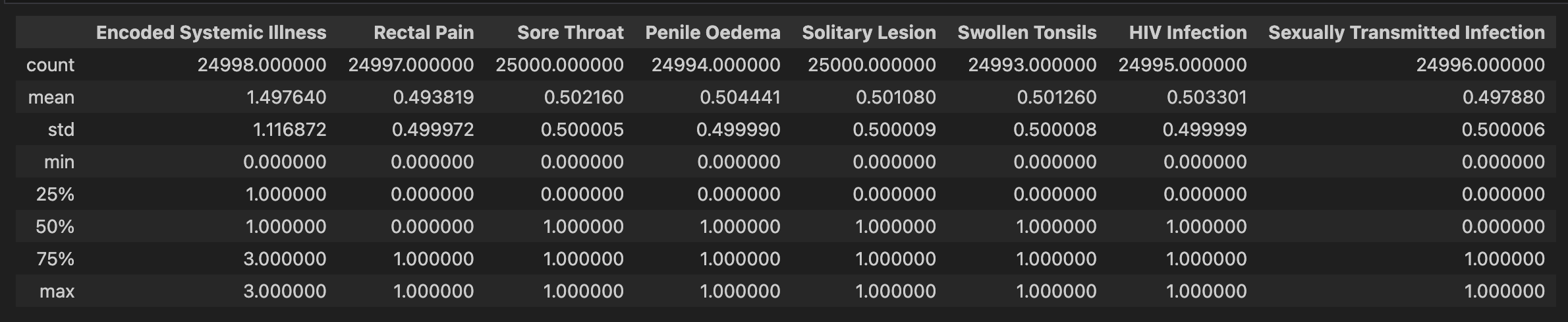
1. Mark the variables logically applicable in the classification modeling of MPOX.

|  |  |
| --- | --- |
| 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

1. Screenshots of the code output

Basic Statistical Descriptions



Measurement Scale Types

A screen shot of a computer screen

Description automatically generated

Distribution of the Result

A graph of a bar graph

Description automatically generated with medium confidence

# Task 03: Cleaning and Transforming the Data

1. Issues identified in the retained data set and possible variables.

|  |  |  |
| --- | --- | --- |
| Dataset or Variable Issue | Name of the Variable | Issue Description |
| Variable Issue | Encoded Systemic Illness | Missing 2 records NaN values |
| Variable Issue | Rectal Pain | Missing 3 Records NaN values |
| Variable Issue | Penile Oedema | Missing 6 records NaN values |
| Variable Issue | Oral Lesions | Missing 15 records of data, 4 records NaN values and 7 records with value No and 4 records with value Yes |
| Variable Issue | Swollen Tonsils | Missing 7 records NaN values |
| Variable Issue | HIV Infection | Missing 5 records NaN values |
| Variable Issue | Age | Issues in 42 records, 36 records NaN values, 1 value with string twenty, having 1 minus value -23 and 4 outliers from 18 to 61 range |
| Variable Issue | Sexually transmitted infection | Missing 4 records NaN values |
| Variable Issue | MPOX PCR Result | Using Negative & Positive as string values |
| Dataset issue | Whole Dataset | 4100 duplicate records found when the ‘Test ID’ column was dropped |

1. Solution to mitigate issues found with justification for using that solution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset or Variable Issue | Name of the Variable | Issue Description | Solution | Justification |
| Variable Issue | Encoded Systemic Illness | Missing 2 records NaN values. | Imputation to replace the NaN with most frequent (mode) value. | When the affected records are smaller in comparison this is a good strategy to avoid loss of data which may be crucial for classification. |
| Variable Issue | Rectal Pain | Missing 3 Records NaN values. | Imputation to replace the NaN with most frequent (mode) value. | When the affected records are smaller in comparison this is a good strategy to avoid loss of data which may be crucial for classification. |
| Variable Issue | Penile Oedema | Missing 6 records NaN values. | Imputation to replace the NaN with most frequent (mode) value. | When the affected records are smaller in comparison this is a good strategy to avoid loss of data which may be crucial for classification. |
| Variable Issue | Oral Lesions | Missing 15 records of data, 4 records NaN values and 7 records with value No and 4 records with value Yes  2 string values 0.0 converted to int. | Imputation to replace 15 records of NaN with most frequent (mode) value. Converted the string values to numeric with Yes = 1 and No = 0 also converted the 0.0 string value to 0 and 1.0 string to numeric values. | When the affected records are smaller in comparison this is a good strategy to avoid loss of data which may be crucial for classification. Converted the string to numerical to facilitate the machine learning inputs. To get higher quality dataset. |
| Variable Issue | Swollen Tonsils | Missing 7 records NaN values. | Imputation to replace the NaN with most frequent (mode) value. | When the affected records are smaller in comparison this is a good strategy to avoid loss of data which may be crucial for classification. |
| Variable Issue | HIV Infection | Missing 5 records NaN values. | Imputation to replace the NaN with most frequent (mode) value. | When the affected records are smaller in comparison this is a good strategy to avoid loss of data which may be crucial for classification. |
| Variable Issue | Age | Issues in 42 records, 36 records NaN values, 1 value with string twenty, having 1 minus value -23 and 4 outliers from 18 to 61 range. | Imputation to replace 36 records the NaN with most frequent (mode) value.  String value is converted to numeric value, minus value is converted to plus value using absolute function. 4 records not within the range based on the scatter plot removed from the dataset. | When the affected records are smaller in comparison this is a good strategy to avoid loss of data which may be crucial for classification.  Converted the string to numerical to facilitate the machine learning inputs. To get higher quality dataset. dropped the outliers, this to ensure higher quality data set which will increase the efficiency and accuracy of the models. |
| Variable Issue | Sexually transmitted infection | Missing 4 records NaN values. | Imputation to replace the NaN with most frequent (mode) value. | When the affected records are smaller in comparison this is a good strategy to avoid loss of data which may be crucial for classification. |
| Variable Issue | MPOX PCR Result | Using Negative & Positive as string values. | Converted using the mapping to positive = 1 and negative = 0. | Classification algorithms require numerical data hence to facilitate efficient model training and evaluation converted to the numerical. |
| Dataset issue | Whole Dataset | 4100 duplicate records found when the ‘Test ID’ column was dropped. | No action. | Since the duplicate record appear only after dropping the primary id column ‘Test ID’ this should be a false positive. |

1. Outputs (Before & After)

|  |  |
| --- | --- |
| Before | After |
| 1. Cleaning the missing values using the imputation for all variables | |
|  |  |
| 1. Fixing Oral Lesions Yes/No values and unique value issues | |
|  |  |
| 1. Fixing the Age-related issues | |
|  |  |
| 1. Converting the MPOX PCR Result to numeric values | |
|  |  |
| 1. Duplicate Records | |
|  |  |
| 1. Full Dataset information | |
|  |  |

# Task 04: Create Predictive Classification Models

1. Classification algorithm details

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm Name | Type of Algorithm | Learnable Parameters | Possible Hyper – Parameters | Python package source code to call the algorithm |
| LR | Parametric | CoefficientsWeights for features and intercept | `solver=’liblinear’ `penalty=’L2’` | `sklearn.linear\_model`  `import LogisticRegression` |
| DT | Non-parametric | Splitting rules for nodes, decision rules based on features | `max\_depth=5` `min\_samples\_split=4` `min\_samples\_leaf=2` `criterion=’entropy` | `sklearn.tree`  `import DecisionTreeClassifier` |
| KNN | Non-parametric | Training Dataset | `n\_neighbors=2` `metric=’minkowski` | `sklearn.neighbors`  `import KneighborsClassifier` |
| SVM (RBF) | Non-parametric | Support vectors, weights for features | `C=2`  `kernal=’rbf’` | `sklearn.svm`  `import SVC` |
| NB | Parametric | Prior probabilities for classes and Features | `var\_somoothing=1e-10` | `sklearn.naive\_bayes`  `import GaussianNB` |

1. Data shape function output
2. List of features used to train the model and the X and y data shape, in this case Y data shape does not show the column count.

A screenshot of a computer program

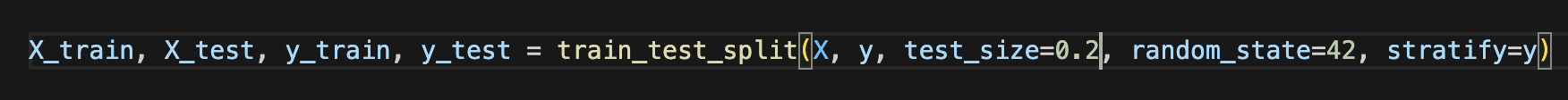
Description automatically generated

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

The choice of the training-test split ratio is a critical aspect of model evaluation, greatly influencing the model's ability to generalize to unseen data (Provost & Fawcett, 2013). In this scenario, a split ratio of 80% training data and 20% testing data was selected for several reasons. Firstly, allocating a larger proportion of the data to training (80%) allows the model to learn more intricate patterns and relationships within the dataset, potentially enhancing its robustness. Secondly, with 20% reserved for testing, there is ample unseen data to comprehensively assess the model's performance without compromising the adequacy of the training set.

This split ratio strikes a balance between the size of the training and testing sets, aiming to mitigate issues such as overfitting or underfitting. Such a decision aligns with established best practices in machine learning, where common split ratios include 70/30, 80/20, or 75/25, tailored to the size and characteristics of the dataset (James et al., 2013, p. 186). The 80/20 split, though slightly different from the stated 75/25, remains widely recommended and has demonstrated success in various studies. It ensures a sufficiently large training set for effective model learning while reserving a substantial test set for robust model evaluation

1. Code line from the source code.



1. MPOX positive-negative ratios (positive = 1 and negative = 0)

A black screen with white text

Description automatically generated

# 

# Task 05: How Good Is the Model

1. Test confusion matrix for each trained model (output screenshots)

|  |  |
| --- | --- |
| Logistic Regression | Decision Tree Classifier |
| A screenshot of a graph  Description automatically generated |  |
| K-nearest Neighbor | Support Vector machines with RBG |
|  |  |
| Naïve Bayes |  |
|  |  |

1. Five different classification evaluation metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric Name | “USE” or “DO NOT USE” | Justification in relation to the success criteria | Model Name | Metric Score |
| Accuracy | DO NOT USE | Since the model should predict as many MPOX positive subjects as possible and it could treat false positives and false negatives equally which is not align with the success criteria. | LR | 0.6726 |
| DT | 0.6758 |
| KNN | 0.6576 |
| SVM (RBF) | 0.6882 |
| NB | 0.6728 |
| Recall | USE | Since importance of correctly detecting MPOX positive subjects by focusing on the true positive rate, which aligns with the success criteria | LR | 0.8966 |
| DT | 0.8456 |
| KNN | 0.7831 |
| SVM (RBF) | 0.9179 |
| NB | 0.9047 |
| Precision | USE | Since this evaluates the correctness of positive predictions and to ensure the predicted MPOX positive subjects are correct, and this align with the success criteria. | LR | 0.6856 |
| DT | 0.7942 |
| KNN | 0.7091 |
| SVM (RBF) | 0.6923 |
| NB | 0.6835 |
| F-Measure | USE | Since this score consider both precision and recall, provides a comprehensive evaluation. Since the success criteria involved with both recall and precision this provides a comprehensive look. | LR | 0.6576 |
| DT | 0.7685 |
| KNN | 0.7443 |
| SVM (RBF) | 0.7893 |
| NB | 0.7787 |
| AUC-ROC | DO NOT USE | Since this does not directly address the specific success criteria provided | LR | 0.5885 |
| DT | 0.6120 |
| KNN | 0.6104 |
| SVM (RBF) | 0.6020 |
| NB | 0.5857 |

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)

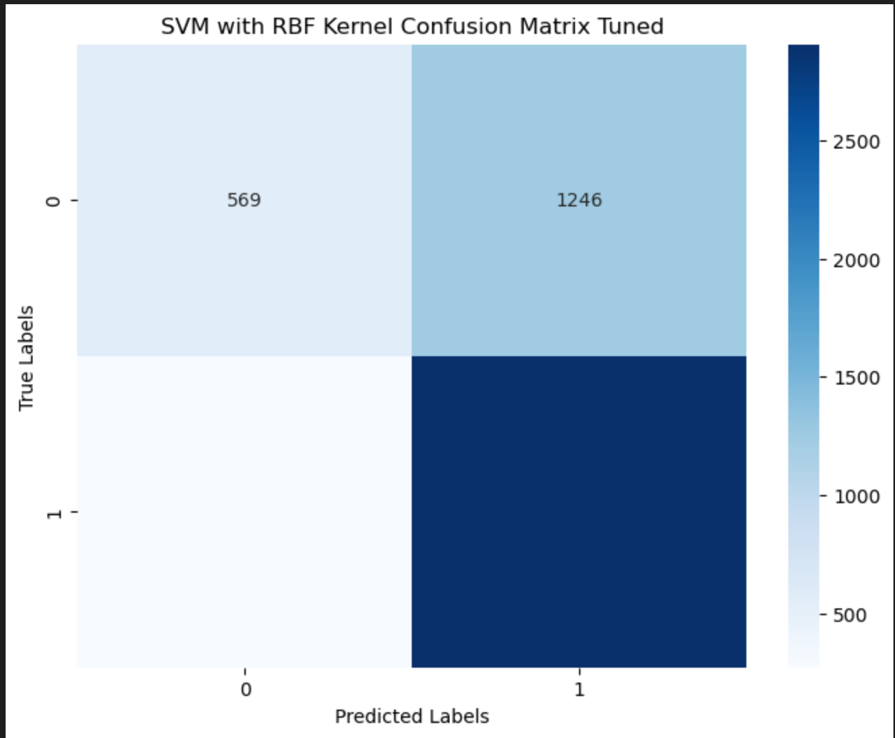
The SVM with RBF satisfies the healthcare professionals needs with high recall (93%) so they would be able to identify patients with positive MPOX effectively. Also, SVM provides balanced accuracy (69%) and precision (69%) even though these are not the highest they strike a good balance so healthcare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