Predicting Erythemato-Squamous Diseases (ESDs) Types

Team: Fabulous Five

Soliyana Ahmed

Hana Awad

Mia Robinson

Sai Rakesh Duggineni

Tejaswini Madivada

INTRODUCTION

Skin diseases are defined as conditions that affect your skin. Erythemato-squamous diseases (ESDs) are common skin diseases that are divided into six different categories. The categories are psoriasis, lichen planus, pityriasis rosea, pityriasis rubra pilaris, seborrheic dermatitis, and chronic/atopic dermatitis. Diagnosing erythemato-squamous diseases poses a challenging issue in dermatology due to the shared clinical characteristics of erythema and scaling, with minimal distinctions among them (Guvenir & Emeksiz, 2000).

Psoriasis

Figure 1: Psoriasis on of darker skin (left) and lighter skin (right)

Psoriasis is a chronic skin condition characterized by an itchy, scaly rash, typically found on the knees, elbows, trunk, and scalp. While there is no cure, it manifests in various ways, ranging from dandruff-like scaling to extensive eruptions. The rash can vary in color, appearing as shades of purple with gray scales on darker skin and pink or red with silver scales on lighter skin **(Figure 1)**. Other symptoms include small scaling spots, dry and cracked skin that may bleed, itching, burning, soreness, and cyclic rashes that flare up for a few weeks or months before subsiding (Mayo Clinic, 2022). Histopathological features of psoriasis include epidermal hyperplasia with club-shaped rete, reduced or nonexistent granular cell layer, and thinning of the suprapapillary plate, amongst others (Lauren & Brinster., 2019).

Lichen planus

Figure 2: Lichen planus flat and raised bumps on wrists

Lichen planus is a condition affecting the skin, hair, nails, mouth, and genitals. Its skin symptoms typically involve the development of purple, itchy, flat bumps. Symptoms vary based on the affected area, such as shiny bumps on the inner forearms, wrists **(figure 2)**, or ankles, rash lines from scratching, lacy white patches on the tongue or inside the cheeks, itchiness, painful sores in the mouth or genitals, occasional hair loss, nail scarring or loss, and dark lines on nails (Mayo Clinic, 2023). The classic histopathological features of lichen planus include the six Ps of LP which are purple, polygon, flat-topped pruritic papules, planar, and plaques (Arnold & Krishnamurthy., 2016).

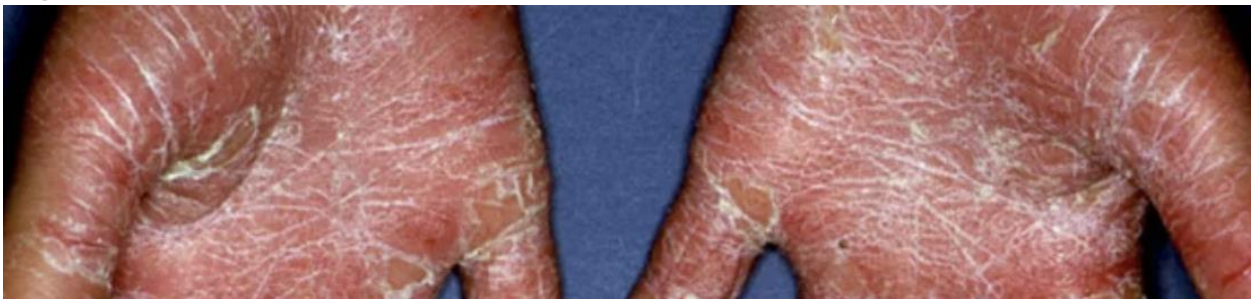
Pityriasis Rosea

Figure 3: Pityriasis rosea herald patch

Pityriasis rosea typically starts as an oval spot **(Figure 3)**, known as a herald patch, on the face, chest, abdomen, or back. These spots can be as big as 4 inches long. Smaller spots that look like drooping pine tree branches may also develop. Oftentimes itching may also be a symptom. Particularly for brown or darker skin individuals, temporary skin discoloration may occur, either darker or lighter than usual (post-inflammatory hyperpigmentation or hypopigmentation) (Mayo Clinic, 2022). Histopathological features of pityriasis rosea are erythrocytes in the papillary dermis/epidermis, papillary dermis collagen homogenization, and reduction or nonexistence of granular cell layer (Panizzon & Block., 1982)

Pityriasis rubra pilaris (PRP)

Figure 4: PRP scaly patches on the palm of hands



PRP encompasses a range of skin conditions characterized by chronic scaling and inflammation. Individuals with PRP exhibit reddish, scaly patches that can appear throughout the body or be localized to specific areas **(figure 4).** The histopathological features of PRP include alternating parakeratosis and orthokeratosis in both horizontal and vertical directions, thick suprapapillary plates, broad rete ridges, focal/confluent hypergranulosis, narrow dermal papillae, and sparse superficial perivascular infiltration. Psoriasis does exhibit some histologic similarities with PTP, however, there are some differentiating features. These include the presence of neutrophils in mounds of parakeratosis, thin rete ridges, thin suprapapillary plates, broad dermal papillae, and mixed inflammatory cell infiltrates with varying densities (Soeprono., 1986).

Seborrheic dermatitis

Figure 5: Seborrheic dermatitis on scalp/ forehead

Seborrheic dermatitis is a common skin condition primarily affecting the scalp. Common signs and symptoms include areas of oily skin with flaky white or yellow scales or crust, observed on the scalp **(figure 5)**, face, sides of the nose, eyebrows, ears, eyelids, chest, armpits, groin, or under the breasts. Additional features encompass flaking skin (dandruff) on the scalp, hair, eyebrows, beard, or mustache, along with itchiness (pruritus). The rash may appear darker or lighter in individuals with brown or Black skin and redder in those with white skin. Some cases may present as a ring-shaped (annular) rash, known as petaloid seborrheic dermatitis (Mayo

Clinic, 2022). The histopathological features of seborrheic dermatitis include acanthosis, focal spongiosis, infundibular and surface epidermis displaying a superficial perivascular infiltrate of lymphocytes, and focal parakeratosis (Tucker & Masood., 2023).

Chronic/ Atopic Dermatitis

Figure 6: Eczema rash on arm

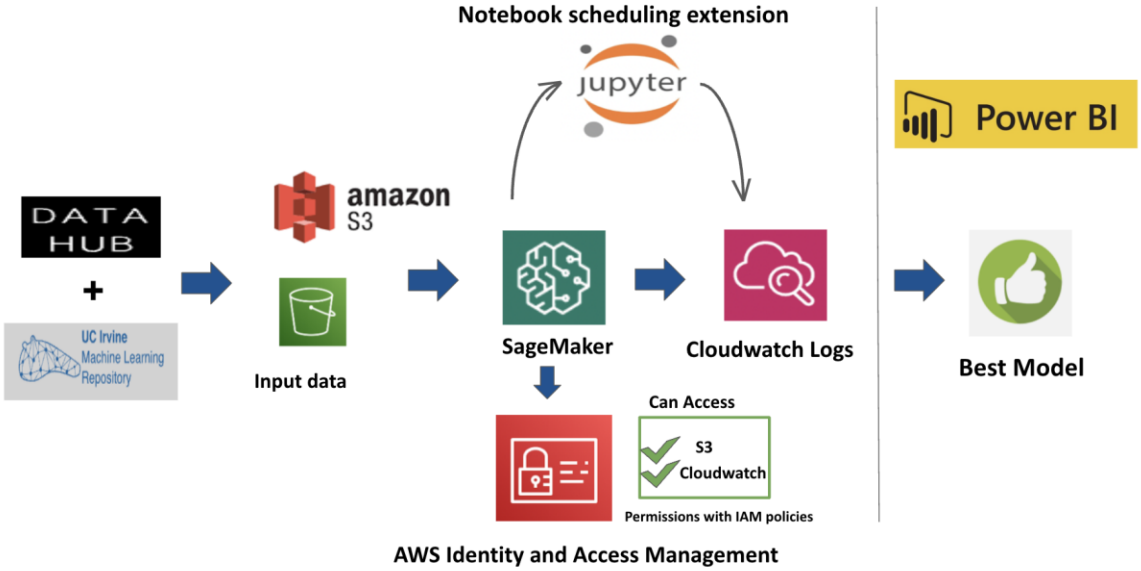
Atopic dermatitis, or eczema, is a condition characterized by dry, inflamed, and itchy skin **(Figure 6)**. While it is common in young children, it can affect individuals of any age. Symptoms of atopic dermatitis vary and can be seen anywhere on the body. These symptoms include dry and cracked skin, a rash on swollen skin with color variations based on skin tone, itchiness, thickened skin, raw and sensitive skin from scratching, small raised bumps on black or brown skin, crusting and oozing, and darkening of the skin around the eyes (Mayo Clinic., 2023).

Histopathological features of eczema include the widening of spaces between keratinocytes and elongation of intercellular bridges, formation of parakeratosis above the spongiosis areas. Also, there is edema and a superficial perivascular infiltrate containing lymphocytes, histiocytes, and occasional neutrophils and eosinophils (DermNet., No date).

These skin diseases are typically diagnosed through a combination of clinical assessment, medical history evaluation, and, if needed, diagnostic tests. Dermatologists visually examine the skin for characteristic symptoms, consider the patient's medical history, and may perform a skin biopsy, culture, utilize diascopy, and dermoscopy, amongst other tests to confirm the specific disease type. Response to treatment can also provide diagnostic clues. The diagnostic approach varies based on the patient's symptoms and the type of Erythemato-Squamous disease suspected (Cleveland Clinic, no date).

Our project aims to create a predictive model for accurately identifying and classifying distinct types of Erythemato-Squamous diseases using clinical and histological data. Healthcare professionals can utilize this tool to aid in disease diagnosis, potentially reducing misdiagnoses, accelerating diagnosis, optimizing healthcare resource allocation, and ultimately improving patient outcomes through early and precise interventions.

ARCHITECTURAL FRAMEWORK



Data Collection:

Datasets are sourced from reputable source UCI Irvine website, ensuring diverse and high-quality data for our machine learning project.

Data Storage in S3 Bucket:

The collected datasets are stored in an Amazon S3 bucket, a scalable and durable object storage service. S3 provides a centralized and secure repository for our data, facilitating easy access and retrieval.

SageMaker Notebook Instance:

Amazon SageMaker notebook instances are utilized for data exploration, model development, and experimentation. These instances provide a collaborative environment with adjustable computing resources based on project requirements.

Model Training and Deployment:

Machine learning models are trained using SageMaker's built-in algorithms or custom code. Trained models can be deployed as endpoints for real-time predictions, facilitating integration into applications.

CloudWatch Logs for Monitoring:

CloudWatch Logs capture detailed logs and metrics from the SageMaker notebook instance and model training processes. This monitoring mechanism provides insights into the performance and health of the system, aiding in debugging and optimization.

Security and IAM Roles:

IAM roles are employed to grant the necessary permissions for the SageMaker notebook instance to access S3 buckets and other AWS services securely. This ensures a controlled and secure environment for data processing.

Extract Transform Load

Data undergoes an ETL process, involving extraction from S3, transformation as needed for model training, and loading into the SageMaker notebook for analysis. This ensures that the data is in a suitable format for effective machine learning model development.

Our project incorporates a robust ETL process to prepare and preprocess datasets for machine learning. This process involves:

Extraction:

Datasets are extracted from external sources such as Data Hub and UCI Irvine websites. These sources provide diverse and relevant data for our machine learning tasks.

Transformation:

Extracted data undergoes transformation to make it suitable for model training. This includes handling missing values, encoding categorical variables, and scaling features as needed.

Loading:

Transformed data is loaded into the Amazon S3 bucket, a scalable and secure storage solution. This step ensures centralized data storage accessible by SageMaker for analysis and model development. By implementing a robust ETL process, we ensure that our machine learning models receive clean, structured, and well-prepared data, laying the foundation for successful model training and deployment.

Data Quality Checks:

Robust data quality checks are integrated into the transformation phase, guaranteeing the integrity and accuracy of our datasets. These checks identify and address any inconsistencies, ensuring the reliability of the data.

Handling Imbalanced Data:

Special attention is given to addressing imbalanced data concerns. Through strategic techniques such as oversampling, undersampling, or specific algorithmic approaches, we ensure a balanced distribution of classes for more accurate model training.

Data Versioning:

We've incorporated a robust data versioning strategy, allowing us to track changes and updates to our datasets over time. This ensures traceability and reproducibility, essential for maintaining data consistency in our machine learning pipeline.

Metadata Logging:

Our ETL process includes comprehensive metadata logging. This captures vital details such as the data source, applied transformation steps, and key data quality metrics. This metadata serves as a valuable documentation resource for understanding data lineage.

Scalability Considerations:

To handle large datasets efficiently, our ETL process incorporates scalable solutions. Techniques such as parallel processing and distributed computing ensure that our system can effectively manage the volume of data we're working with.

Error Handling and Logging:

Robust error handling mechanisms and logging strategies are embedded in our ETL process. Any issues or anomalies are systematically logged, providing a clear trail for investigation and troubleshooting.

Logging and Monitoring with CloudWatch:

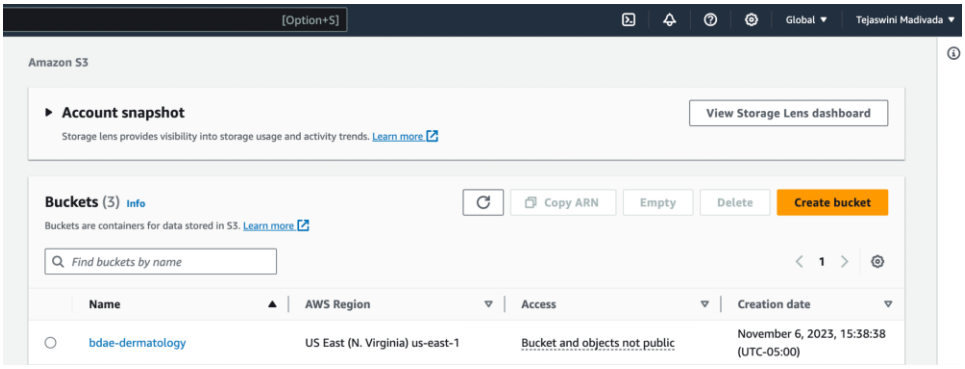
We configure CloudWatch to capture and monitor logs from the SageMaker notebook instances. This step is essential for tracking execution details, performance, and any potential issues throughout the project. This architecture offers a seamless data workflow, enabling us to access, process, and analyze datasets with a powerful combination of Amazon S3 for storage and SageMaker for data processing and model development.

Leveraging the resources and data management capabilities provided by this architecture, we are actively engaged in the critical task of developing and predicting the best machine learning model for our project. Through rigorous experimentation and analysis, we aim to identify the model that offers the highest predictive accuracy and relevance to our specific use case. This

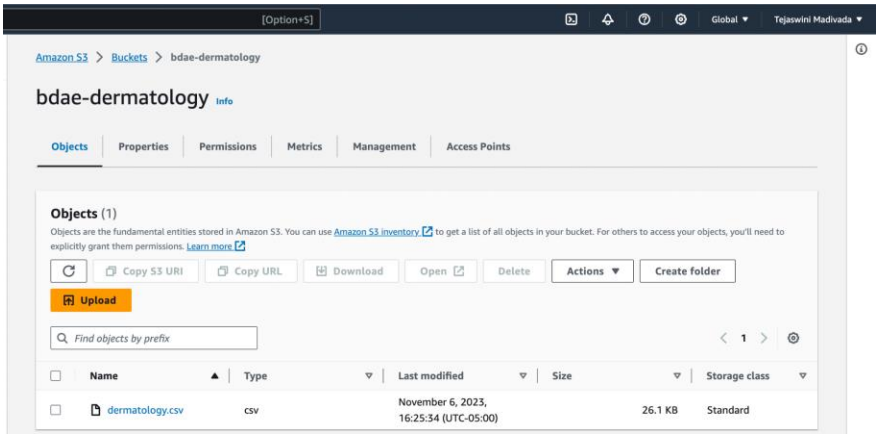
architecture ensures efficient data management, secure storage, robust monitoring, and a clear path toward determining the most effective model for our machine learning project.

Amazon S3 was chosen for its scalable, cost-effective data storage and seamless integration with AWS services, enabling secure, compliant, and accessible data management essential for the project's success.

Created an S3 Bucket to store input data.

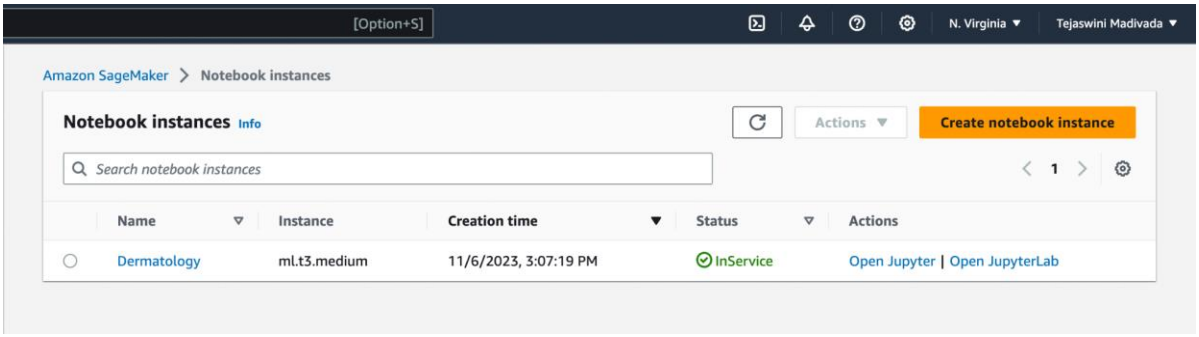


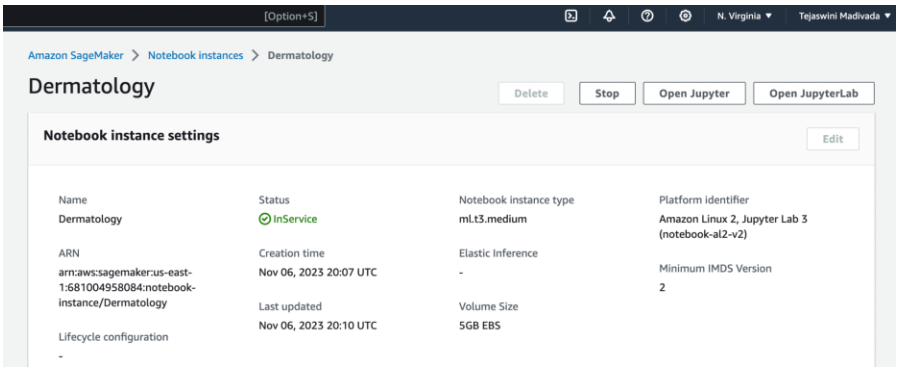
Loaded input CSV into the created bucket.



Amazon SageMaker offers a fully managed and integrated environment for developing, training and deploying machine learning models. This means you can focus more on the machine learning tasks and less on managing infrastructure, which can save time and effort SageMaker integrates seamlessly with other AWS services, such as S3 for data storage, Lambda for automation, and CloudWatch for monitoring, allowing you to build end-to-end solutions within the AWS ecosystem. Here we created a notebook instance in sagemaker.

Amazon Sage Maker instance:

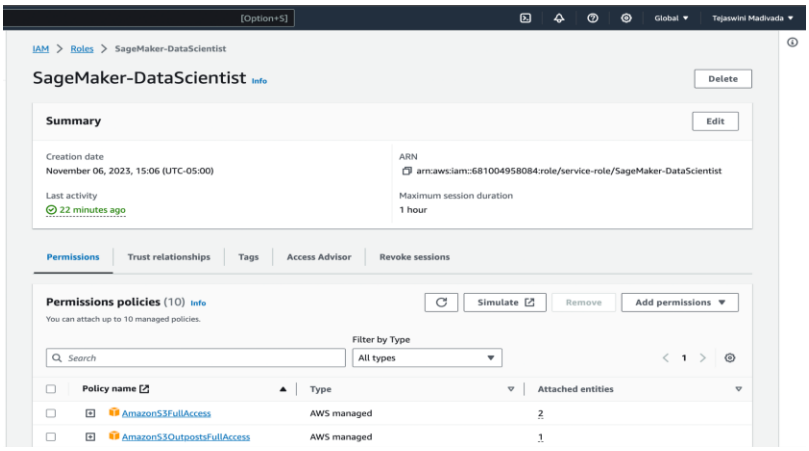




To facilitate the smooth execution of our machine learning project in Amazon SageMaker, we established a dedicated IAM (Identity and Access Management) role. This IAM role was carefully configured to grant the necessary permissions for our notebook instance to interact with critical AWS services. Specifically, we assigned permissions to access Amazon S3 and CloudWatch. This is pivotal for our project, as our dataset resides in an Amazon S3 bucket, and we need to read data from there. Additionally, we rely on CloudWatch to monitor the execution and gather valuable log data generated during the run of our Jupyter Notebook.

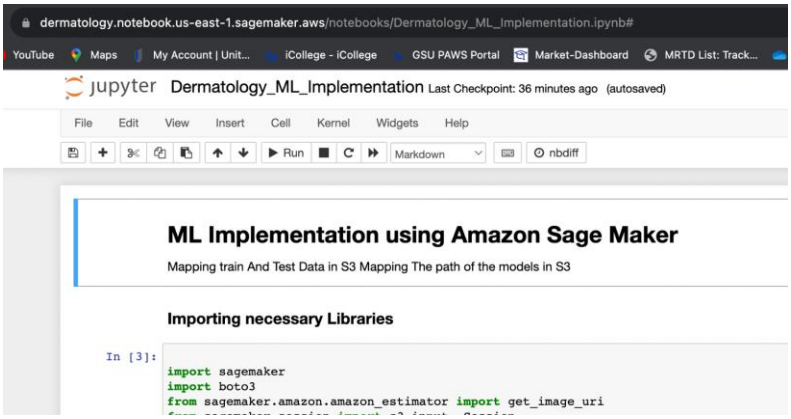
By setting up this IAM role with the appropriate permissions, we ensure that our project can seamlessly access the required resources while adhering to best practices in data security and AWS service integration. This configuration streamlines our project's workflow and enables efficient data retrieval and monitoring, ultimately contributing to the success of our machine-learning endeavor.

IAM ROLE:



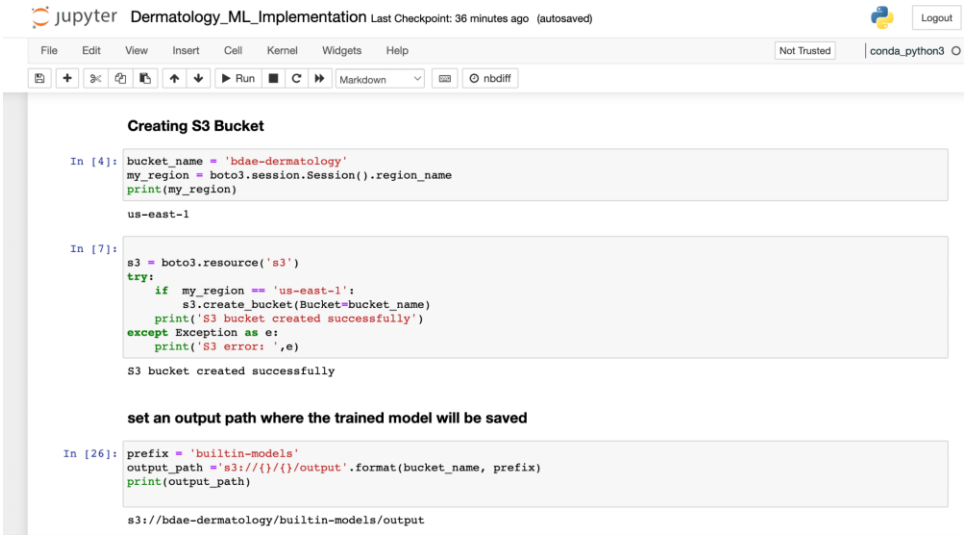
The below code segment sets up the environment for working with Amazon SageMaker, an AWS service for building, training, and deploying machine learning models

- Import sagemaker: This imports the sagemaker library, which provides the core functionalities for working with SageMaker. It includes tools for model training, deployment, and data management.
- Import boto3: The boto3 library is used to interact with various AWS services, including SageMaker. It allows you to create, configure, and manage AWS resources programmatically.
- From sagemaker.amazon.amazon_estimator import get_image_uri: This line imports the get_image_uri function from the amazon_estimator module within SageMaker. This function is used to obtain the URI (Uniform Resource Identifier) for a specific built-in Amazon SageMaker algorithm image. It's commonly used when specifying the algorithm image for training.
- From sagemaker. session import s3_input, Session: These imports bring in the s3_input and Session classes from the sagemaker.session module.
    - s3_input is used to specify the location of training data stored in Amazon S3 and how it should be processed.
    - A session is used to create a SageMaker session, which helps manage the interactions with SageMaker resources and configurations



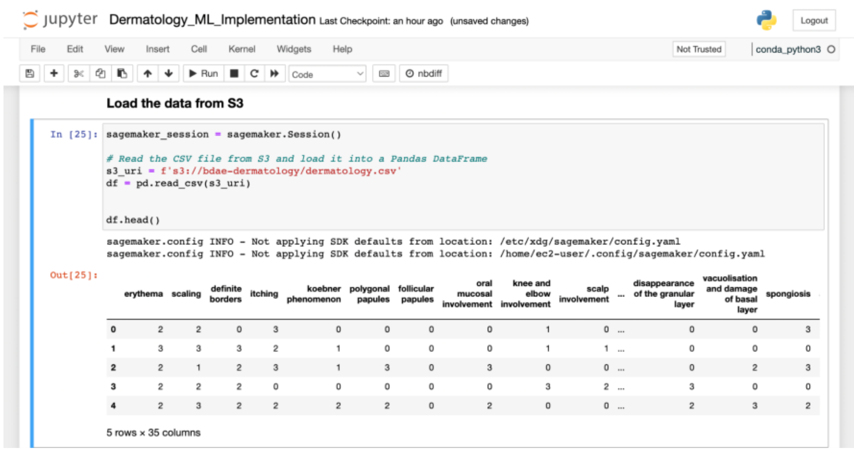
In this portion of our project, we set up an Amazon S3 (Simple Storage Service) bucket to efficiently manage our project's data. This code achieves the following key objectives:

- Bucket Creation: We created an S3 bucket named 'bdae-dermatology' by leveraging the Amazon S3 resource and the boto3 Python library. The choice of the bucket's region is determined dynamically by our current AWS region, and the bucket is created in the 'us-east-1' region, if applicable.
- Output Path Definition: We defined a 'output_path' variable that specifies the destination for our project's output. This path is structured as 's3://bdae-dermatology/builtin-models/output,' which will be used to store the results of our machine-learning models and related outputs.



In this part of the project, we're utilizing Amazon SageMaker, an AWS service that simplifies the machine learning workflow. The code can be explained as follows:

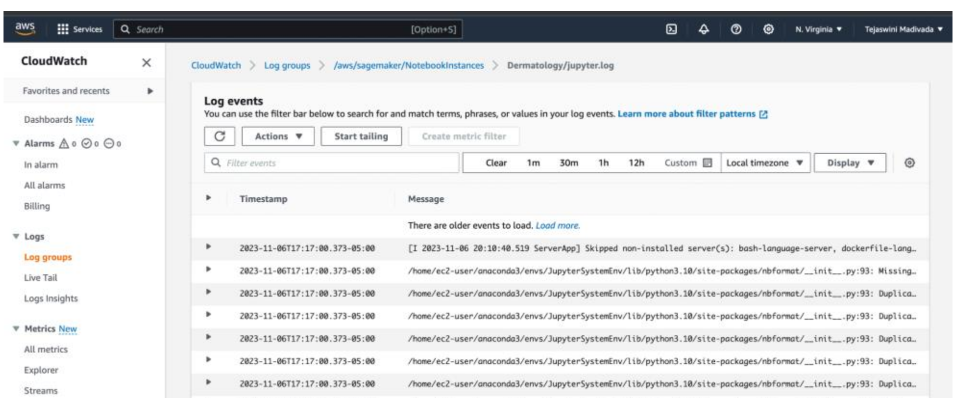
- SageMaker Session Initialization: We create a sagemaker_session to interact with SageMaker, allowing us to manage and execute machine learning tasks seamlessly within AWS.
- Data Retrieval from S3: We read our project's dataset from Amazon S3, which is a reliable and scalable object storage service, using the s3_uri variable. This dataset is stored in the 'bdae-dermatology' S3 bucket. By loading the data directly from S3, we ensure efficient access and use of our dataset.
- Pandas DataFrame Creation: The data from S3 is loaded into a Pandas DataFrame named 'df.' Pandas is a popular Python library for data manipulation and analysis. Loading the data into a DataFrame format allows us to work with and analyze the dataset effectively.



	erythema	scaling	definite borders	itching	koebner phenomenon	polygonal papules	follicular papules	oral mucosal involvement	knee and elbow involvement	scalp involvement	...	disappearance of the granular layer	vacuolisation and damage of basal layer	spongiosis
0	2	2	0	3	0	0	0	0	1	0	...	0	0	3
1	3	3	3	2	1	0	0	0	1	1	...	0	0	0
2	2	1	2	3	1	3	0	3	0	0	...	0	2	3
3	2	2	2	0	0	0	0	0	3	2	...	3	0	0
4	2	3	2	2	2	2	0	2	0	0	...	2	3	2

5 rows × 35 columns

Cloudwatch logs:



This image captures the CloudWatch Logs interface, an integral part of our project's monitoring and logging strategy. CloudWatch Logs enable real-time monitoring and tracking of events, providing insights into the performance, errors, and execution details of our machine learning workflows. These logs play a vital role in ensuring the reliability and stability of our project,

allowing us to diagnose issues, optimize performance, and maintain data integrity. With CloudWatch Logs, we gain transparency and visibility into our machine learning operations, facilitating efficient troubleshooting and informed decision-making.

DATA

The dataset originates from the UCI Machine Learning Repository, a pivotal resource for the empirical analysis of machine learning algorithms. Established in 1987 by UCI PhD student David Aha, the repository has been extensively utilized by millions of students, educators, and researchers. It includes 366 records with 34 features separated by commas; this dataset provides a comprehensive view of clinical and histopathological findings.

Features 1 to 11 contain clinical data, while features 12 to 33 capture histopathological findings. The age feature, numbered 34, exhibits eight unknown values and 358 values ranging from 0 to 75. The dataset encompasses patients with the six ESD categories. The target variables, or our y label, are the skin disease types, psoriasis, lichen planus, pityriasis rosea, pityriasis rubra pilaris, seborrheic dermatitis, and chronic/atopic dermatitis labeled 1-6 respectively.

Data Features

Clinical Features:
1. Erythema (0-3): Severity of redness in wounds.
2. Scaling (0-3): Amount of dandruff peeling off the skin or in lesions.
3. Definite Borders (0-3): Clarity of circumscribed borders in wounds.
4. Itching (0-3): Intensity of itching in wounds.
5. Koebner Phenomenon (0-3): Limited manifestation of dermatological disease due to traumatic skin stimulation.
6. Polygonal Papules (0-3): Raised, multi-edged lesions less than 1 cm in diameter on the skin.
7. Follicular Papules (0-3): Swellings less than 1 cm in height, evenly distributed.
8. Oral Mucosal Involvement (0-3): Lesions forming in the oral mucosa.
9. Knee and Elbow Involvement (0-3): Lesions forming on knees and elbows.
10. Scalp Involvement (0-3): Lesions forming on the scalp.
11. Family History (0-1): Presence of family history (binary).

Histopathological Attributes: (take values 0, 1, 2, 3)

12: melanin incontinence

13: eosinophils in the infiltrate

14: PNL infiltrate

15: fibrosis of the papillary dermis

16: exocytosis
17: acanthosis
18: hyperkeratosis
19: parakeratosis
20: clubbing of the rete ridges
21: elongation of the rete ridges
22: thinning of the suprapapillary epidermis
23: spongiform pustule
24: munro microabcess
25:focal hypergranulosis
26: disappearance of the granular layer
27: vacuolisation and damage of basal layer
28: spongiosis
29: saw-tooth appearance of retes
30: follicular horn plug
31: perifollicular parakeratosis
32: inflammatory monoluclear inflitrate
33: band-like infiltrate

DATA CLEANING AND PREPROCESSING

Data cleaning is a crucial process within the broader realm of data preparation. It involves identifying and sorting errors, inaccuracies, and inconsistencies within a database. We are doing this step to enhance quality, reliability, and usability. The primary goal of data cleaning is to ensure that the data accurately reflects the real-world data we seek to represent. This process involves handling missing values, correcting typos, and addressing images found to contain missing pixels or metadata attributes which will be flagged during the data import process.

One fundamental aspect of data cleaning is the identification and treatment of missing values. Missing data can arise due to various reasons, such as data entry errors, system failures, or intentional omissions. Data cleaning techniques involve making informed decisions about how to handle missing values, whether through imputation methods, removal of incomplete records, or other strategies.

In the data preprocessing phase, we executed several steps to ensure the quality and integrity of our dataset. Initially, we conducted an outlier check by examining the minimum, maximum, and average values for each dataset to identify any values falling outside the normal range. To

address missing values, predictive models trained on fully observed data points were employed for imputation. Notably, our dataset revealed missing values specifically for the "age" variable, prompting the utilization of a random forest model during data exploration. This model effectively predicted the missing ages, allowing us to replace the gaps with the imputed age data. Additionally, we performed a check for duplicate entries to eliminate redundancy and maintain data accuracy throughout the preprocessing process. These steps contribute to the reliability and suitability of the data for subsequent analyses and modeling.

Data Imbalance:

One key characteristic of the skin disease dataset that impacts model performance is the imbalance in the distribution of the target classes. Specifically, there is a substantially higher proportion of cases representing psoriasis (target value 1) compared to the other skin conditions. Psoriasis diagnosis makes up 30% of the dataset whereas the other skin diseases make up at most 19% or lower.

This skewed distribution of target classes can adversely impact model evaluation metrics and real-world performance. Since psoriasis patients likely dominate the dataset, there is a high risk of the model becoming biased towards simply predicting every case as psoriasis. This would artificially inflate overall accuracy, but provide little usable information to distinguish among the conditions. The less prevalent conditions would likely suffer from low recall due to the model rarely predicting these minority classes. To mitigate these issues, special techniques like over/under-sampling, loss weighting, and threshold optimization may be required during model development. Additionally, evaluation metrics that account for class imbalance like F1-score, precision, recall, and ROC-AUC will be more informative than simple accuracy.

In terms of real-world usage, the predictability gap between the common psoriasis condition and rarer disorders is problematic. A model will likely perform reasonably well flagging potential psoriasis cases, but remain unreliable at identifying less frequent conditions that are equally important to diagnose accurately. Thus, while psoriasis prediction serves an important clinical purpose, improving the detection of all conditions including seborrheic dermatitis, pityriasis rosea, etc. remains an area needing improvement through focused efforts to address data imbalance. Overall, despite the disproportionate disease distribution, strategically addressing this imbalance can yield a useful multi-class skin diagnosis model with real-world clinical utility across all target conditions.

To address the issue of the higher target values for psoriasis in our data set, firstly, we split the data into training and testing sets using the scikit-learn library. To tackle the imbalance in the target variable, we applied the Synthetic Minority Over-sampling Technique (SMOTE) specifically to the training set. This technique generated synthetic samples for the minority classes, ensuring a more balanced representation. After training a Decision Tree classifier on this

balanced data, the model was able to make predictions on the test set. The result was a training set with equal instances of each class, creating a more robust and unbiased model that can make predictions on new data with improved accuracy.

The key points are:

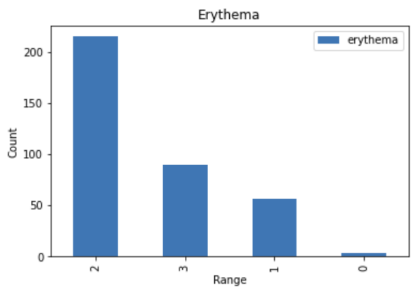
- Imbalance risks model bias towards predicting only psoriasis
- Obscures real-world performance in rare classes
- Needs over/under-sampling, loss weighting, optimized thresholds
- Evaluate with metrics accommodating imbalance (F1, ROC-AUC)
- The goal is to improve the detection of all conditions, not just psoriasis

---

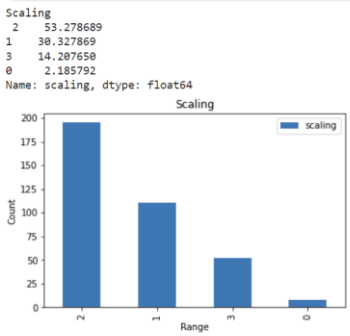
EXPLORATORY DATA ANALYSIS

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases have been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, and 2 indicate the relative intermediate values.

```
Erythema
 2    58.743169
3    24.590164
1    15.573770
0     1.092896
Name: erythema, dtype: float64
```




```
Scaling
 2    53.278689
1    30.327869
3    14.207650
0     2.185792
Name: scaling, dtype: float64
```



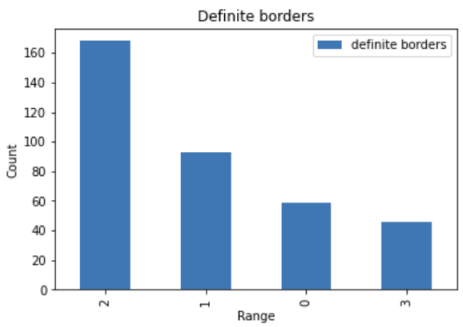
Erythema

Erythema refers to redness of the skin, often caused by inflammation, irritation, or increased blood flow to the affected area. It is a common symptom and can be associated with various skin conditions, allergies, infections, or other underlying health issues. There are higher instances of a severity score of two for erythema followed by a severity score of three, which is less than half of the severity score of two.

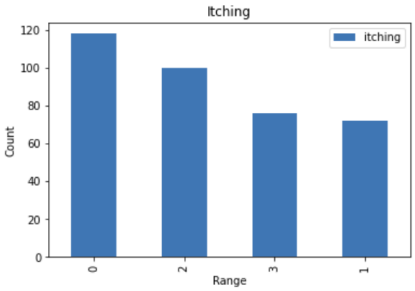
Scaling

Scaling is often associated with conditions such as psoriasis, seborrheic dermatitis, or fungal infections, where an abnormal rate of skin cell turnover leads to the visible shedding of skin flakes. There are higher instances of a severity score of two for scaling followed by a severity score of one, which is more than half of the severity score of two.

```
Definite borders
 2    45.901639
1    25.409836
0    16.120219
3    12.568306
Name: definite borders, dtype: float64
```




```
Itching
 0    32.240437
2    27.322404
3    20.765027
1    19.672131
Name: itching, dtype: float64
```



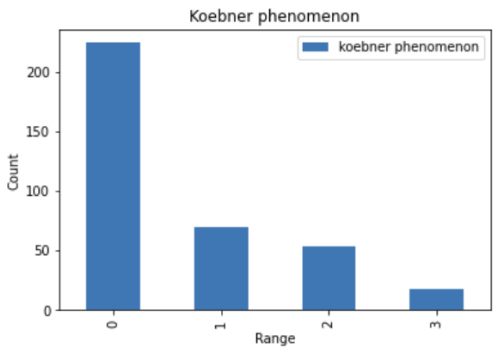
Definite Borders

Definite Borders refer to well-defined and clear boundaries or edges of a skin lesion, rash, or abnormality. There are higher instances of a severity score of two for definite borders followed by a severity score of one, which is more than half of the severity score of two.

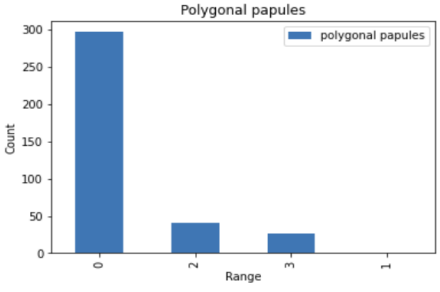
Itching

Itchy skin, also known as pruritus, is an irritating and uncontrollable sensation that makes you want to scratch to relieve the feeling. There are more instances of very low severity, followed by a severity level of two.

```
Koebner phenomenon
 0    61.202186
1    19.125683
2    14.754098
3     4.918033
Name: koebner phenomenon, dtype: float64
```




```
Polygonal papules
 0    81.147541
2    11.202186
3     7.377049
1     0.273224
Name: polygonal papules, dtype: float64
```



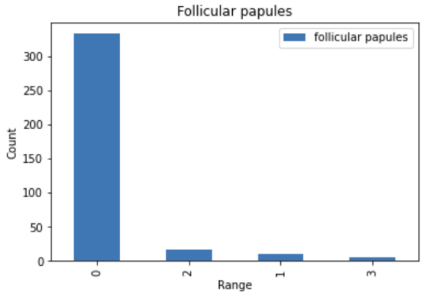
Koebner phenomenon

The Koebner phenomenon refers to the development of new skin lesions or the worsening of existing ones in areas of the skin that have been traumatized or injured. There are higher instances of low severity in this feature.

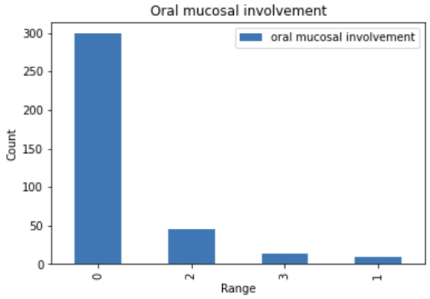
Polygonal papules

Polygonal papules refer to small, raised, and flat-topped skin lesions with multiple straight edges, resembling the shape of polygons. Very similar to the Koebner phenomenon, it also showed the lowest severity level as 0.

```
Follicular papules
 0    90.983607
2     4.371585
1     3.005464
3     1.639344
Name: follicular papules, dtype: float64
```




```
Oral mucosal involvement
 0    81.693989
2    12.295082
3     3.551913
1     2.459016
Name: oral mucosal involvement, dtype: float64
```



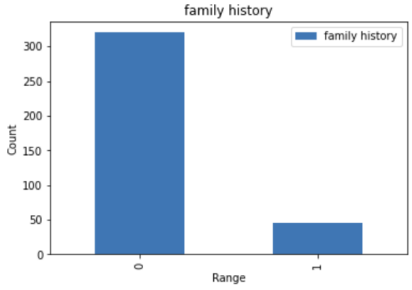
Follicular papules

Follicular papules are small, raised bumps on the skin that develop around hair follicles. The highest severity score for this feature is also zero.

Oral mucosal involvement

Oral mucosal involvement refers to the presence of symptoms, lesions, or abnormalities affecting the mucous membranes inside the mouth. The highest severity score for this feature very similar to most other features, is also zero.

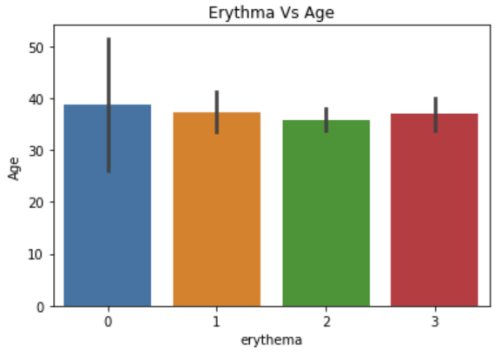
```
family history
 0    87.431694
1    12.568306
Name: family history, dtype: float64
```



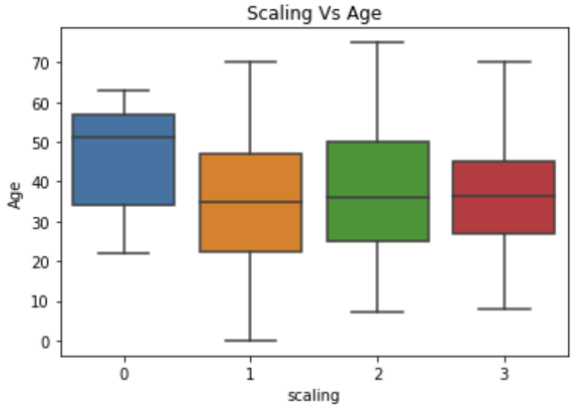
Family history

It is apparent that most of the patients in this data have a history of family members who have at least one skin disease. This suggests there may be a genetic factor at play when looking at these skin diseases, however, correlation is not causation.

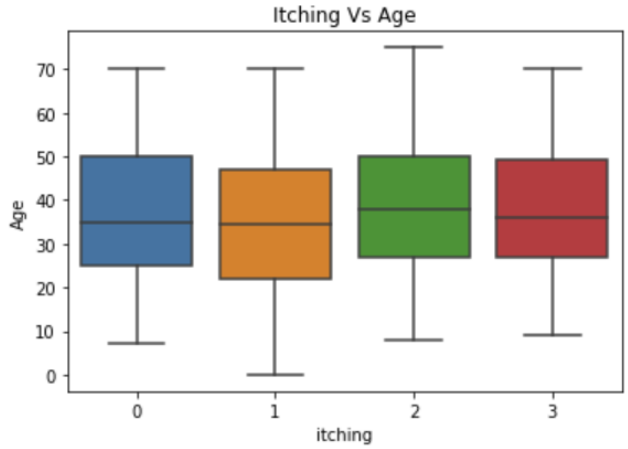
Bi-Variate Analysis



The above bar plot shows the Erythema with respect to Age. According to data, -People with no erythema have an age of around 39. -People with severe erythema have age around 38. Hence, Age is not a factor that influences Erythema severity.



The Boxplot shows the Age distribution with respect to Scaling.

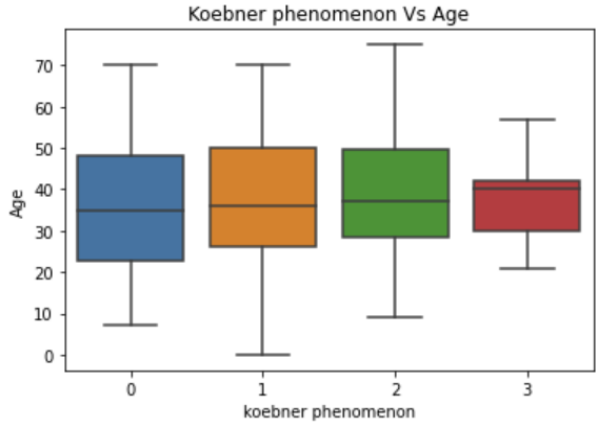


The Boxplot shows the Age distribution with respect to Itching.

According to the Data,

- The IQR range of no Itching in the skin is from 25 to 50.
- The IQR range of severe Itching in the skin is from 28 to 48.

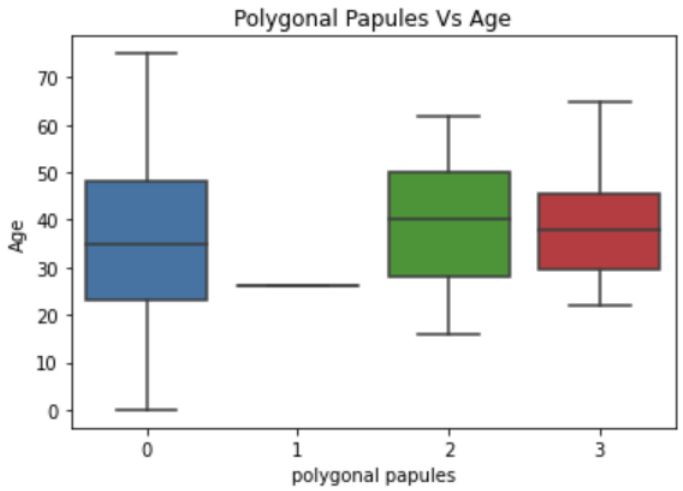
Koebner phenomenon Vs Age



The Boxplot shows the Age distribution with respect to Itching.

According to the Data,

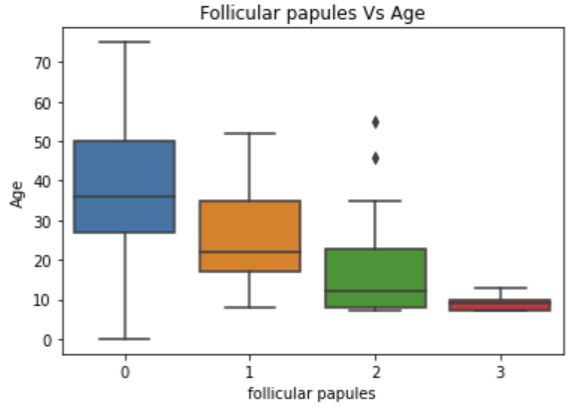
- The people who have no itching on the body have an equidistant IQR range.
- The people who have severe itching problems between 50 percentile to 75 percentile are very less.



The Boxplot shows the Age distribution with respect to Polygonal Papules.

According to the Data,

- The people who have no Polygonal Papules on the body have IQR range from 23 to 48 age.
- The people who have severe Polygonal Papules on the body have IQR ranging from 28 to 35 age

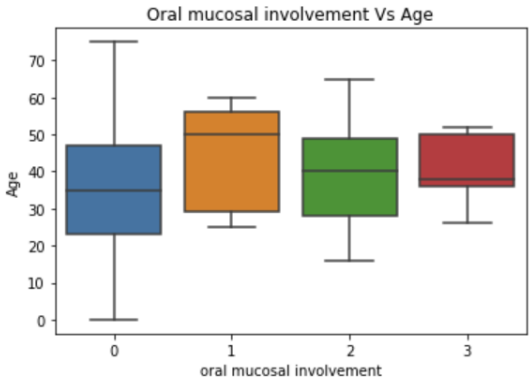


The Boxplot shows the Age distribution with respect to Follicular papules.

According to the Data,

- The people who have no Follicular papules on the body have IQR range from 28 to 50 age.
- The people who have severe Follicular papules on the body have IQR ranging from 8 to 12 years old. It seems that the problem comes to young children rather than adults.

Oral mucosal involvement vs. age



The Boxplot shows the Age distribution with respect to oral mucosal involvement.

According to the Data,

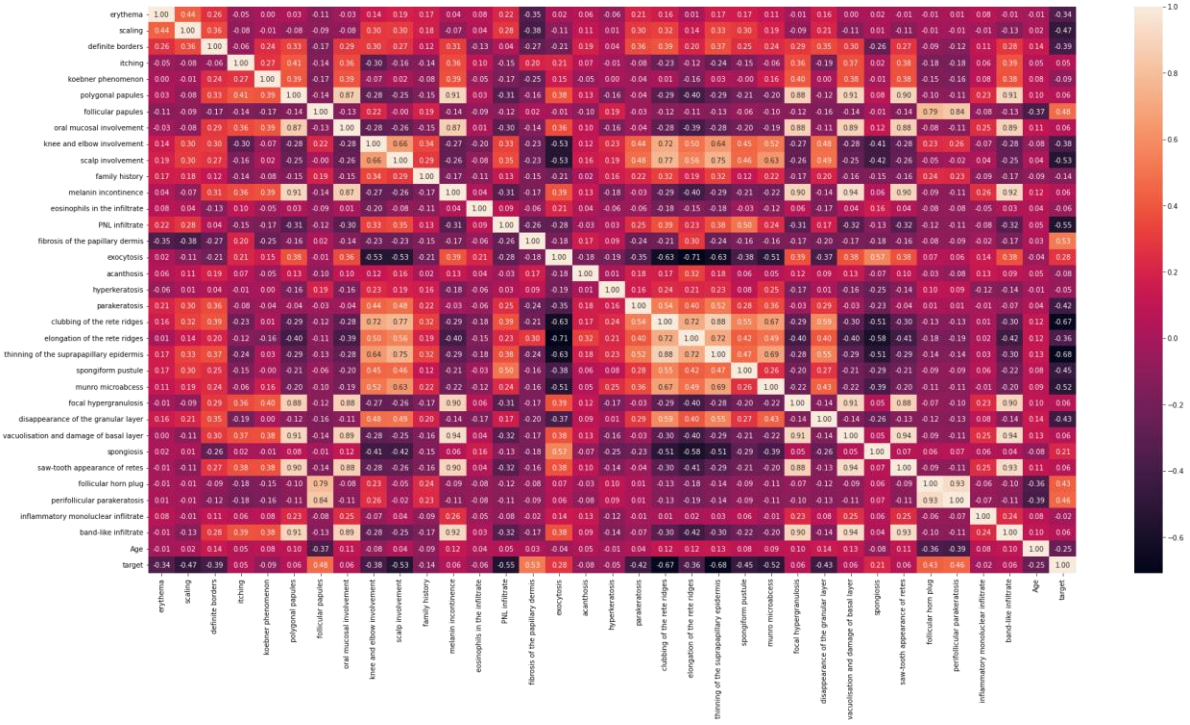
- The people who have no oral mucosal involvement have IQR range from 23 to 48 age.
- The people who have severe oral mucosal involvement are between 50 percentile to 75 percentile which is from 37 to 49 age.

FEATURE SELECTION

Cross-correlation is an operation used to measure the similarity between two sets of data, often comparing how one set of values changes in relation to another set. In simple terms, it helps us understand how patterns in one set of data correspond or match with patterns in another set.

Imagine you have two sequences of numbers, like two lists of temperature readings over time from two different locations. Cross-correlation allows you to determine if there's a relationship between the temperature changes in one location compared to the changes in the other. It helps identify whether changes in one set of data are mirrored, delayed, or unrelated to changes in the other set. In essence, cross-correlation is a tool that helps us find similarities or connections between different sets of data by examining how they vary together over time or space.

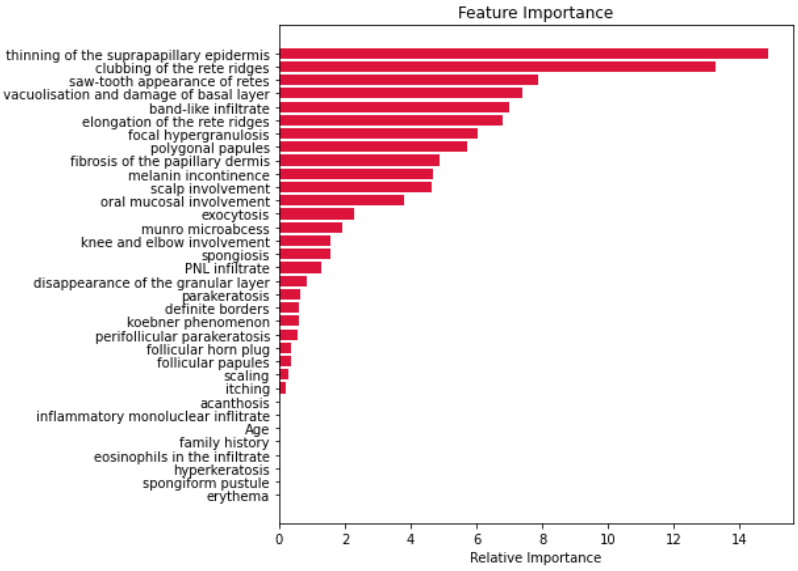
Correlational Matrix



Feature importance

Feature importance refers to the measure of the impact or contribution of different features (variables or attributes) in a model toward predicting the target outcome. It is a concept commonly used in machine learning and statistical modeling to understand which features are most influential in making predictions.

We utilized scikit-learn Random Forest classifier to dig into our data and gauge feature importance. This tool helped us analyze which specific features carry the most weight when it comes to making predictions using the Random Forest classifier. The generated chart served as a visual guide, offering valuable insights into the significance of each feature and how they contribute to the predictive power of our model. This way, we could pinpoint the key factors that play a crucial role in the decision-making process of the Random Forest algorithm. After discovering the most important features, we removed the least important features (features at relative importance 0) from our selection for modeling.



MODELING

Since It was a multi-class classification problem we tried to understand the assumption of each ML algorithm that can perform well in this setting. We used the following ML Models:Support Vector Machine Classifier using Linear Kernel, K- Nearest Neighbor with K = 6, Categorical Naïve Bayes, Random Forest, and Gradient Boosting Classifier.

SVM-LINEAR

Support Vector Machines (SVM) with a linear kernel, often referred to as SVM with a linear classifier, is a type of machine learning model used for classification and regression tasks. The linear kernel is one of the simplest types of kernels used in SVM.

Accuracy:

- Accuracy on Train Data: 99%
- Accuracy on Test Data: 98%

Explanation:

- Accuracy represents the ratio of correctly predicted instances to the total number of instances. It is a general measure of the model's overall correctness.
- A high accuracy on both the training and test datasets indicates that the model is making accurate predictions.

Precision:

- Precision on Train Data: 99%
- Precision on Test Data: 98%

Explanation:

- Precision is the ratio of correctly predicted positive observations to the total predicted positives. It measures the model's ability to avoid false positives.
- A high precision indicates that when the model predicts a positive outcome, it is likely correct. It is particularly relevant in situations where false positives are costly.

Recall (Sensitivity or True Positive Rate):

- Recall on Train Data: 99%
- Recall on Test Data: 98%

Explanation:

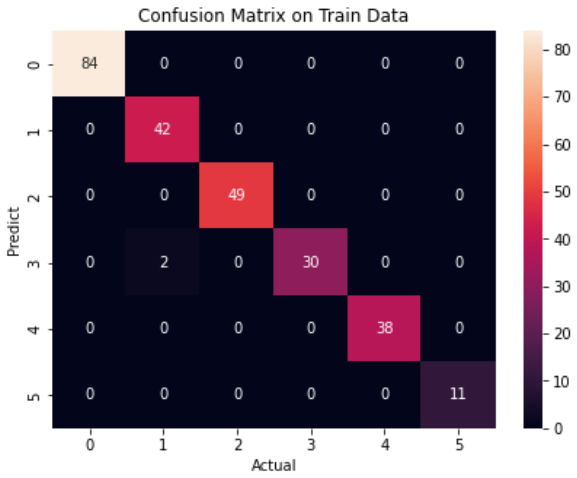
- Recall is the ratio of correctly predicted positive observations to all actual positives. It measures the model's ability to capture all relevant instances.
- High recall indicates that the model is effective at identifying most of the positive instances. It is particularly important when the cost of false negatives is high.

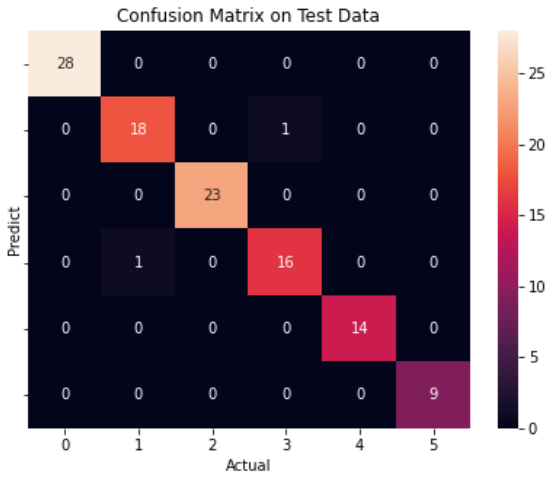
F1 Score:

- F1 on Train Data: 99%
- F1 on Test Data: 98%

Explanation:

- The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall.
- A high F1 score indicates a model that achieves both high precision and high recall. It is useful when there is an uneven class distribution.

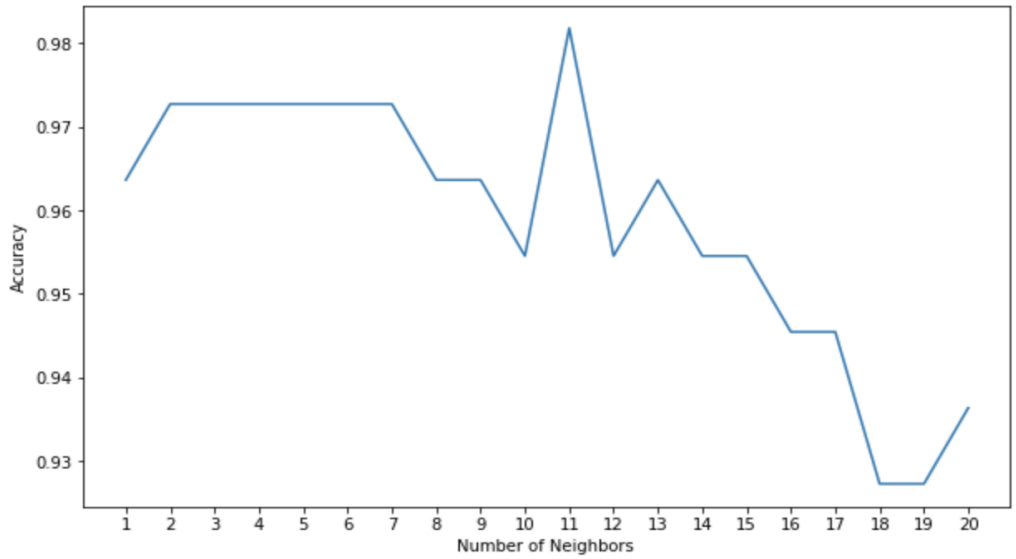




KNN Classifier

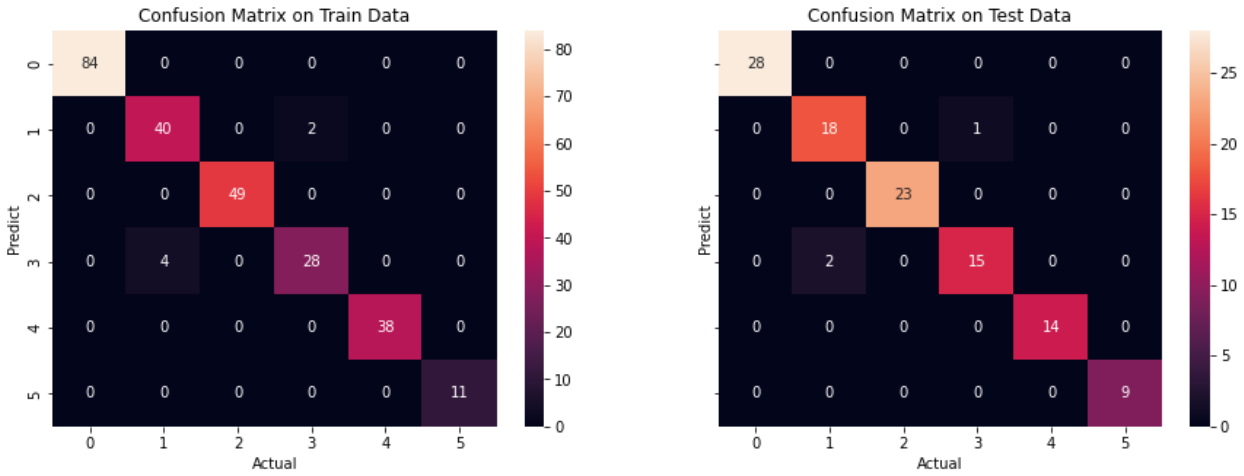
K-Nearest Neighbors (KNN) is a simple and intuitive classification algorithm that falls under the category of supervised learning. It's a non-parametric and lazy learning algorithm, meaning it doesn't make assumptions about the underlying data distribution and doesn't explicitly build a model during the training phase. Instead, it classifies new data points based on their proximity to existing data points in the feature space.

We implemented a code function that iterates through values of k ranging from 1 to 20. For each iteration, the function trains a K-Nearest Neighbors model, and through this process, we identify the value of k that yields the highest F1 score. This approach allows us to see different numbers of neighbors to determine the optimal configuration for our KNN model in terms of predictive accuracy. As the model shows, a K = 11 gives us the highest accuracy.



- Accuracy:

    - The proportion of correctly classified instances among the total instances. In both training and test datasets, the model correctly predicts the class labels for 97% of instances.
- Precision:
    - The ratio of correctly predicted positive instances to the total instances predicted as positive. In both datasets, when the model predicts a positive outcome, it is correct 97% of the time.
- Recall:
    - The ratio of correctly predicted positive instances to the total actual positive instances. In both datasets, the model captures 97% of the actual positive instances.
- F1 Score:
    - The harmonic mean of precision and recall. It provides a balance between precision and recall. In both datasets, the F1 score is 97%.



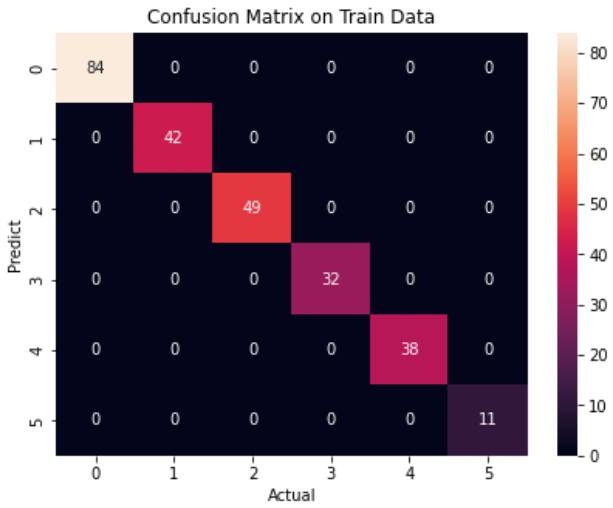
Tuning K values using Grid Search CV:

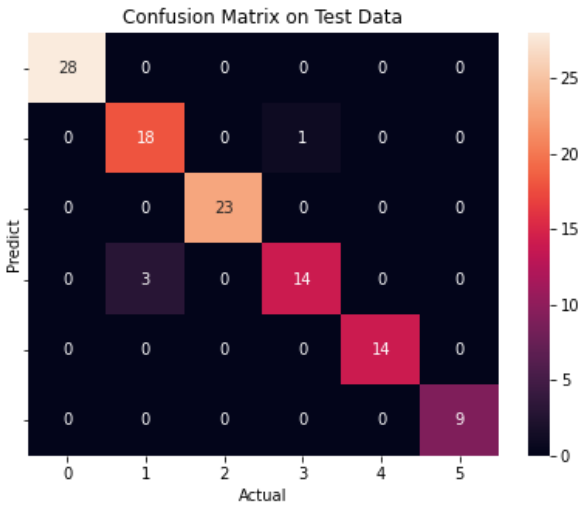
Tuning the value of K in a k-nearest neighbors (KNN) model can significantly impact its performance. Grid Search Cross-Validation (Grid Search CV) is a systematic approach to finding the best hyperparameter values by evaluating the model's performance across various combinations of hyperparameters.

- Accuracy:
    - Accuracy represents the ratio of correctly predicted instances to the total number of instances. The model correctly predicts the class labels for 97% of instances in both the training and test datasets.
- Precision:

- Precision is the ratio of correctly predicted positive instances to the total instances predicted as positive. In both datasets, when the model predicts a positive outcome, it is correct 97% of the time.

- Recall (Sensitivity or True Positive Rate):
  - Recall is the ratio of correctly predicted positive instances to all actual positive instances. The model captures 97% of the actual positive instances in both the training and test datasets.
- F1 Score:
  - The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall. In both datasets, the F1 score is 97%.



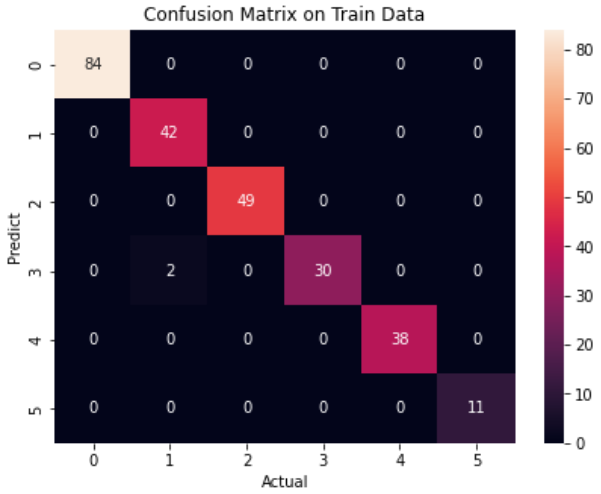


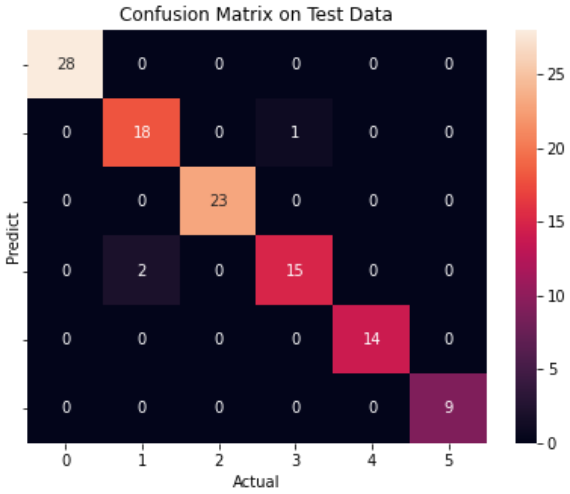
CATEGORICAL NAIVE BAYES

Categorical Naive Bayes, also known as Multinomial Naive Bayes, is a variant of the Naive Bayes algorithm designed for handling discrete features. It is particularly well-suited for classification tasks where the input features are categorical or represent counts of occurrences in different categories. This model is commonly used in natural language processing (NLP) for tasks like text classification.

- Accuracy:
  - The model correctly predicts the class labels for 99% of instances in the training dataset and 97% in the test dataset.
- Precision:
  - In the training dataset, when the model predicts a positive outcome, it is correct 99% of the time. In the test dataset, this drops to 97%.
- Recall (Sensitivity or True Positive Rate):
  - The model captures 99% of the actual positive instances in the training dataset, while in the test dataset, it captures 97%.

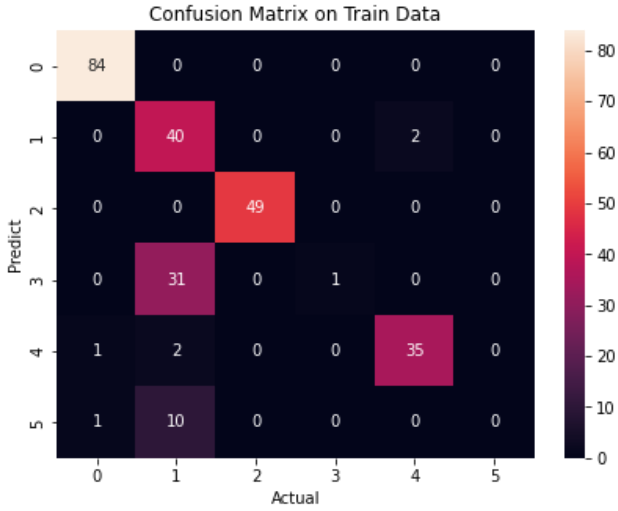
- F1 Score:
    - In the training dataset, the F1 score is 99%, and in the test dataset, it is 97%.

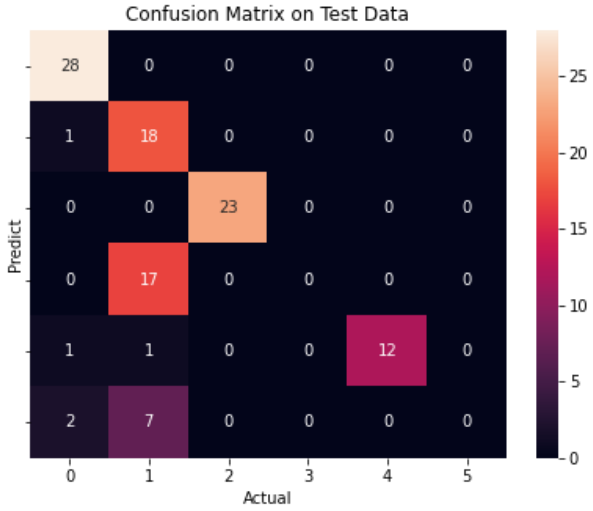




RANDOM-FOREST CLASSIFIER

Random Forest is an ensemble learning method that can be used for both classification and regression tasks. It builds multiple decision trees during the training phase and merges their predictions to provide a more accurate and robust result. The "random" in Random Forest comes from the fact that it introduces randomness in the tree-building process.





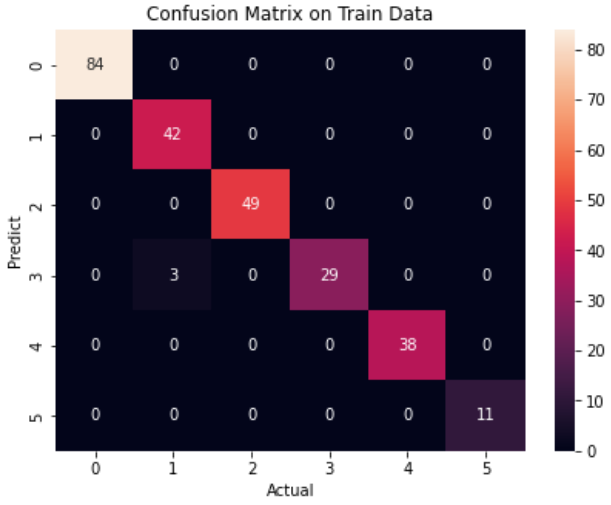
HYPERPARAMETER TUNING FOR RANDOM FOREST CLASSIFIER:

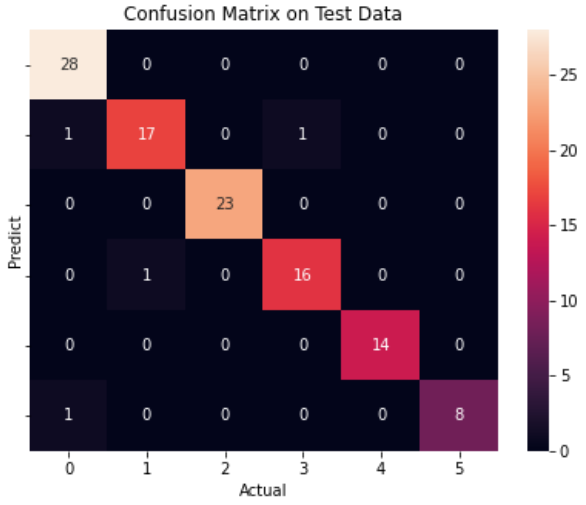
Hyperparameter tuning is a crucial step in optimizing the performance of a Random Forest Classifier. Adjusting the hyperparameters can significantly impact the model's accuracy, robustness, and generalization to new data. Here are some key hyperparameters for a Random Forest Classifier and considerations for tuning them:

We used a hyperparameter grid for a RandomForestClassifier with options for 'n_estimators', 'max_features', and 'max_depth'. It prints the current time and initializes a GridSearchCV object to perform hyperparameter tuning using 5-fold cross-validation on the RandomForestClassifier. The tuning aims to find the optimal combination of hyperparameters for the best model performance. Here below showcases the best parameters for tuning our random forest model.

```
▼ RandomForestClassifier
RandomForestClassifier(max_depth=7, n_estimators=300, random_state=18)
```

- Accuracy:
  - The model correctly predicts the class labels for 98% of instances in the training dataset and 96% in the test dataset.
- Precision:
  - In the training dataset, when the model predicts a positive outcome, it is correct 98% of the time. In the test dataset, this drops to 96%.
- Recall (Sensitivity or True Positive Rate):
  - The model captures 98% of the actual positive instances in the training dataset, while in the test dataset, it captures 96%.
- F1 Score:
  - In the training dataset, the F1 score is 98%, and in the test dataset, it is 96%.



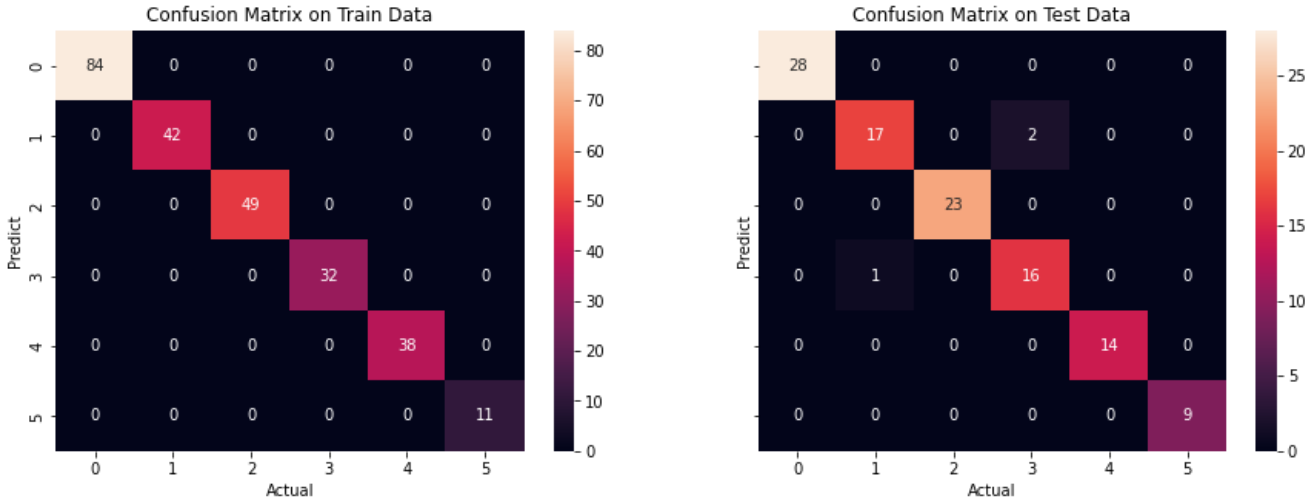


GRADIENT BOOSTING CLASSIFIER

Gradient Boosting is an ensemble learning technique that builds a series of weak learners, typically decision trees, sequentially, with each tree correcting the errors of its predecessor. Gradient Boosting is a powerful and widely used algorithm for both classification and regression tasks. One popular implementation of Gradient Boosting is the Gradient Boosting Classifier.

Gradient Boosting is a powerful algorithm with high predictive accuracy and flexibility. It is widely used in practice and has shown success in various machine-learning competitions.

- Accuracy:
  - The model correctly predicts the class labels for 100% of instances in the training dataset and 97% in the test dataset.
- Precision:
  - In the training dataset, when the model predicts a positive outcome, it is correct 100% of the time. In the test dataset, this drops to 97%.
- Recall (Sensitivity or True Positive Rate):
  - The model captures 100% of the actual positive instances in the training dataset, while in the test dataset, it captures 97%.
- F1 Score:
  - In the training dataset, the F1 score is 100%, and in the test dataset, it is 97%.



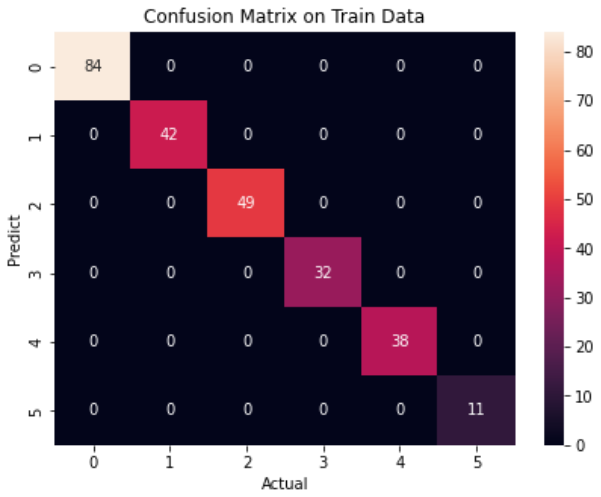
HYPERPARAMETER TUNING FOR GRADIENT BOOSTING CLASSIFIER:

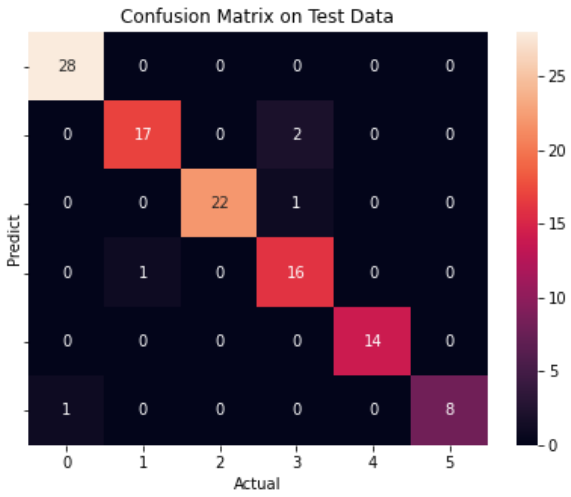
When tuning hyperparameters, it's important to strike a balance between model complexity and generalization to avoid overfitting. Hyperparameter tuning for a Gradient Boosting Classifier involves optimizing the parameters that control the behavior and complexity of the model. Gradient Boosting models, such as XGBoost, LightGBM, and Scikit-learn's GradientBoostingClassifier, have several hyperparameters that can be fine-tuned for optimal performance.

We used a grid search for hyperparameter tuning on a GradientBoostingClassifier. It sets up a parameter grid including options for 'n_estimators', 'max_depth', 'min_samples_split', 'learning_rate', and 'random_state'. The start time is printed, then the GridSearchCV function is used to fit the model using 5-fold cross-validation on training data (X_train, y_train). Finally, the end time is printed to show the duration of the grid search process.

```
GradientBoostingClassifier
GradientBoostingClassifier(min_samples_split=50, n_estimators=200,
                           random_state=18)
```

- Accuracy:
  - The model correctly predicts the class labels for 100% of instances in the training dataset and 95% in the test dataset.
- Precision:
  - In the training dataset, when the model predicts a positive outcome, it is correct 100% of the time. In the test dataset, this drops to 95%.
- Recall (Sensitivity or True Positive Rate):
  - The model captures 100% of the actual positive instances in the training dataset, while in the test dataset, it captures 95%.
- F1 Score:
  - In the training dataset, the F1 score is 100%, and in the test dataset, it is 95%





MODAL EVALUATION:

Model Name	F1-Train	F1-Test
SVM-Linear	98%	97%
KNN	97%	96%
Categorical Naive Bayes	99%	97%
Random Forest	81%	71%
Random Forest Tuned	98%	96%
Gradient Boosting Classifier	100%	97%
Gradient Boosting Classifier Tuned	100%	96%

SVM-Linear:

- F1-Train: 98%
- F1-Test: 97%

The SVM-Linear model demonstrates strong performance on both training and test datasets, with a slightly lower but still excellent F1 score on the test dataset. This suggests good generalization.

KNN:

- F1-Train: 97%
- F1-Test: 96%

The KNN model performs well on both training and test datasets, with consistent F1 scores. It indicates good generalization and balanced precision and recall.

KNN Tuned:

- F1-Train: 100%
- F1-Test: 97%

The tuned version of the KNN model shows a perfect F1 score on the training data and strong performance on the test data, indicating effective parameter tuning and good generalization.

Categorical Naive Bayes:

- F1-Train: 99%
- F1-Test: 97%

The Categorical Naive Bayes model performs well on both training and test datasets, with high F1 scores, suggesting a good balance between precision and recall.

Random Forest:

- F1-Train: 81%
- F1-Test: 71%

The Random Forest model shows a notable drop in F1 score from training to test datasets, indicating potential We utilized a scikit-learn Random Forest classifier to dig into our data and gauge feature importance. This tool helped us analyze which specific features carry the most weight when it comes to making predictions using the Random Forest classifier. The generated chart served as a visual guide, offering valuable insights into the significance of each feature and how they contribute to the predictive power of our model. This way, we could pinpoint the key factors that play a crucial role in the decision-making process of the Random Forest algorithm.

After discovering the most important features, we removed the least important features from our selection for modeling. need for model refinement.

Random Forest Tuned:

- F1-Train: 98%
- F1-Test: 96%
- The tuned Random Forest model exhibits improved performance on both training and test datasets compared to the untuned version, but there is still a gap between training and test F1 scores.

Gradient Boosting Classifier:

- F1-Train: 100%
- F1-Test: 97%
- The Gradient Boosting Classifier shows a perfect F1 score on the training data and strong generalization to the test data, indicating effective learning.

Gradient Boosting Classifier Tuned:

- F1-Train: 100%
- F1-Test: 96%
- The tuned Gradient Boosting Classifier maintains strong performance on both training and test datasets, with a slight decrease in F1 score on the test data.

Upon reviewing the overall observations, specific models such as SVM-Linear, KNN, Categorical Naive Bayes, and the Gradient Boosting Classifier do well with the test data. Tuning helps KNN and Random Forest models perform better. However, Random Forest models, even with tuning, see a bit of a drop in F1 scores when moving from training to test datasets. On the other hand, the Gradient Boosting models keep up a strong performance, even after tuning. These observations give us a clear idea of how each model handles the test data and how tuning affects their performance. Consider further exploration of the models with lower F1 scores, investigating potential causes for the observed gaps between training and test performance, and fine-tuning hyperparameters to optimize overall model performance. Out of all the SVM-Linear performed well on the train and test datasets.

POWER BI EXPLORATIONS

Power BI is a business analytics service by Microsoft that provides interactive visualizations and business intelligence capabilities with an interface that is easy to use for end users to create their

reports and dashboards. It enables users to connect to a wide variety of data sources, visualize and share insights across the organization, or embed them in an app or website.

Purpose and Use of Power BI:

Power BI is used for transforming raw data into meaningful information, and to make informed decisions. The platform can connect to various sources, including databases, Excel files, and online services like Databricks. The primary purpose of Power BI is to create compelling visualizations and reports for data analysis. Some features include data transformation, modeling, and creating relationships between different data tables.

Importing Dermatology CSV into Power BI:

To begin the analysis, we imported the dermatology data in CSV format into Power BI. This involved navigating to the "Home" tab, selecting "Get Data," and choosing "Text/CSV" as the data source. We then specified the file path and loaded the data into Power BI.[

Using DAX Formulas for Analysis:

DAX (Data Analysis Expressions) is a formula language used in Power BI for creating custom calculations in tables and matrices. In this analysis, we employed DAX formulas to project clinical attributes over a 10-year period based on various dermatological indicators. We wanted to get an idea of a future state skin disease clinical feature expectations based on a number of factors including climate.

Using Climate Change:

Recognizing the intrinsic connections between dermatological conditions and environmental influences, it became imperative to avoid an incomplete understanding by neglecting the impacts of shifting climate patterns. Sourcing climate change datasets became a critical aspect of this endeavor, requiring the identification and acquisition of reputable datasets from leading scientific institutions. The key variables encompassed temperature, humidity, UV radiation, precipitation, and atmospheric CO2 concentrations. The integration of these multi-sourced datasets into the Power BI platform involved significant efforts in terms of data transformation and normalization to ensure compatibility across diverse sources. A pivotal aspect of this process involved careful consideration of the credibility, recency, and geographic granularity of the sourced climate data. Ensuring that the data used in the analysis was not only reliable and up-to-date but also specific to relevant geographic regions, played a crucial role in enhancing the accuracy and applicability of the integrated climate change factors within the dermatology trends model.

Clinical Projection with DAX:

Leveraging DAX, we employed functions like CALCULATE and VAR to analyze historical data related to dermatology clinical attributes. By factoring in variables such as clinical attributes, and climate-related factors, we created a dynamic model to project the amount of individuals exhibiting dermatology clinical attributes over the next 10 years. DAX's flexibility allowed for adjusting parameters and refining the model as needed.

```
Projected Cases =
VAR MaxAge = MAX(Derm[VAR YearsProjected])
VAR MinAge = MIN(Derm[VAR YearsProjected])
VAR YearsProjected = 10

RETURN
```

Histopathological Projection with DAX:

Beyond individual projections, we utilized DAX to forecast the overall number of dermatology cases in the coming years. The SUMX function, coupled with appropriate filtering, enabled the creation of a robust predictive model. Integrating external factors, including climate change data, allowed for a more comprehensive understanding of potential shifts in dermatology prevalence.

Temporal Dynamics and Iterative Refinement:

Understanding that the dermatology landscape is not static, we incorporated temporal dynamics into the DAX formulas. This ensured that the projections were not merely extrapolations but dynamic reflections of changing conditions. The iterative refinement process, a hallmark of DAX analysis, allowed for constant adaptation, keeping the model relevant and reliable in the face of evolving circumstances.

PowerBI Cards for Instant Insights:

Leveraging PowerBI cards, strategically presented key insights tailored to the nuances of the dermatology landscape. Displaying real-time and static data on current case numbers, these cards served as efficient tools for stakeholders to swiftly grasp essential information. The dynamic updating feature ensured that decision-makers were consistently equipped with the latest data, aligning seamlessly with the principles of agile and responsive decision-making. These cards, succinctly delivering critical data points, became indispensable in fostering a clear understanding of the dermatological scenario.

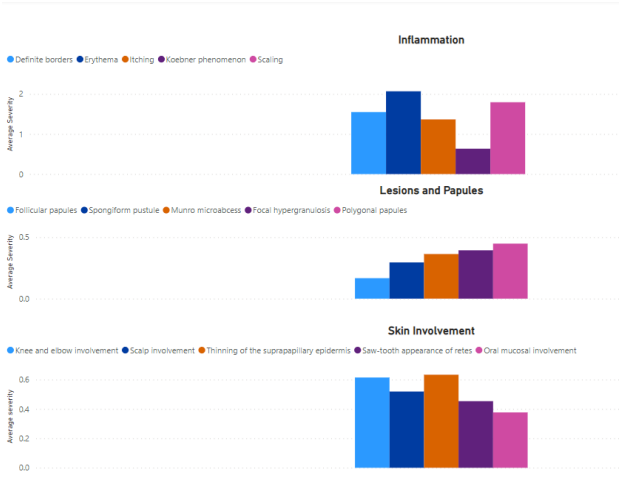
Impact of Climate Change on Dermatology Projections

By incorporating climate change data into our projection models, we gained insights into how shifts in climate patterns could impact the future prevalence of dermatological conditions. This

forward-looking analysis provided stakeholders with a more holistic view, allowing for proactive planning and mitigation strategies.

Insightful Bar Charts:

To elevate the visual storytelling aspect, we incorporated bar charts and dynamic line graphs into the PowerBI dashboards. Bar charts effectively conveyed distribution, bringing forth trends and anomalies with clarity. We grouped the bar charts by dermatology symptom groups which allowed me to drill down and showcase the average severity per group. This provided additional granular insights into the relationships between specific symptoms and severity levels over time. Simultaneously, line graphs became instrumental in depicting the trajectory of dermatology cases over time. This multi-layered visual representation facilitated the identification of patterns and potential correlations with climate change variables, offering a comprehensive narrative.



Tackling Challenges Head-On:

Circular References:

In the intricate dance of data modeling, circular references emerged as a notable challenge. DAX, being sensitive to such dependencies, demanded a meticulous approach. Through iterative refinement and restructuring of formulas, we successfully mitigated circular references, ensuring the stability and accuracy of my calculations. This process not only underscored the importance

of a systematic approach but also highlighted the robustness of DAX in handling complex scenarios.

Optimizing 'SUMX' for Efficient Performance:

As the complexity of the model increased, challenges related to the 'SUMX' function arose, particularly when dealing with too many causes. Recognizing the potential impact on performance, we delved into optimizing DAX expressions. Streamlining calculations and implementing techniques such as the use of summary tables became essential strategies. These optimizations not only addressed performance concerns but also showcased the adaptability and problem-solving capabilities inherent in the integration of DAX and PowerBI.

Navigating the Dermatology Data Landscape:

In conclusion, the amalgamation of DAX formulas and PowerBI not only facilitated a meticulous analysis of dermatology data but also paved the way for informed decision-making in an ever-changing environment. The projection models and case forecasts, enriched by climate change considerations, provided stakeholders with not just a snapshot but a dynamic panorama of the dermatology landscape.

By effectively utilizing the capabilities of DAX and PowerBI, we navigated the intricacies of dermatology data analysis, shedding light on both current trends and future trajectories. The challenges encountered, whether in the form of circular references or performance optimizations, underscored the need for a nuanced approach. The solutions found not only resolved specific issues but also contributed to the overarching narrative of adaptability and resilience in the face of complex data scenarios.

As we stand at the intersection of data analysis and dynamic visualization, the journey undertaken exemplifies the potential encapsulated within DAX and PowerBI. Beyond being mere tools, they become enablers of insight, foresight, and strategic decision-making in the complex realm of dermatology and beyond.

GROUP CONCLUSION

In conclusion, this project aims to develop an artificial intelligence-powered system for automated diagnosis of skin conditions that can enhance accessibility, efficiency, consistency and accuracy of dermatological care. Leveraging state-of-the-art deep learning techniques along with rigorous validation protocols, we will build robust models capable of classifying diverse

manifestations of skin diseases from data inputted into the model- providing predictive analysis of what will most likely be the skin disease the patient is exhibiting, with respect to clinical and histopathological data. By augmenting the capabilities of dermatologists and reducing the barriers to expert diagnosis, this initiative has immense potential to reshape the landscape of dermatology practices globally and improve patient outcomes through early interventions.

Real-world Impact:

If successfully implemented, this project can have a far-reaching impact on clinical dermatology workflows and medical outcomes. Automating time-consuming diagnosis procedures will greatly alleviate the burden on dermatologists, allowing them to focus on more complex decision-making and patient interactions. Patients in remote areas with scarce dermatology resources can benefit immensely from the geographical reach of this system. Flagging potential malignancies early via periodic screening can quite literally save lives through timely treatment. Together with telemedicine infrastructure, the system can extend specialized dermatology services to underprivileged regions. Beyond these direct clinical impacts, the dataset accumulated through this initiative will spur dermatology research.

Trust and Compliance:

However, realizing this vision requires carefully addressing challenges around model interpretability, patient privacy and regulatory compliance. Trust is paramount - especially in healthcare - and opaque black box models often undermine confidence. Hence it is critical to engineer transparent models and complement them with expert-in-the-loop frameworks that combine human and machine intelligence for enhanced accountability. Additionally, managing sensitive patient health data necessitates stringent controls around deidentification, encryption and access restrictions to safeguard confidentiality. Regulatory principles of ethics, safety and reliability will be integral to technology design.

Future Work:

There remain ample opportunities to build on this foundation across multiple dimensions. The system can be enhanced to not just classify conditions but also estimate severity scores. Reinforcement learning can enable dynamic Bayesian treatments. Longitudinal patient monitoring can provide insights into progression and personalized care. Deploying the system on edge devices like smartphones can enable in-home monitoring and advice. Ultimately, we envision this project as the cornerstone of an ecosystem for AI-driven dermatology care - with patient wellbeing at its core.

Broader Perspective:

Stepping back, this initiative exemplifies the transformational potential of artificial intelligence in augmenting and enhancing healthcare. With sufficient precautions and purposeful design, these engineering innovations can democratize quality care. We believe that pragmatic and

ethical deployment of technology can uplift health standards worldwide. This project represents a modest yet meaningful step in that broad direction. We look forward to feedback and participation from the community.

Hana Awad:

As a student studying applications of artificial intelligence, I welcomed the chance to build classification models on a real healthcare dataset for a course project. Getting hands-on practice moving beyond textbook concepts to apply machine learning using Python appealed to my learning style.

We received a dataset containing medical records with dozens of patient attributes and a target variable noting one of multiple possible disease diagnoses. Beyond predicting this outcome, we were instructed to demonstrate several modeling approaches for comparison and recommend the best overall technique considering accuracy, computational efficiency, and reliability. After importing the necessary Python packages for data manipulation, visualization and machine learning, I first examined relationships within the dataset to inform feature selection. Choices balancing predictive power and overcomplexity required experience I aimed to build.

With features identified, I split the thousands of records into exclusive partitions for training versus final holdout testing. This enabled tuning models strictly based on training data without leakage compromising the blind test set I would use for impartial final evaluation.

I began modeling with a random forest ensemble classifier, leveraging the efficient Scikit-Learn library. This combined predictions from many distinct decision trees to improve stability and

performance over individual models. Optimizing parameters like tree count, depth and leaf size yielded strong multi-class accuracy despite some label skew across diagnoses.

With features identified, the modeling phase commenced with a random forest ensemble classifier, making use of the efficient Scikit-Learn library. Random forests are an ensemble learning method that leverages the strength of multiple decision trees. Each tree in the forest makes a prediction, and the final result is determined by a majority vote or averaging, leading to enhanced stability and performance.

Optimizing parameters such as tree count, depth, and leaf size was instrumental in achieving strong multi-class accuracy, particularly in the face of potential label skewness across various diagnoses. Random forests are known for their versatility, handling both classification and regression tasks adeptly. However, their interpretability might be limited, and understanding the individual decision-making processes of each tree in the forest can be challenging.

Seeking potential improvements in predictive performance, the next modeling technique involved the implementation of a gradient boosting model, once again relying on decision trees. Gradient boosting is an ensemble technique that builds trees sequentially, with each subsequent tree correcting the errors of the previous ones. This iterative process typically leads to a strong predictive model.

In this phase, light parameter tuning was employed to mitigate concerns related to overfitting. Overfitting occurs when a model learns the training data too well, capturing noise

and idiosyncrasies that do not generalize to unseen data. Parameter tuning involves finding the right balance to ensure the model generalizes effectively to the test partition while avoiding excessive complexity.

Feature importance statistics were also readily available with the gradient boosting model. These statistics offer insights into the contribution of each feature to the model's predictions. Understanding feature importance can be crucial in identifying the key factors influencing the target variable and, in turn, informing decision-making in the healthcare context.

Despite the incremental improvement in scoring over the random forest model, the comparative analysis of the two highlighted the nuanced trade-offs between the models. While gradient boosting demonstrated enhanced interpretability and the ability to capture complex relationships, the marginal improvement in accuracy prompted a deeper exploration of alternative approaches.

The final modeling technique involved the implementation of extreme gradient boosting via XGBoost, a state-of-the-art gradient boosting algorithm with advanced enhancements to the gradient tree concept. XGBoost has gained popularity for its efficiency, scalability, and ability to handle diverse datasets.

Additional optimizations were introduced to further refine variable usage and regularization, addressing concerns related to overfitting while simultaneously enhancing model

complexity. The concept of regularization helps prevent the model from becoming too complex, striking a balance between fitting the training data well and generalizing to new, unseen data.

The implementation of XGBoost resulted in a noticeable lift in test accuracy, underlining the effectiveness of this model in the healthcare context. The advanced features of XGBoost, such as parallel computing and regularization techniques, contributed to its superior performance. However, it's important to note that the choice of the best model depends not only on accuracy but also on considerations such as computational efficiency, interpretability, and the specific requirements of the healthcare application.

The comparative analysis of the three models—random forest, gradient boosting, and XGBoost—offers a nuanced understanding of their strengths, limitations, and suitability for healthcare applications.

Random forests, with their ensemble approach, provided a robust and stable model with commendable accuracy. However, the challenge lies in interpreting individual decision trees within the forest, making it less transparent compared to other models. The versatility of random forests, handling both classification and regression tasks effectively, makes them a valuable choice for diverse applications.

Gradient boosting, while offering improved interpretability and the ability to capture complex relationships, presented only a marginal improvement in accuracy over the random forest model. Feature importance statistics provided valuable insights into the contribution of

each feature, aiding in the identification of key factors influencing the target variable. The iterative nature of gradient boosting allowed for the sequential correction of errors, potentially capturing nuanced patterns in the data.

XGBoost, as an advanced implementation of gradient boosting, demonstrated superior performance in terms of accuracy. The additional optimizations, including parallel computing and regularization techniques, contributed to the model's efficiency and prevented overfitting despite increased complexity. The lift in test accuracy highlighted the potential of XGBoost in healthcare applications, especially when striking a balance between accuracy and model complexity is crucial.

In conclusion, the exploration and comparison of classification models in healthcare underscore the complexity of decision-making in model selection. The random forest, gradient boosting, and XGBoost models each bring their unique strengths to the table, and the choice among them should be guided by the specific requirements of the healthcare application.

The random forest, with its stability and versatility, is a reliable choice when interpretability is not the primary concern. Gradient boosting, offering improved interpretability and the ability to capture complex relationships, becomes valuable when understanding the underlying mechanisms of the model is crucial. XGBoost, with its advanced features and superior accuracy, emerges as a compelling option when striking a balance between predictive performance and model complexity is paramount.

This project not only provided a practical application of artificial intelligence in healthcare but also highlighted the importance of thoughtful model selection and parameter tuning. The journey from data exploration and feature selection to the implementation and optimization of diverse models showcased the intricate decision-making involved in real-world applications of machine learning in healthcare. As the field continues to evolve, the ability to navigate the landscape of healthcare modeling becomes increasingly vital, with each model offering a unique lens through which to uncover insights and make informed decisions for improved patient outcomes.

Citations

1. Guvenir, H.A. and Emeksiz, N. (2000) *An expert system for the differential diagnosis of erythemato-squamous diseases, Expert Systems with Applications*. Available at:https://www.academia.edu/18418572/An_expert_system_for_the_differential_diagnosis_of_erythemato_squamous_diseases.

2. *Psoriasis* (2022) *Mayo Clinic*. Available at: https://www.mayoclinic.org/diseases-conditions/psoriasis/symptoms-causes/syc-20355840.

3. Lauren, P. and Brinster, N. (2019) *Eosinophils among the histological features of psoriasis : The American Journal of Dermatopathologym*, *Lippincott Walters Kluwer*. Available at: https://journals.lww.com/amjdermatopathology/abstract/2019/05000/eosinophils_among_the_histological_features_of.3.aspx#:~:text=The%20diagnosis%20of%20psoriasis%20vulgaris%20is%20typically%20based%20on%20the,thinning%2C%20and%20neutrophils%20in%20the.

4. *Lichen Planus* (2023) *Mayo Clinic*. Available at: https://www.mayoclinic.org/diseases-conditions/lichen-planus/symptoms-causes/syc-20351378#:~:text=Overview,that%20develop%20over%20several%20weeks.

5. Arnold DL, Krishnamurthy K. Lichen Planus. [Updated 2023 Jun 1]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK526126/

6. *Pityriasis rosea* (2022) *Mayo Clinic*. Available at: https://www.mayoclinic.org/diseases-conditions/pityriasis-rosea/symptoms-causes/syc-20376405.

7. Renato Panizzon, Peter H. Block; Histopathology of Pityriasis rosea Gibert: Qualitative and Quantitative Light-Microscopic Study of 62 biopsies of 40 patients. *Dermatologica* 1 June 1982; 165 (6): 551–558. https://doi.org/10.1159/000250021

8. *Pityriasis rubra pilaris - about the disease* (2023) *Genetic and Rare Diseases Information Center*. Available at: https://rarediseases.info.nih.gov/diseases/7401/pityriasis-rubra-pilaris/.

9. Soeprono, F F. “Histologic criteria for the diagnosis of pityriasis rubra pilaris.” *The American Journal of dermatopathology* vol. 8,4 (1986): 277-83. doi:10.1097/00000372198608000-00001

10. *Seborrheic dermatitis* (2022) *Mayo Clinic*. Available at: https://www.mayoclinic.org/diseases-conditions/seborrheic-dermatitis/symptoms-causes/syc20352710#:~:text=Seborrheic%20(seb%2Do%2DREE,%2C%20ears%2C%20eyelids%20and%20chest.

11. Tucker D, Masood S. Seborrheic Dermatitis. [Updated 2023 Feb 16]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551707/#

12. *Atopic dermatitis (eczema)* (2023) *Mayo Clinic*. Available at: https://www.mayoclinic.org/diseases-conditions/atopic-dermatitis-eczema/symptoms-causes/syc-20353273.

13. *Eczema pathology* (no date) *DermNet*. Available at: https://dermnetnz.org/topics/eczema-pathology.

14. *Skin diseases: Types of, symptoms, treatment & prevention* (no date) *Cleveland Clinic*. Available at: https://my.clevelandclinic.org/health/diseases/21573-skin-diseases