PLATFORMS FOR DATA ANALYTICS PROJECT

Module Code: B8IT154-TMD1S

Names: Jack Maguire; Sean Carroll; Shane Crowley

Student Numbers: 20026112; 20024157; 20024425

Group: 2

Table of Contents

Framing the business problem 3

Analytics Problem Framing 5

Data 7

Methodology 8

Model Building 9

Deployment 12

Conclusion 14

References 15

Appendixes 18

Framing the business problem

The “Monkey-Pox PATIENTS Dataset” (Ahmed, 2022) from Kaggle.com was the chosen open data source for the basis of this project. The dataset is a synthetic collection of data based on the article by Patel et al. (2022). The research, carried out between May and July 2022, is based on patients tested for monkeypox in south London in a regional high consequence infectious disease centre. Considering this outbreak there is a need to analyse this dataset for an efficient solution.

The primary business problem is to use the dataset to create a model that will accurately classify whether patients will test positive or negative for monkeypox based on previous testing for nine symptoms. The model aims to reduce the number of false positives and false negatives in the prediction of monkeypox amongst the patients tested in the future.

Through the analysis of the data, it is the hope that it will enhance the stakeholders’ understanding of Monkeypox, improve prevention strategies, and optimise healthcare interventions. The stakeholders involved include healthcare providers and workers, public health authorities (e.g., the NHS), researchers and scientists, the patients, insurance providers (e.g., AXA), and government agencies (e.g., Department of Health and Social Care).

The dataset is well-suited for an analytics solution as an analysis of the collected data can provide valuable insights into the disease's patterns, risk factors, and potential interventions (Pastorino et al., 2019). An analysis of the problem could help in early detection of monkeypox, improve resource allocation, and enhance overall healthcare strategies.

Addressing the ongoing monkeypox outbreak within the dataset certain constraints must be considered. The business problem is informed by constraints discussed by Kimball & Ross (2013, p. 341-342). An initiative should be developed that is ethically sound, not breach privacy regulations, and should be adaptable to varying healthcare infrastructures.

In terms of a business perspective there are many benefits (Batko & Ślęzak, 2022). These include:

1. Early detection of monkeypox.
2. Improving the quality of healthcare services.
3. Improved treatment protocols that will lead to better patient outcomes.
4. Public health education about monkeypox.
5. A reduction in healthcare costs.
6. Supporting the work of medical personnel.

Analytics Problem Framing

The analytics problem at hand is to design and implement a machine learning solution that effectively identifies cases of monkeypox within “Monkey-Pox PATIENTS Dataset.” This solution must navigate significant constraints, including compliance with data privacy regulations, ethical considerations, and should be adaptable across diverse healthcare infrastructures.

The ABT contains the following attributes and outputs:

- **Patient_ID:** Unique identifier for each patient record.
- **Systemic_Illness:** Type of illness (None/Fever/Swollen Lymph Nodes/Muscle Aches and Pain).
- **Rectal_Pain:** Do they have rectal pain (True/False).
- **Sore_Throat:** Do they have sore throat (True/False).
- **Penile_Oedema:** Do they have penile edema (True/False).
- **Oral_Lesions:** Do they have oral lesions (True/False).
- **Solitary_Lesion:** Do they have solitary lesions (True/False).
- **Swollen_Tonsils:** Do they have swollen tonsils (True/False).
- **HIV_Infection:** Do they have HIV infection (True/False).
- **Sexually_Transmitted_Infection:** Do they have any sexually transmitted infection (True/False).
- **Monkeypox:** Do they have MonkeyPox (Positive) or not (Negative).

The assumptions in relation to the “Monkey-Pox PATIENTS Dataset” (Ahmed, 2022) are that it accurately represents the 2022 monkeypox outbreak in south London. Chosen features, e.g., ‘Oral Lesions,’ are relevant. Patient information is consistent throughout, and factors influencing monkeypox remain stable over time. The dataset is coherent with privacy regulations, ensuring anonymity. The features in the dataset certainly represent key components, and insights from the dataset can be scalable.

The definitions of the key metrics of success, based on the dimensions of data quality by Teradata (2023) include:

Representational Accuracy: The “Monkey-Pox PATIENTS Dataset” (Ahmed, 2022) should reflect the 2022 monkeypox outbreak in south London accurately.

Relevant Features: The chosen features should capture the monkeypox outbreak.

Data Consistency: The consistency of patient information throughout the dataset.

Privacy Compliance: Ensure patient anonymity.

Relevance: Give insights into real-world issues surrounding the monkeypox outbreak.

Data

The Monkeypox dataset consists of over twenty-five thousand instances and contains eleven features. The features include ‘Patient_id,’ ‘Systemic Illness,’ ‘Rectal Pain,’ ‘Sore Throat,’ ‘Penile Oedema,’ ‘Oral Lesions,’ ‘Solitary Lesion,’ ‘Swollen Tonsils,’ ‘HIV Infection,’ ‘STI,’ and ‘Monkeypox.’

The dataset is a classification dataset. “Classification is the kind of learning where the algorithm needs to map the new data that is obtained to any one of the 2 classes that we have in our dataset” (Dublin Business School, 2023). Each of the eleven features can be classified as categorical data. The ‘Patient_id’ feature is a list of the patients’ unique ID. The ‘Systemic Illness’ feature classifies patients as having one of three illnesses or no illness. Patients are categorised as testing either ‘positive’ or ‘negative’ in the ‘Monkeypox’ feature. In each of the other eight features patients’ testing for symptoms is classified as being ‘true’ or ‘false.’

The x variables for analysis were ‘Systemic Illness,’ ‘Rectal Pain,’ ‘Sore Throat,’ ‘Penile Oedema,’ ‘Oral Lesions,’ ‘Solitary Lesion,’ ‘Swollen Tonsils,’ ‘HIV Infection,’ ‘STI,’and the y variable for analysis was ‘Monkeypox.’

According to McKinney (2022, p. 203), when preparing data there may be the need for data cleaning, correcting missing values, and/or transforming variables. The “Monkey-Pox PATIENTS Dataset” does not contain any missing values or incorrect values. Therefore, the dataset is brought straight to the modelling stage as there was no need to clean the data, correct missing values, or transform variables.

Methodology

To predict the Monkeypox outcomes an array of classification models were used which were chosen from a scientific article 'Accurate Machine Learning Algorithms for Monkeypox based on COVID-19' (Sri, Chowdary, Navya, & Reeja, 2023). This article provides a roadmap for model selection, guiding our focus on the exploration of effective machine learning algorithms for Monkeypox.

The featured scientific work shows an accuracy metric of 94% to 99% across various models employed in predicting Monkeypox outcomes. This robust performance highlights the suitability of machine learning algorithms for predictive modelling in the specific context of Monkeypox.

Based on the article's emphasis on similarities in symptoms between COVID-19 and Monkeypox, our approach uses proven machine learning algorithms from COVID-19 predictions. Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines are chosen for their accuracy, given the shared symptoms between the two diseases.

In the modelling phase, RapidMiner was used to execute the selected models. The Auto Model function within RapidMiner was important in streamlining the modelling process, automating the preprocessing steps.

To strengthen the research, an additional model (Naive Bayes) was included, providing an additional set of outcomes, and diversifying our predictive capabilities. RapidMiner ran the selected models, splitting the training and test data in a 60:40 ratio, allowing for robust evaluation and analysis.

Model Building

For the model building phase, below are the steps that were carried out in Rapid Miner's Auto Model feature to obtain each model's results:

1. **<u>Data Loading and Target Variable Selection:</u>**

   The analysis was initiated by loading the Monkeypox dataset into RapidMiner and identifying the target variable, which, in this case, is the column representing the Monkeypox outcome (see Appendix A).

2. **<u>Feature Selection:</u>**

   Following data loading, relevant input features were selected. It's noteworthy that RapidMiner flagged the "Patient ID" column as having minimal correlation with the outcome. Consequently, this column was excluded from the subsequent modelling steps (see Appendix B).

3. **<u>Model Selection and Execution:</u>**

   Subsequently, the machine learning models chosen were deemed suitable for Monkeypox prediction. With the feature selection in place, the selected models were executed, and results obtained (see Appendix C).

In accordance with Tigercchiold (2022) the evaluation metrics utilised included ‘Accuracy,’ ‘Precision,’ ‘Recall,’ and the ‘F1 Score,’ each providing a nuanced perspective on the models' effectiveness.

Accuracy gauges the proportion of correct predictions out of the total predictions made, providing a holistic view of model performance.

Precision focuses on the accuracy of positive predictions, revealing how many predicted positive outcomes were genuinely positive.

Recall delves into the model's ability to correctly identify all actual positive instances, offering insights into sensitivity.

The **F1 Score** serves as a balanced measure, incorporating both precision and recall, providing a comprehensive evaluation.

An approach was adapted based on the article's authors and opted for ‘Accuracy’ as an evaluation metric. The rationale, as articulated in the article, lies in the simplicity of accuracy for comparing various machine learning models (Sri, Chowdary, Navya, & Reeja, 2023). Additionally, accuracy proves to be a fitting metric when dealing with a balanced dataset, where the instances for each class (in this case positive and negative) are similar.

F1 score was also a focus and used as an evaluation metric. In healthcare, it's essential to strike a balance between overdiagnosis and underdiagnosis. The F1 score provides a balanced measure of how well a model is performing in terms of identifying positive cases while minimising both false positives and false negatives (Hicks, et al., 2022). This is crucial for building trust in the model among healthcare professionals.

Table 1:

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	67.8%	71.6%	82.0%	76.5%
Decision Trees	63.6%	63.6%	100% ⭐	77.7%
Random Forest	66.4%	70.0%	82.6%	75.8%
Support Vector Machines	69.2% ⭐	71.7% ⭐	86.1%	78.2% ⭐
Naive Bayes	67.7%	71.5%	82.0%	76.4%

In evaluation, the assessment table (*See Table 1*) scored Support Vector Machines as the standout performer, showing the highest accuracy at 69.2% and F1 Score at 78.2%. However, it is crucial to note that across all models, the achieved accuracy did not surpass 69.2%. This emphasises a general limitation in the reliability of these models for accurately predicting Monkeypox cases.

Despite Support Vector Machines showing the highest accuracy and F1 Score, it is important to communicate the overarching finding that none of the models, including the top performer, can be trusted to predict Monkeypox with high certainty.

Deployment

Machine learning can be used for clinical decisions in relation to healthcare, so it is paramount to ensure patient safety, for early diagnosis and reduce onward transmission when thinking of the deployment of the model.

There are a number of factors that should be considered when deciding to deploy a machine learning model. In terms of healthcare, being accurate with the prediction is crucial.

Machine learning approaches to diagnosis are purely associative, identifying diseases (in this case monkeypox) that are strongly correlated to a patient with symptoms (Mohd, 2022). Therefore, there is a risk that the inability to disentangle correlation from causation can result in suboptimal or dangerous diagnosis (inaccurate results).

The accuracy obtained during the model evaluation overall fell below the desired threshold for healthcare applications. An accuracy of 69.2% and F1-score of 78.2% does not suffice when trying to accurately predict an outcome for a serious condition like Monkeypox. The accuracy should aim for 99% (Mantelakis, 2021) to enable reliable outcomes.

Therefore the model as a whole would have to be re-evaluated.

Ways of improving accuracy of the model include Additional Data Collection such as more inputs within the data (Bohr, 2020). The more data collected, the more precise the model will be leading to more accurate results. Another way to improve accuracy is ensemble methods which involves combining several models to take one reliable model (Dietterich, 2000) which could also be an option.

Conclusion

To conclude, the analysis of the Monkeypox data between May and July 2022 using Machine Learning techniques, has provided for interesting insights. The goal of our model was to accurately predict the outcome of a patient who had symptoms which had correlations to the disease Monkeypox. Our machine learning algorithms were chosen in relation to an article about Covid-19 where similarities were made with Monkeypox.

They were as follows, Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines. Through the use of RapidMiner, SVM was the standout performer with an accuracy of 69.2% and an F1-score of 78.2%. We focused on accuracy as it provides for a comprehension view of model performance, as well as F1-score which integrates precision and recall to gain a better understanding of model performance which is important for healthcare application.

From the results, we gathered that none of the models were sufficient enough in accurately predicting Monkeypox.

The results fell short of the expected 99% accuracy that healthcare demands. Therefore we came to the conclusion that our model would have to be re-evaluated.

References

Ahmed, M. (2022). Monkey-Pox PATIENTS Dataset. [Dataset] Available at: https://www.kaggle.com/datasets/muhammad4hmed/monkeypox-patients-dataset

Batko, K., & Ślęzak, A. (2022) 'The use of Big Data Analytics in healthcare', *Journal of Big Data*, 9(1), 3. https://doi.org/10.1186/s40537-021-00553-4.

Bohr, A., & Memarzadeh, K. (2020) *Current healthcare, big data, and machine learning*, Artificial Intelligence in Healthcare, Academic Press.

Dietterich, T. G. (2000) *Ensemble Methods in Machine Learning*. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Dublin Business School. (2023) 'SUPERVISED LEARNING', B8IT154: Platforms for Data Analytics. [Lecture Slides] Available at: https://elearning.dbs.ie/mod/folder/view.php?id=1475095 (Accessed: 20 December 2023).

Hicks, S., Strümke, I., Thambawita, V., Hammou, M., Riegler, M., Halvorsen, P., & Parasa, S. (2022) *Evaluation metrics for medical applications of artificial intelligence*. National Center for Biotechnology Information.

Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022) *Significance of machine learning in healthcare: Features, pillars and applications,* International Journal of Intelligent Networks.

Kimball, R. & Ross, M. (2013) *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd edn. [Online] Available at: *https://www.vlebooks.com/Product/Index/1027730?page=0&startBookmarkId=-1* (Accessed 21 December 2023).

Mantelakis, A., Assael, Y., Sorooshian, P., & Khajuria, A. *Machine Learning Demonstrates High Accuracy for Disease Diagnosis* Glob Open. 2021 Jun 24;9(6):e3638. doi: 10.1097/GOX.0000000000003638.

Pastorino, R., De Vito, C., Migliara, G., Glocker, K., Binenbaum, I., Ricciardi, W., & Boccia, S. (2019) ‘Benefits and challenges of Big Data in healthcare: an overview of the European initiatives’, *European Journal of Public Health*, 29(3), 23-27. doi: 10.1093/eurpub/ckz168.

Patel, A., Bilinska, J., Tam, J. C. H., Da Silva Fontoura, D., Mason, C. Y., Daunt, A., ... et al. (2022) ‘Clinical features and novel presentations of human monkeypox in a central London centre during the 2022 outbreak: descriptive case series’, *BMJ*. doi:10.1136/bmj-2022-072410.

Sri, T., Chowdary, L., Navya, & Reeja, R. (2023). Accurate Machine Learning Algorithm for Monkey Pox Based on Covid-19. *International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, 380-383.

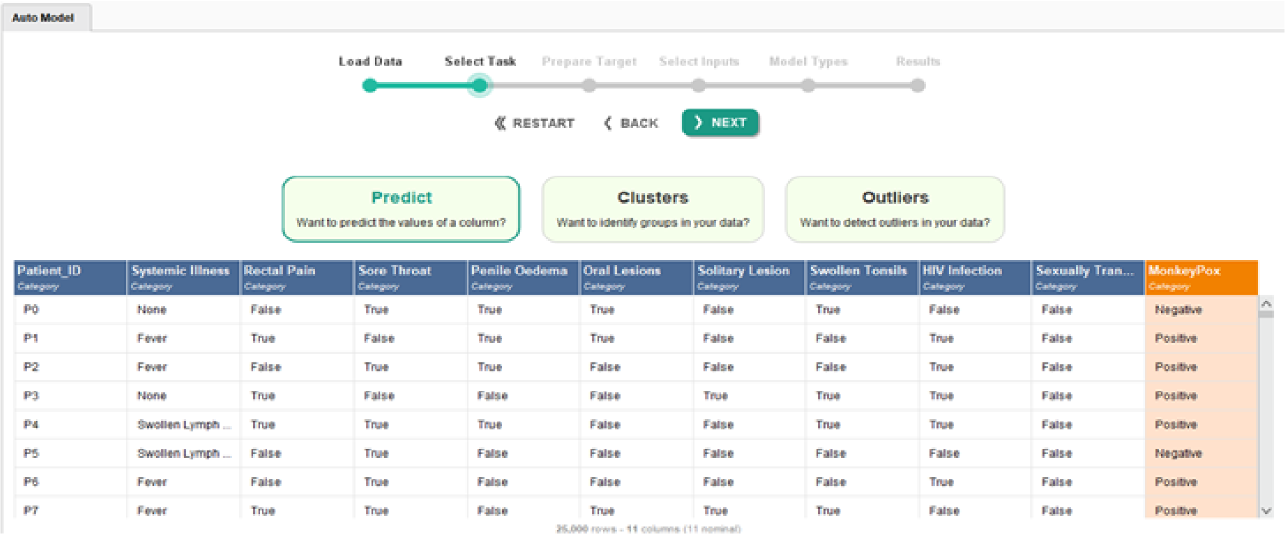
Teradata. (2023). 'What is data quality?’, Teradata Insights, Available at: https://www.teradata.com/insights/data-platform/data-quality-for-informed-decision-making [Online] (Accessed: 22nd December 2023).

Tigercchiold, T. (2022, November 7). *What is Accuracy, Precision, Recall and F1 Score?* Retrieved from Labelf.ai: https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score [Online] (Accessed: 16 December 2023).

Appendixes

Appendix A

Data Loading and Target Variable Selection

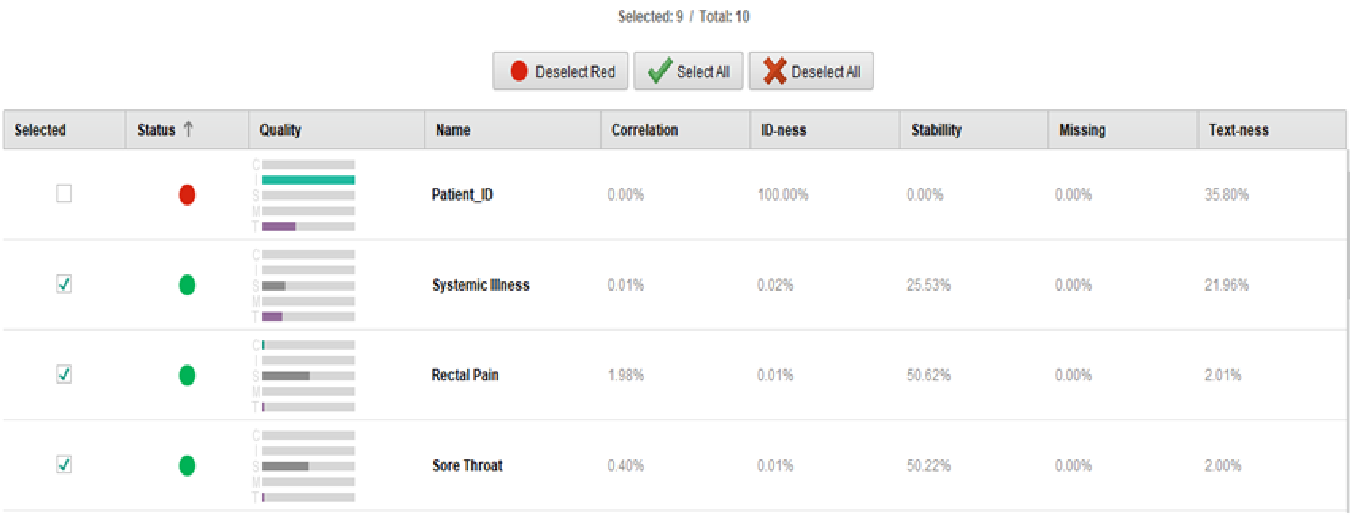

Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT
Predict
Want to predict the values of a column?
Clusters
Want to identify groups in your data?
Outliers
Want to detect outliers in your data?

Patient_ID	Systemic Illness	Rectal Pain	Sore Throat	Penile Oedema	Oral Lesions	Solitary Lesion	Swollen Tonsils	HIV Infection	Sexually Tran...	MonkeyPox
Category	Category	Category	Category	Category	Category	Category	Category	Category	Category	Category
P0	None	False	True	True	True	False	True	False	False	Negative
P1	Fever	True	False	True	True	False	False	True	False	Positive
P2	Fever	False	True	True	False	False	False	True	False	Positive
P3	None	True	False	False	False	True	True	True	False	Positive
P4	Swollen Lymph ...	True	True	True	False	False	True	True	False	Positive
P5	Swollen Lymph ...	False	True	False	False	False	False	False	False	Negative
P6	Fever	False	True	False	False	False	False	True	False	Positive
P7	Fever	True	True	False	True	True	True	False	False	Positive

25,000 rows - 11 columns (11 nominal)

Appendix B

Feature Selection

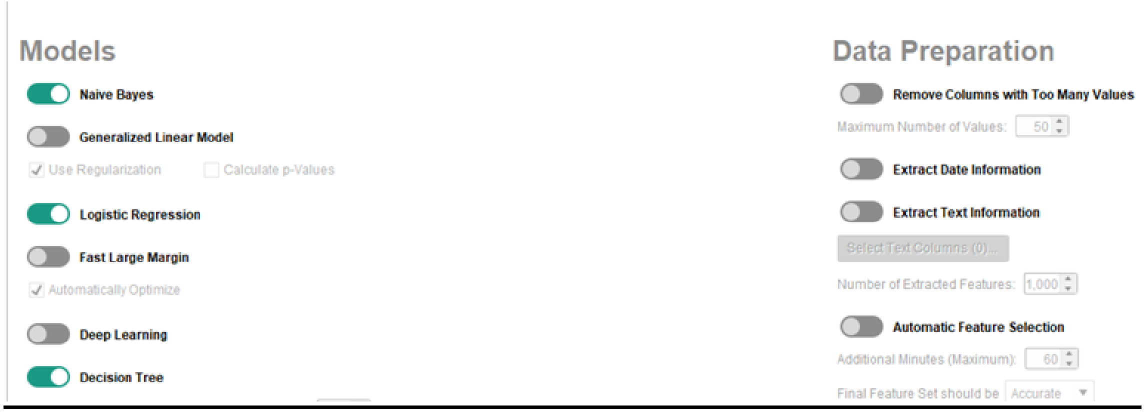


Selected: 9 / Total: 10

Selected	Status ↑	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
□	🔴	C I S M T	Patient_ID	0.00%	100.00%	0.00%	0.00%	35.80%
☑	🟢	C I S M T	Systemic Illness	0.01%	0.02%	25.53%	0.00%	21.96%
☑	🟢	C I S M T	Rectal Pain	1.98%	0.01%	50.62%	0.00%	2.01%
☑	🟢	C I S M T	Sore Throat	0.40%	0.01%	50.22%	0.00%	2.00%

Appendix C

Model Selection and Execution



Models

● Naive Bayes

○ Generalized Linear Model

☑ Use Regularization   □ Calculate p-Values

● Logistic Regression

○ Fast Large Margin

☑ Automatically Optimize

○ Deep Learning

● Decision Tree

Data Preparation

○ Remove Columns with Too Many Values

Maximum Number of Values: 50

○ Extract Date Information

○ Extract Text Information

Select Text Columns (0)...

Number of Extracted Features: 1,000

○ Automatic Feature Selection

Additional Minutes (Maximum): 60

Final Feature Set should be Accurate