Jefferson Veliz

Professor Jairam Jairam

Data mining for Business Analytics

August 8th, 2023

Project #2 Report

My motivation is to develop a model that can distinguish between cases of Monkeypox and other skin conditions. Monkeypox is a rare skin ailment that often resembles other similar issues, making accurate identification crucial. By utilizing machine learning models and images of skin problems, my goal is to create a model that can reliably determine whether an individual has Monkeypox or a different condition. While this project is primarily an academic endeavor, its potential implications are promising. Improved diagnostic accuracy could lead to better patient care and contribute to the prevention of further spread of the illness.

The K-Nearest Neighbors (KNN) model exhibited a distinct performance profile for predicting "Monkeypox" (M) and "Non-Monkeypox" (NM) classes. The weighted average precision stood at 0.83, signifying a well-rounded accuracy in predictions. The model's ability to correctly identify instances from both classes was evident from the weighted average recall of 0.76, indicating a balanced classification capability. The F1-scorewas 0.74 for both classes, showcasing a balanced trade-off between these metrics. Collectively, these findings underscore the model's capacity to achieve equilibrium in predictive accuracy. The Jaccard Accuracy Score of 75.63 further validates the model's proficiency in delivering accurate classifications.

The Gaussian Naive Bayes (GNB) model showcased a distinctive performance pattern in predicting (M) and (NM) classes. The weighted average precision was 0.68, denoting a balanced accuracy in predictions across classes. The model's ability to accurately capture instances from both classes was evident from the weighted average recall of 0.68, reflecting a harmonious classification capability. The F1-score, encapsulating the balance between precision and recall, was 0.67 for both classes, underscoring a stable equilibrium of these metrics. These findings collectively underscore the model's capability to achieve a balanced predictive accuracy. The Jaccard Accuracy Score of 67.51 further substantiates the model's competence in delivering accurate classifications.

The Decision Tree (DT) model demonstrated balanced performance for both (M) and (NM) classes. The weighted average precision was 0.76, indicating an overall accurate prediction across classes. The model's ability to capture instances from both classes was evident from the weighted average recall of 0.76, showcasing a balanced classification ability. The F1-score, which combines precision and recall, reached a value of 0.76 for both classes, highlighting a harmonious balance between the two metrics. These results collectively indicate the model's capability to achieve balanced accuracy in its predictions. Additionally, the Jaccard Accuracy Score of 76.19 further corroborated the model's reliable classification accuracy.

The Sigmoid Kernel SVM model displayed balanced performance for both (M) and (NM) classes. The weighted average precision was 0.78, indicating an accurate overall prediction across classes. The model's ability to capture instances from both classes was evident from the weighted average recall of 0.77, showcasing effective classification. The F1-score attained a value of 0.77 for both classes, signifying a well-rounded performance. These metrics collectively demonstrate the model's proficiency in achieving balanced accuracy in its predictions. Additionally, the Jaccard Accuracy Score of 77.45 reinforced the model's strong classification accuracy.

The RBF SVM model displayed distinct performance outcomes for the (M) and (NM) classes. With a focus on precision, it achieved high precision (1.00) for NM, indicating accurate identification of NM cases. However, precision for M was lower (0.55), suggesting some false positives. In terms of recall, the model excelled in capturing M cases (1.00), while its recall for NM was lower (0.15), indicating challenges in identifying NM cases. The F1-score, a balance between precision and recall, was 0.49, underscoring the overall effectiveness of the model. The model's accuracy was 58.54%, reflecting its ability to make accurate predictions but also highlighting class-specific disparities in performance. This suggests that while the model is adept at certain aspects, there is room for improvement in its performance on both M and NM cases

The Polynomial SVM model exhibited distinct performance characteristics for the (M) and (NM) classes. It achieved high precision (0.82) for NM, indicating accurate identification of NM cases. However, precision was lower (0.65) for M cases. The model excelled in recall for both classes, with 0.99 for NM and 0.50 for M, indicating its ability to effectively capture NM cases. The F1-score reached 0.72. Overall accuracy was 74%, highlighting the model's capacity for accurate predictions across the dataset. At last, the Polynomial SVM model demonstrated class-specific performance variations, with notable strengths in capturing NM cases and opportunities for improvement in predicting M cases.

The Linear SVM model's performance was evaluated using a classification report based on the weighted average of precision, recall, F1-score, and accuracy. The model exhibited balanced precision and recall for both (M) and (NM) classes. With a precision of 0.82 and recall of 0.81, the model demonstrated its ability to make accurate positive predictions for both categories. The F1-score of 0.81 signifies a balance between precision and recall. Overall, the model achieved an accuracy of 81%, indicating its competence in correctly classifying instances from both classes.

The Logistic Regression model's performance in predicting (M) and (NM) classes is detailed in the classification report. It demonstrates balanced precision, recall, and F1-scores for both classes: 0.86 precision, 0.77 recall, and 0.81 F1-score for (M) and 0.78 precision, 0.87 recall, and 0.82 F1-score for "(NM) The model's overall accuracy is 82%, and it has the Jaccard Accuracy Score of 81.65. This indicates the model's efficacy in classifying instances from both classes with reliable accuracy, which can prove valuable in practical monkeypox classification tasks.

The classification report provides an overview of the Random Forest Classifier's performance in predicting two classes: (M) and (NM). The Jaccard Accuracy Score of 0.85.29 indicates the model's overall accuracy in classifying instances into the respective classes. Looking specifically at the "Monkeypox" class, the precision of 0.87 highlights the proportion of correctly predicted "Monkeypox" cases out of all instances classified as "Monkeypox." The recall of 0.84 indicates the model's effectiveness in capturing actual "Monkeypox" cases, considering all instances that are actually "Monkeypox."

Shifting to the "Non-Monkeypox" class, the precision of 0.84 illustrates the proportion of accurately predicted "Non-Monkeypox" cases out of all instances classified as "Non-Monkeypox." The recall of 0.87 signifies the model's capability to identify actual "Non-Monkeypox" cases, considering all instances that truly belong to the "Non-Monkeypox" class. The F1-score of 0.85 for the "Non-Monkeypox" class reflects the balance between precision and recall.In terms of overall accuracy, the RFC achieved an accuracy rate of 85.29%, signifying its ability to correctly classify instances from both classes. The weighted average precision of 0.85 demonstrates a balanced accuracy considering both "Monkeypox" and "Non-Monkeypox" classes.

I've selected the Random Forest classifier as the prime candidate for deployment in our monkeypox image classification project due to its compelling strengths and relevance to real-world challenges. Its consistent accuracy across a variety of monkeypox images assures dependable predictions, which is crucial for distinguishing between healthy and infected diseases accurately. This model's ability to handle different image formats, sizes, and features enhances its usability for both technical and non-technical users. By deploying this model, we Doctors enthusiasts to make informed decisions in managing monkeypox outbreaks .The Random Forest model addresses a critical problem-solving task by enabling the automated identification of monkeypox cases in images, streamlining diagnosis and decision-making processes.

Moreover, the performance of the K-means clustering algorithm applied to the dataset of monkeypox images.The dataset comprises both infected (m) and non-infected (nm) images, each represented by distinct features. The K-means clustering resulted in clusters of images, yielding a purity score of 50.42%. This score indicates that approximately half of the images within each cluster correspond to the same true class. While the achieved score suggests that clusters captured meaningful patterns, there is room for improvement. Challenges in accurately distinguishing between infected and non-infected images are evident, impacting the overall clustering performance. To enhance the results, considerations include visualizing clusters, optimizing the number of clusters (K), exploring alternative clustering algorithms, and investigating feature engineering techniques.

To wrap up, this project aimed to create a machine learning model that can tell the difference between Monkeypox and other skin conditions. We tested different machine learning tools and found that the Random Forest model is the best fit. It's accurate and easy to understand, making it useful for doctors and enthusiasts. By using this model, we can help doctors make better decisions and control the spread of Monkeypox. We also looked at grouping similar images using the K-means method. Although the clusters weren't perfect, they still give us insights. This can guide us to improve the groups by adjusting the features we use.The impact of accurate disease identification on patient health and preventing illness is huge. This project isn't just for the classroom, it's a step towards better healthcare and solving real problems. As I move ahead, I can build on what I’ve learned to make healthcare even better and keep exploring the possibilities of technology in solving important challenges.

**Visualizations:**

A graph with blue rectangular bars

Description automatically generated with medium confidence

**Figure 1:** Dataset of monkeypox and non- monkeypox. NM had more pictures than M.

A blue and white bar graph

Description automatically generated

**Figure 2:** Due to the data being imbalanced, to avoid bias, I made the M and NM equal.

A graph with lines and numbers

Description automatically generated

**Figure 3:** We found out that the distortion score elbow for k means clustering is two.

A graph of different colored bars

Description automatically generated with medium confidence

**Figure 4:** Random forest is the best working machine learning model with a 85% accuracy compared to the others.

**References:**

**Professor’s notes**

<https://chat.openai.com> – To fix grammar and debug code. I learned to ask chatgpt the right questions. I used it as a tutor to refine my skills.