Multi-class classification using Dermatology Data

GROUP 04

S14853 – PRAMUDI RAJAMANTHRI

S15030 – VIDURA CHATHURANGA

S15091 – CHALANI WIJAMUNIGE

ST 4052

Data Analysis Project I

Abstract

This report aims to conduct a comprehensive descriptive analysis of a dataset obtained from Kaggle, focusing on the challenging field of dermatology. The dataset pertains to "Erythemato-squamous" diseases, which comprises of common clinical features of erythema and scaling, with subtle distinctions, making their differential diagnosis challenging. Evaluation involved 12 clinical features and 22 histopathological features. We aim to unveil insights into these diseases, potentially assisting dermatologists in refining their differentiation and advancing diagnostic accuracy for better patient care. This report will present the findings of the descriptive analysis performed on the dataset, which shows the associations and relationships among variables and special characteristics of certain variables in differentiating Erythemato-squamous diseases, setting the foundation for the subsequent predictive modeling phase.

Contents

List of Figures

List of Tables

1. Introduction

Nidangatha Same Roga

Erythemato-squamous diseases are chronic diseases that negatively affect patients' mental and social quality of life as well as cause economic negativities with high treatment costs with high-cost drugs obtained from foreign countries. The differential diagnosis of Erythemato-squamous diseases presents a significant challenge in dermatology due to their shared clinical features of erythema and scaling, with minimal distinguishing characteristics. This group of disorders includes psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris. Accurate diagnosis often requires a biopsy, but even histopathological analysis obtained by the results of the biopsy faces difficulties as these diseases exhibit many overlapping features such as, a disease may show histopathological features of another disease at the beginning stage and may have the characteristic features at the following stages. In this data analysis project, we aim to address this diagnostic challenge by utilizing a dataset obtained from Kaggle. The dataset comprises 12 clinical and 22 histopathological features, including age and family history with various degrees of presence for each feature along with the response variable *class* which includes the above mentioned 6 erythemato-squamous disease categories. By employing a multi-class classification approach, we aim to develop a predictive model to aid in the differential diagnosis of these Erythemato-squamous diseases.

details about the skin's appearance under a microscope

2. Description of the problem

In this data analysis project, we seek to address the challenging task of differentiating the features between erythemato-squamous disease categories in dermatology. Our main objectives are to analyze the characteristics of the dataset, identifying any patterns or correlations among features, and assess the impact of age on the occurrence of these features. By gaining insights from this descriptive analysis, we aim to develop a predictive model that can improve the differential diagnostic accuracy of Erythemato-squamous disease and ultimately enhance patient care in dermatology.

3. Description of the data set

The Dermatology Dataset consists of 366 records, each with 35 attributes. Among these attributes, 32 are ordinal, while the *family_history* variable is nominal. The *age* variable is the only continuous variable, where the response variable *class* is a categorical variable with 6 levels, indicating the type of Erythemato-squamous disease. Initially patients have been first examined with 12 clinical features, after which the assessment of 22 histopathological attributes was performed using skin disease samples. Histopathological features have been identified by analyzing the samples under a microscope. If any diseases were found in the family, the family history attribute in the dataset constructed for the domain has a value of 1 (one), and if not, the value is 0 (zero). All other ordinal attributes (clinical and histopathological both) were assigned a value in the range from 0 to 3 (0 = absence of features; 1, 2 = comparative intermediate values; 3 = highest value).

Dataset: Dermatology Dataset (Multi-class classification) | Kaggle

Variable Name	Variable Type	Clinical / Histopathological	Description
erythema	Qualitative	Clinical	skin redness
scaling	Qualitative	Clinical	scaly skin.
definite_borders	Qualitative	Clinical	clear sharp border separating it from its surroundings.
itching	Qualitative	Clinical	unpleasant sensation on the skin that provokes the desire to rub or scratch the area.
koebner_phenomenon	Qualitative	Clinical	refers to when people with a specific dermatological disease manifest disease lesion in other skin lesions.
polygonal_papules	Qualitative	Clinical	presence of shiny, flat-topped and firm on palpation circumscribed elevations.
follicular_papules	Qualitative	Clinical	presence of skin lesion, less than one centimeter in diameter, circumscribed, elevated, with well-defined borders and solid content
oral_mucosal_involvement	Qualitative	Clinical	presence of skin lesions inside the mouth.
knee_and_elbow_involvement	Qualitative	Clinical	skin lesions in the knee and/or the elbow
scalp_involvement	Qualitative	Clinical	skin lesions in the scalp
family_history	Qualitative	Clinical	whether there is a family history of similar dermatological conditions
age	Quantitative	Clinical	age of the patient in years
melanin_incontinence	Qualitative	Histopathological	spillage of melanin from the basal keratinocytes into the underlying connective tissue.
eosinophils_infiltrate	Qualitative	Histopathological	bone marrow-derived cells that infiltrate skin and mucous membrane.
PNL_infiltrate	Qualitative	Histopathological	pure neuritic leprosy, no skin lesions but larger nerve trunks or their branches are enlarged accompanied with a sensory loss in the areas
fibrosis_papillary_dermis	Qualitative	Histopathological	excess development of fibrous connective tissue in the papillary dermis
exocytosis	Qualitative	Histopathological	passage to the epidermis of cells foreign to it
acanthosis	Qualitative	Histopathological	Presence of dark, velvety skin areas in body creases
hyperkeratosis	Qualitative	Histopathological	thickening of the outer layer of the skin
parakeratosis	Qualitative	Histopathological	a mode of keratinization characterized by the retention of nuclei in the stratum corneum
clubbing_rete_ridges	Qualitative	Histopathological	the epithelial extensions that project into the underlying connective tissue in both skin and mucous membranes
elongation_rete_ridges	Qualitative	Histopathological	hyperpigmentation of the basal layer in the papillary dermis
thinning_suprapapillary_epidermis	Qualitative	Histopathological	a thinning of the granular layer at the tips of the papillae
spongiform_pustule	Qualitative	Histopathological	an epidermal pustule formed by infiltration of neutrophils into necrotic epidermis in pustular psoriasis
munro_microabcess	Qualitative	Histopathological	is an abscess in the stratum corneum of the epidermis due to the infiltration of neutrophils from papillary dermis into the epidermal stratum corneum
focal_hypergranulosis	Qualitative	Histopathological	is an increased thickness of the stratum granulosum
disappearance_granular_layer	Qualitative	Histopathological	disappearance of the skin granular layer
vacuolization_damage_basal_layer	Qualitative	Histopathological	presence of vacuolisation and damage of skin basal layer
spongiosis	Qualitative	Histopathological	presence of intercellular edema
saw_tooth_appearance_retes	Qualitative	Histopathological	appearance of saw tooth patterns under the skin tissue
follicular_horn_plug	Qualitative	Histopathological	presence of follicular horn plugs
perifollicular_parakeratosis	Qualitative	Histopathological	keratinization characterized by the retention of nuclei in tissues surrounding skin follicles

inflammatory_monoluclear_inflitrate	Qualitative	Histopathological	increase in the number of infiltrating mononuclear cells in the skin
band_like_infiltrate	Qualitative	Histopathological	basal epidermis in a banded pattern
class	Qualitative	Response	the type of Erythemato-squamous disease (6 different Erythemato-squamous skin diseases)

Table 3.1

Here, class variable contains 6 different categories which are.

(1) **Psoriasis:** Chronic skin condition with red, scaly patches, sometimes affecting joints. May have psychiatric and bowel complications. Auspitz sign on removal of scales. Histopathology shows elongated rete ridges and lymphocytic infiltration.
(2) **Seborrheic Dermatitis:** Chronic inflammatory disease with oily, scaly patches on sebaceous-rich areas. Recurs with stress, depression, and fatigue. Histopathology shows epidermal spongiosis and inflammatory cell infiltration.
(3) **Lichen Planus:** Papulosquamous inflammatory disease affecting skin, nails, and mucous membranes. More common in women. Histopathology shows saw-tooth rete ridges and melanin incontinence.
(4) **Pityriasis Rosea:** Acute, self-limiting inflammatory disease with pink, scaly patches on trunk and extremities. Histopathology shows spongiosis and exocytosis.
(5) **Chronic Dermatitis (Eczema**): Recurrent, chronic inflammatory skin disease, often starting in childhood. Histopathology shows elongated rete ridges and hyperkeratosis.
(6) **Pityriasis Rubra Pilaris (PRP):** Rare inflammatory disease with unknown cause. Affects men and women, divided into five groups based on age and characteristics. Histopathology shows psoriasiform epidermis with parakeratosis and follicular infundibulum plugs.

4. Data pre-processing

The Dermatology Dataset comprises 366 records, each containing 34 attributes. Originally, 33 variables were represented as integers, while one attribute was categorical (nominal). Subsequently, the 33 integer variables were transformed into ordinal scale, allowing for ordered comparisons between their values.

Among these attributes, the "age" variable had 8 values, which were marked with '?', indicating missing data. To handle these missing values, we replaced the '?' symbols with NaN signifying missing values. Afterward, imputation was performed by replacing these NaN values based on the mean age of the training set.

During our background research on seborrheic dermatitis and its association with the koebner phenomenon, we encountered an unexpected outlier in our dataset. The koebner phenomenon, characterized by skin lesions forming at injury sites, exhibited an outlier value of "2" in the data for seborrheic dermatitis cases. To ensure data integrity and consistency in our analysis, we filtered the dataset, excluding instances where both the "class" attribute (indicating seborrheic dermatitis) and the *koebner_phenomenon* attribute was equal to 2.

With the removal of that data point, the number of records in the dataset reduced to 365.

Finally, the dataset was randomly split into training and test datasets which contains 80% and 20% of the entries from original dataset respectively.

5. Important Results of the Descriptive Analysis

The objective of the descriptive analysis in this study is to gain a comprehensive understanding of the characteristics and inter - associations of the dermatology dataset. This descriptive analysis involves summarizing and exploring the data using various statistical and visual techniques.

Figure 5.1 represents the percentage distribution of the *class* variable in the training set where the counts of the categories psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris are 85,49,59,42,43,14 respectively.

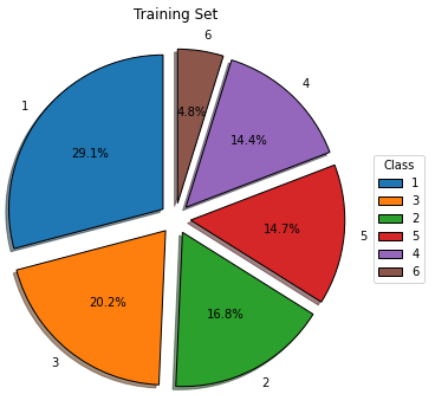

Figure 5.1

The pie chart shows that class distribution in the training set is slightly unbalanced as the representation of the minority category; pityriasis rubra pilaris (6^{th} category) is considerably lower than that of the other categories.

The *age* distribution plot in Figure 5.2 provides valuable information about the age range of patients represented in the Dermatology Dataset. As we can observe from the plot, the majority of patients fall within 10-60 years range, with a peak in the distribution indicating the most common age group.

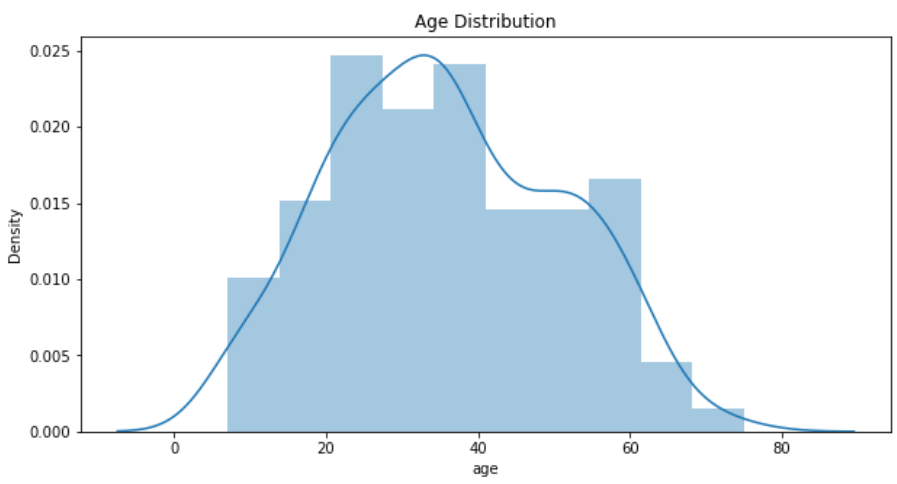

Figure 5.2

The plot exhibits an approximately symmetric bi-modal distribution, suggesting that 20-40 and 50-60 age groups might be more affected by erythemato-squamous diseases compared to others. The age variable's significance in dermatology lies in its potential role as a distinguishing factor among different skin conditions. For instance, psoriasis, a chronic autoimmune condition, often first presents in early adulthood or middle age, between 20-30 years and with a peak onset between 50 to 60 years.

sorayasis

In the article published in the National Library of Medicine in PubMed Central also Staes that ***"Psoriasis is a bi-modally distributed disease with one major age of onset at 20–30 years of age as well as a later smaller peak of onset at 50–60 years."***

saboriyak darmatayitis

On the other hand, seborrheic dermatitis, a common inflammatory skin condition, is more prevalent in infants, adolescents, and adults over the age of 50. Understanding the age distribution allows dermatologists to consider the likelihood of specific diseases based on a patient's age, assisting in narrowing down the potential diagnoses and informing the diagnostic process.

Furthermore, age can influence disease progression and treatment response. For instance, certain skin conditions may become more severe or require different management strategies as patients age.

The grouped box plots in Figure 5.3 represent the distribution of ages across different disease classes in the Dermatology Dataset.

The age distribution plot in Figure 5.3 with respect to disease classes offers valuable insights into the relationship between age and different erythemato-squamous skin conditions. The plot showcases six disease classes which are Psoriasis, Seborrheic Dermatitis, Lichen Planus, Pityriasis Rosea, Chronic Dermatitis (Eczema) and Pityriasis Rubra Pilaris (PRP) by 1 to 6 respectively. Each box plot represents the distribution of ages for patients diagnosed with the respective disease class.

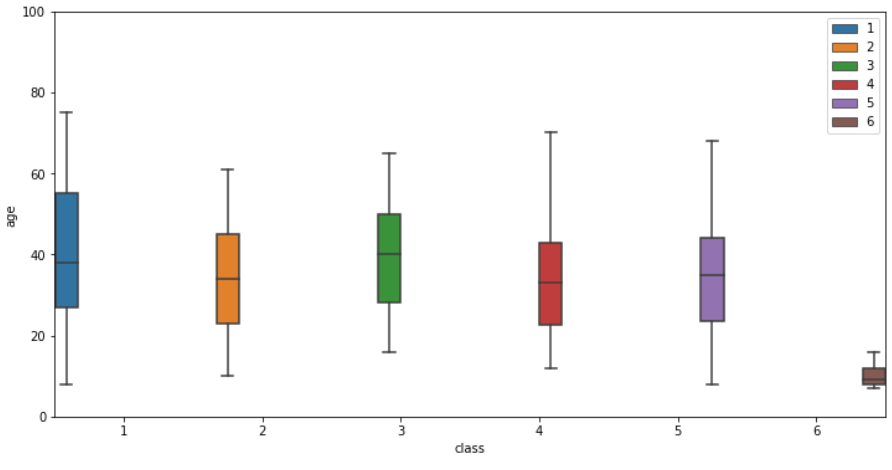

Figure 5.3

Upon analyzing the plot, it is evident that the age distribution varies among the disease classes. For instance, disease class 6 (Pityriasis Rubra Pilaris (PRP)) exhibits a relatively young age profile, with most patients falling in the age range of 7 to 16. In contrast, other disease classes show a broader age range, indicating their potential occurrence across different age groups. Psoriasis disease class shows the broadest distribution in ages from 8 years to 75 years.

The ideas we discussed here align with the findings presented in the report titled 'Diagnosis of Erythemato-Squamous Skin Diseases by Machine Learning Algorithms' authored by Mustafa Necati Bozok and Ali Calhan which states that ***"In the age distribution, the majority of patients are in their 10s, 50s and 60s. When classification is made according to age distribution, it is divided into 5 groups. More than 50% of the patients are adults and are in the first group. 80% of patients in the first group recover within 3 years. The patients in the second group are adults, but the disease progresses chronically. The patients in the third group consist of patients aged 10-50 years, and the disease characteristics are the same as in the first group. The patients in the fourth group are patients between the ages of 10-40. Well defined borders follicular hyperkeratosis and erythema are seen in the knees and elbows. Stress and dissipation are low. The patients in the fifth group are patients between the ages of 5-40 and the disease progresses chronically."***.

According to the objective of our study we examined the clinical and histopathological features which depict significant performance in the differential diagnosis of the Erythemato-Squamous disease. The stacked bar plots effectively visualize the distribution of the feature levels across various disease classes, enabling straightforward comparison and pattern identification. Each bar represents a specific class, divided into segments that depict the percentage distribution of feature levels within that class.

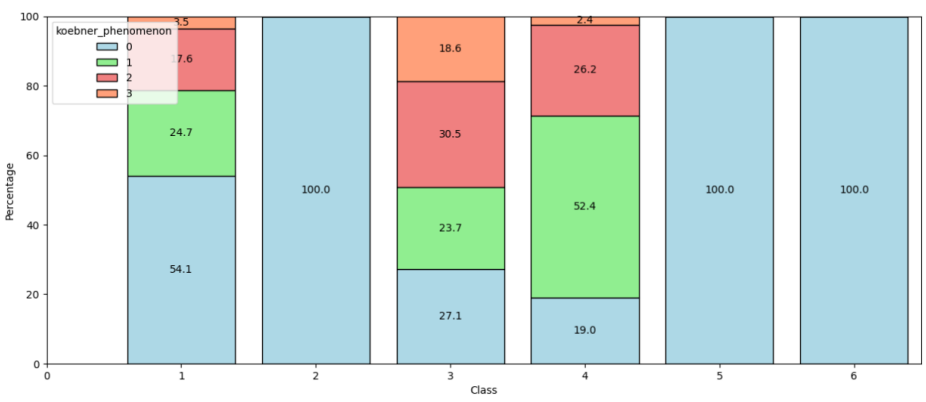

Figure 5.4

Figures 5.4 – 5.7 depicts the specifically identified clinical features namely *koebner_phenomenon, polygonal_papules follicular_papules, oral_mucosal_involvement* which clearly differentiates the disease categories of Erythemato-Squamous while the other clinical features showed equal distribution among the six categories suggesting less significance in the differential diagnosis of the disease.

miyucosal

In Figure 5.4 *koeber_phenomenon* is not found in any observations of categories 2,5 and 6. In 1st category, 54% does not show *koebner phenomenon* and in categories 3 and 4, averagely 25% of the observations do not show it either. categories 1,3 and 4 contains observations with mild to extreme levels of this phenomenon, but the extreme case is very rare. As a whole it is evident that *koeber_phenomenon* is related to the categories 1,3 and 4.

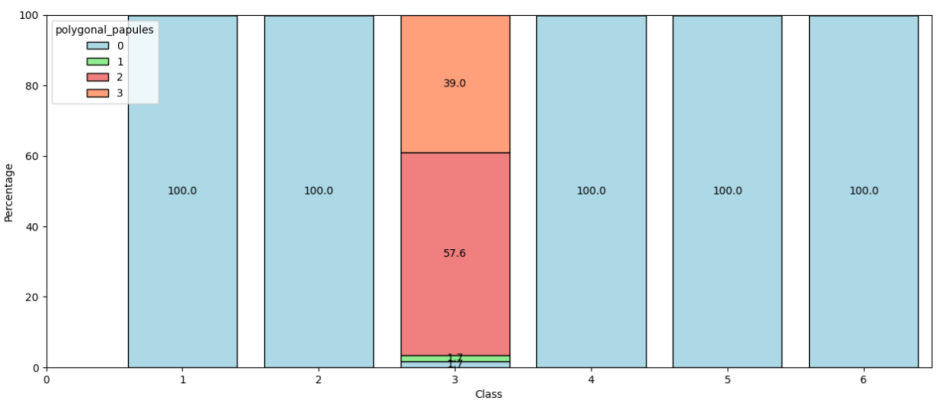

Figure 5.5

Figure 5.5 depicts that *polygonal papules* can only be observed in the 3rd category. Nearly 96% of those observations show moderately to extremely sever polygonal papules in their skin. In the 5th category, nearly 81% of the observations shows absence of this feature and nearly 18% shows mild to moderately severe cases. All the observations of disease 6 show *follicular_papules*, suggesting that follicular papules depict a significant association with category 6.

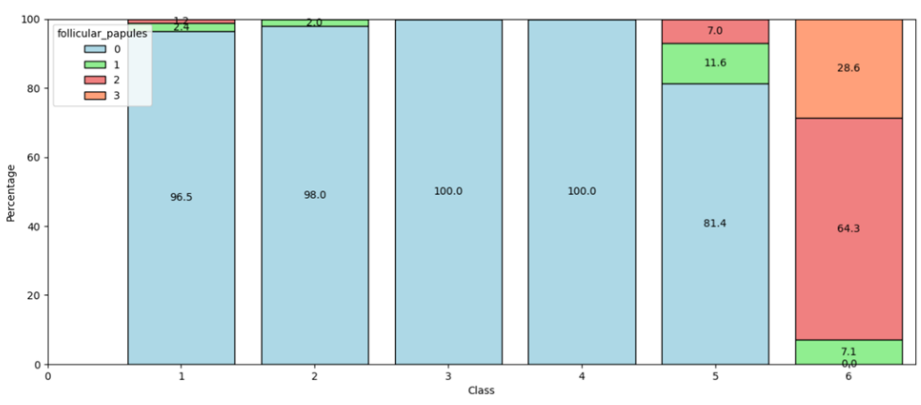

Figure 5.6

According to Figure 5.7, oral mucosal involvement is only observed within the observations of 3rd category. Nearly 93% of the observations of the 3rd category show mild to severe oral mucosal involvement and the majority of it only shows moderately severe conditions.

Other than the identified clinical features which show significance in the differential diagnosis, considering the 22 histopathological features the identified features which significantly differentiates the categories of Erythemato-Squamous disease are depicted from the stacked bar charts in Figure 5.8- 5.17.

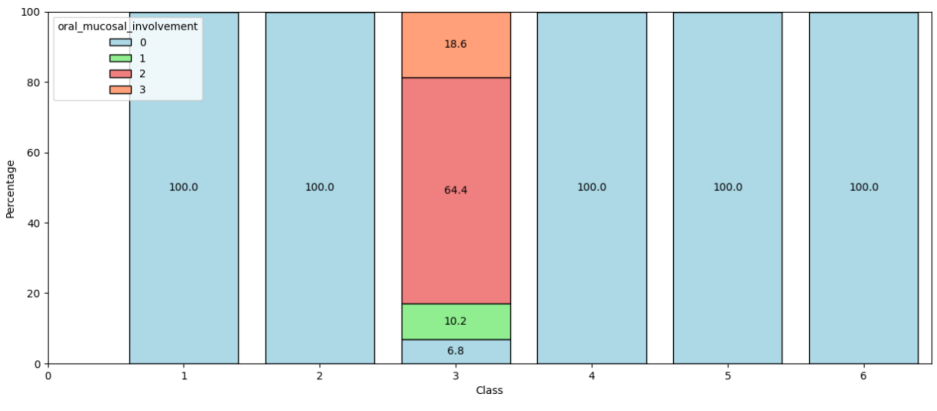

Figure 5.7

According to Figure 5.8, No patients under categories 1,2,4 and 6 showed this *fibrosis_papillary_dermis* and nearly 3% of patients under the 3rd category showed moderately severe fibrosis papillary dermis. All the observations in the 5th category showed mild to severe fibrosis papillary dermis and there were only 7% of mild cases, suggesting *fibrosis_papillary_dermis* being significant in differentiating the 5th category.

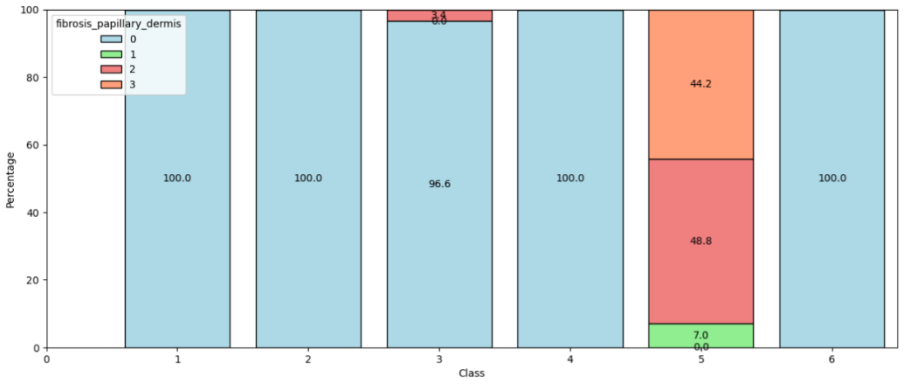

Figure 5.8

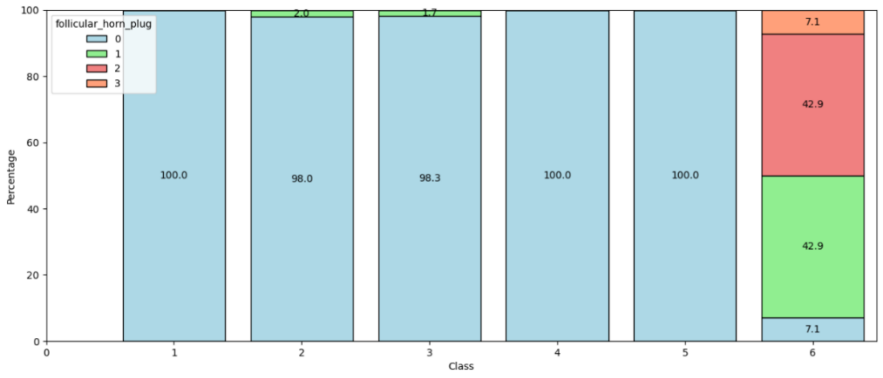

Figure 5.9

In Figure 5.9, we come across a stacked bar plot that illustrates the association between the *follicular_horn_plug* variable and the class variable. Across several classes, namely Psoriasis, Seborrheic Dermatitis, Lichen Planus, Pityriasis Rosea, and Chronic Dermatitis (Eczema), the *follicular_horn_plug* variable is predominantly characterized by level 0. In these cases, level 1 of the variable accounts for less than 5% of the data, which can be considered negligible in comparison. However, class 6, representing Pityriasis Rubra Pilaris, exhibits a starkly different behavior. This class encompasses all four levels of the *follicular_horn_plug* variable, indicating the significance of *follicular_horn_plug* feature in differentiating the 6th category.

Figure 5.10 provides insights into the association between the class variable and the *vacuolisation_damage_basal_layer* variable, a significant pattern becomes apparent. In all categories, namely Psoriasis, Seborrheic Dermatitis, Pityriasis Rosea, Chronic Dermatitis (Eczema), and Pityriasis Rubra Pilaris (PRP), the percentage representing *vacuolisation_damage_basal_layer* level 0 is 100%, depicting the significance of *vacuolisation_damage_basal_layer* in the differential diagnosis of the 3rd category. Close examination of the stacked bar plot illustrating the connection between the *focal_hypergranulosis* variable and the class variable in Figure 5.11, intriguing patterns emerge, shedding significant light on their relationship. Among the various classes represented, namely Psoriasis, Seborrheic Dermatitis, Pityriasis Rosea, Chronic Dermatitis (Eczema), and Pityriasis Rubra Pilaris, it becomes apparent that the *focal_hypergranulosis* variable predominantly exhibits level 0, showing that focal hypergranulosis is directly associated with the 3rd category.

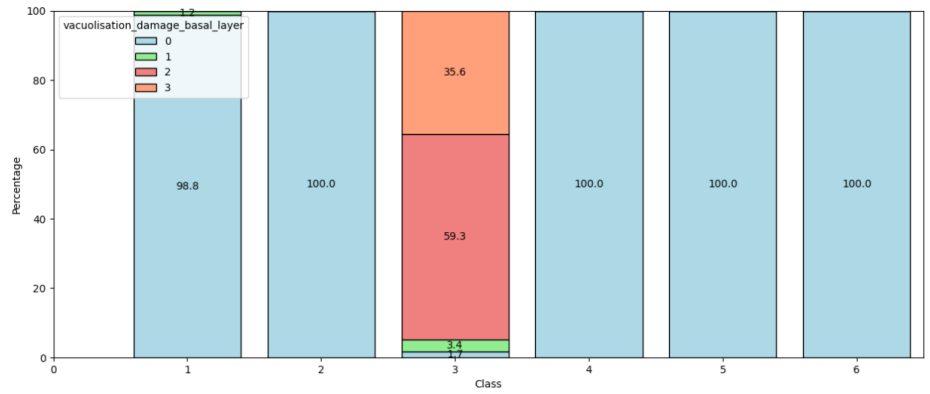

Figure 5.10

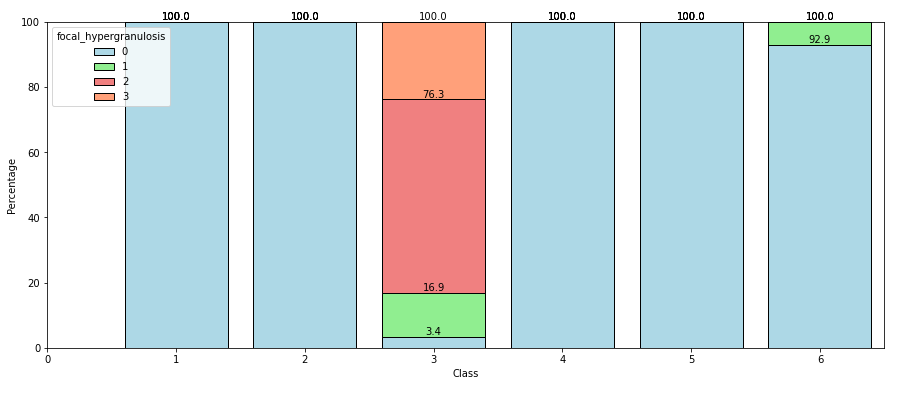

Figure 5.11

According to Figure 5.12, *melanin_incontinence* was only observed among the observations of the 3rd category. Nearly 10% of it shows mild melanin incontinence while nearly 86% of the observations show moderately severe to severe melanin incontinence. Next, we have a bar plot that illustrates the association between the class variable and the *band_like_infiltrate* variable. Upon analysis, it becomes evident that the classes, namely Psoriasis, Pityriasis Rosea, and Chronic Dermatitis (Eczema), are predominantly characterized by *band_like_infiltrate* level 0. Similarly, in the Seborrheic Dermatitis and Pityriasis Rubra Pilaris classes, over 90% of the patients also exhibit *band_like_infiltrate* level 0. Notably, the Lichen Planus class stands out as it lacks any patients belonging to level 0, with only class 2 and 3 being present in this category. These observations lead us to recognize the significant influence of the *band_like_infiltrate* variable in determining the 3rd category.

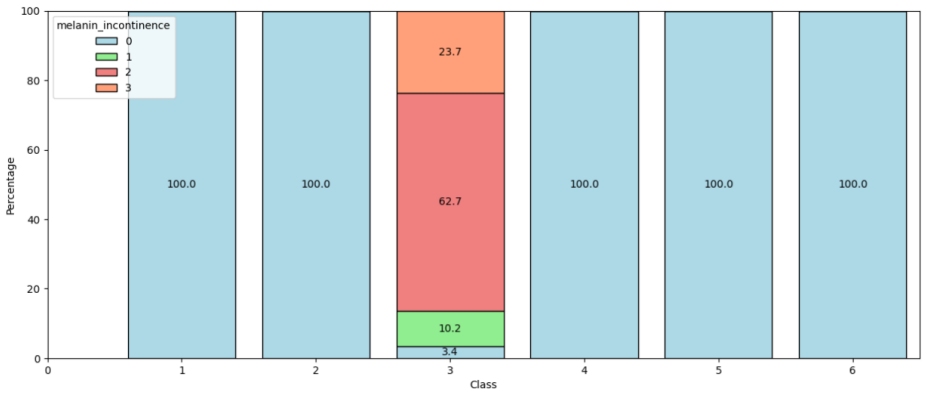

Figure 5.12

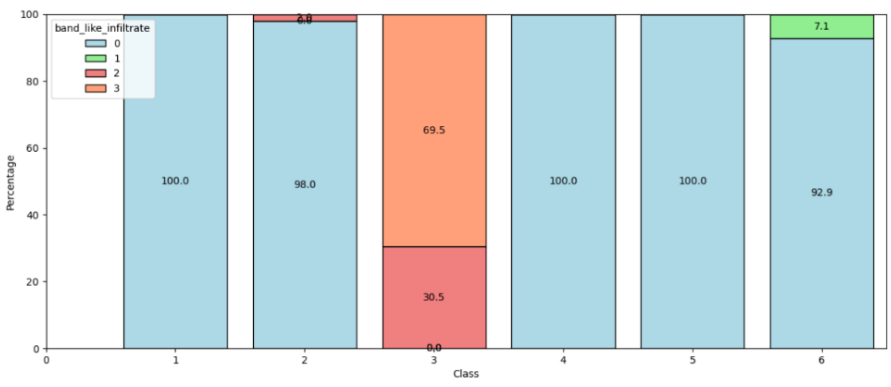

Figure 5.13

As depicted in Figure 5.14, *clubbing_rete_ridges* were not observed among patients having disease categories 2,3, and 4. Few patients (5% and 14%) having disease categories 5 and 6 only showed mild clubbing rete ridges and the rest showed none. Only, almost all of the observations under the 1st category showed mild to severe cases and 52% of them showed moderately severe clubbing rete ridges. Then, we encounter a stacked bar plot that reveals the association between the *saw_tooth_appearance_retes* variable and the class variable in Figure 5.15. Upon a thorough analysis of this plot, distinctive patterns emerge, shedding light on their relationship. Among several classes, such as Psoriasis, Seborrheic Dermatitis, Pityriasis Rosea, Chronic Dermatitis (Eczema), and Pityriasis Rubra Pilaris, the *saw_tooth_appearance_retes* variable predominantly exhibits level 0. Notably, in class 4, the occurrence of level 1 for the *saw_tooth_appearance_retes* variable is minimal, amounting to less than 5% of the data, which can be considered negligible in comparison. However, class 3, representing Lichen Planus, displays a contrasting behavior. This particular class encompasses all four levels of the *saw_tooth_appearance_retes* variable, indicating a more diverse distribution compared to other classes. These observations strongly suggest a meaningful association between the *saw_tooth_appearance_retes* variable and the 3rd category. The prevalence of level 0 across most classes implies its potential significance as a distinctive feature.

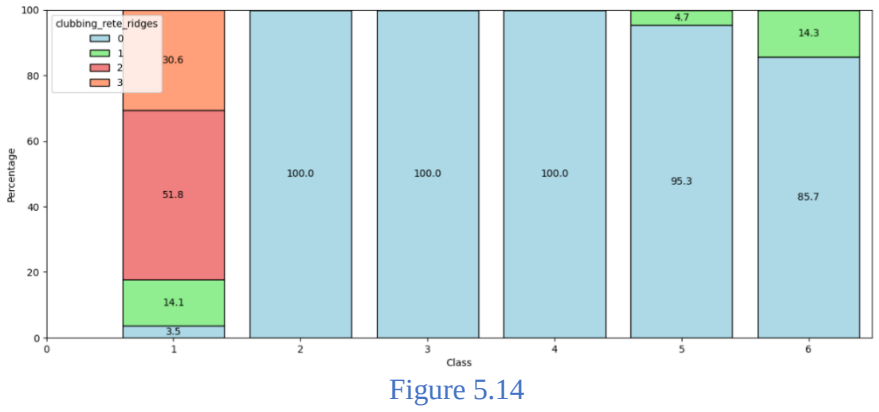

Figure 5.14

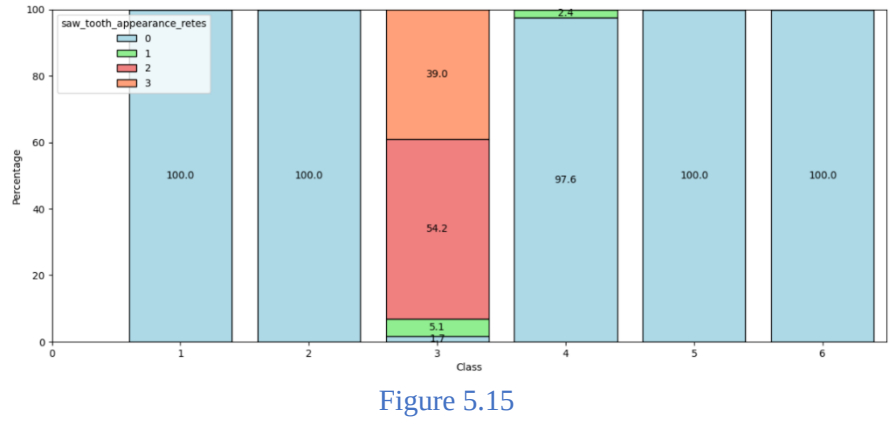

Figure 5.15

Next, we encounter another stacked bar plot that visualizes the relationship between the *perifollicular_parakeratosis* variable and the class variable in Figure 5.16. Upon careful examination of the stacked bar chart, we observe a pattern similar to the one previously observed for the *follicular_horn_plug* variable and its association with the class variable. In this case, we find that the first five classes of the class variable are predominantly characterized by *perifollicular_parakeratosis* level 0. However, a notable distinction emerges in class 6 of the class variable, where level 0 is entirely absent, and only levels 1, 2, and 3 are present. This intriguing behavior strongly suggests a distinctive association between *perifollicular_parakeratosis* and the 6th category. Thinning of suprapapillary epidermis was not observed among any patient having disease categories 3,4,5 and 6 and only 2% of the patients having disease 2 showed this condition, and that also in mild level. Almost all the observations under 1st category show this condition and the majority of it is in moderately severe level.

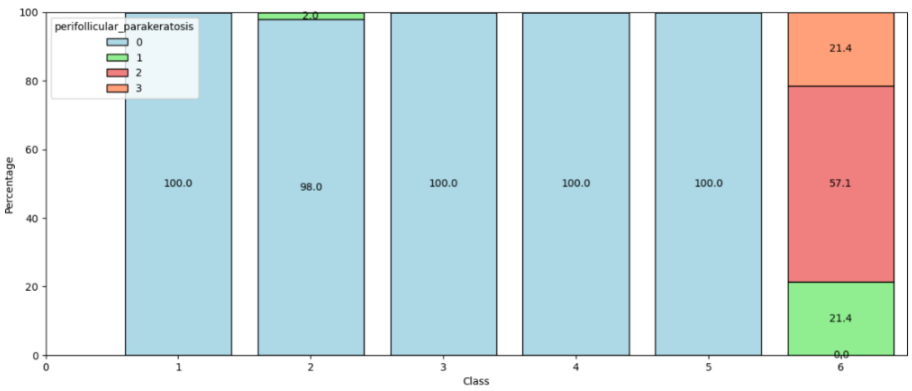

Figure 5.16

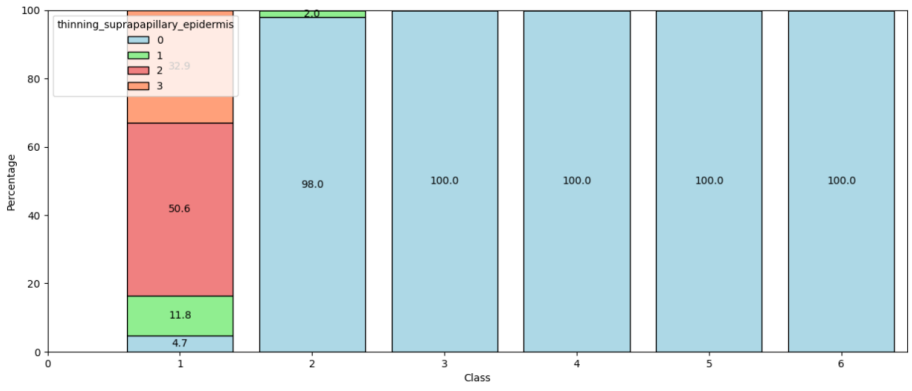

Figure 5.17

Coinciding with the above resulted findings through the stacked bar charts, Artificial Intelligence in Medicine 13 (1998) 147 – 165(provided by Bilkent University Institutional Repository) by H. Altay Gu¨venir, Gu¨ls¸en Demiro¨z, Nilsel I: lter states that ***"The koebner phenomenon is present only in psoriasis, lichen planus and pityriasis rosea. Itching and polygonal papules are for lichen planus, whereas follicular papules are for pityriasis rubra pilaris. Melanin incontinence is a diagnostic feature for lichen planus, fibrosis of the papillary dermis is for chronic dermatitis, exocytosis may be seen in lichen planus, pityriasis rosea and seboreic dermatitis. Acanthosis and parakeratosis can be seen in all of the diseases at different levels. Clubbing of the rete ridges, thinning of the suprapapillary epidermis are diagnostic for psoriasis. The disappearance of the granular layer, vacuolization and damage of basal layer, saw-tooth appearance of retes and aband-like infiltrate are diagnostic for lichen planus. Follicular horn plug and perifollicular parakeratosis are hints for pityriasis rubra pilaris."***

Moreover, Journal of Clinical & Experimental Dermatology Research states that ***"The histopathological findings of psoriasis are regular elongation of the rete ridges, elongation of the dermal papillae, edema of the dermal papillae, dilated blood vessels, thinning of the suprapapillary plate, intermittent parakeratosis, and absence of a granular layer, perivascular and dermal infiltrates of lymphocytes. The histopathological findings of seboreic dermatitis are parakeratosis, epidermal spongiosis, lymphocytic exocytosis, dermal inflammatory cell infiltration. The histopathological findings of lichen planus are saw-tooth rete ridges, atrophy, acanthosis, hyperorthokeratosis, and melanin incontinence. The histopathological findings of pityriasis rosea are spongiosis and exocytosis. The histopathological findings of cronic dermatitis are elongation of the rete ridges, prominent hyperkeratosis, and minimal spongiosis. The histopathological findings of pityriasis rubra pilaris are psoriasis form epidermis with areas of parakeratosis, plugs of the follicular infundibulum."***

Multiple Correspondence Analysis

According to Figure 5.18 a total variability of 17.46% of the categorical predictor variables is addressed by the main two components of MCA, which further depicts how the encoded variables(columns) and the records(rows) are dispersed within the two components and how they are related to each other based on their relative locations within the plot.

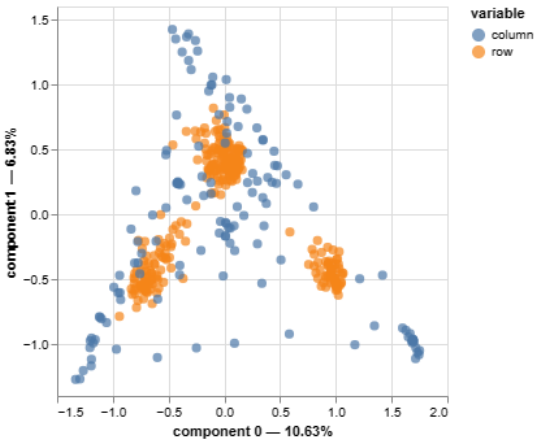

Figure 5.18

With the objective of investigating the clinical and histopathological features which are prominent in differentiating the disease categories of the Erythemato-Squamous disease, the findings from MCA in Figure 5.19 provide more insightful facts clarifying the hidden picture of Figure 5.18 as the most critical features in the diagnosis are located closer to each disease class while the less critical ones are scattered around the plot.

Particularly, the clinical features like *erythema, scaling, itching* are visible to be scattered around which further approves the results obtained from the bar charts above indicating their less prominence in the differential diagnosis of the disease.

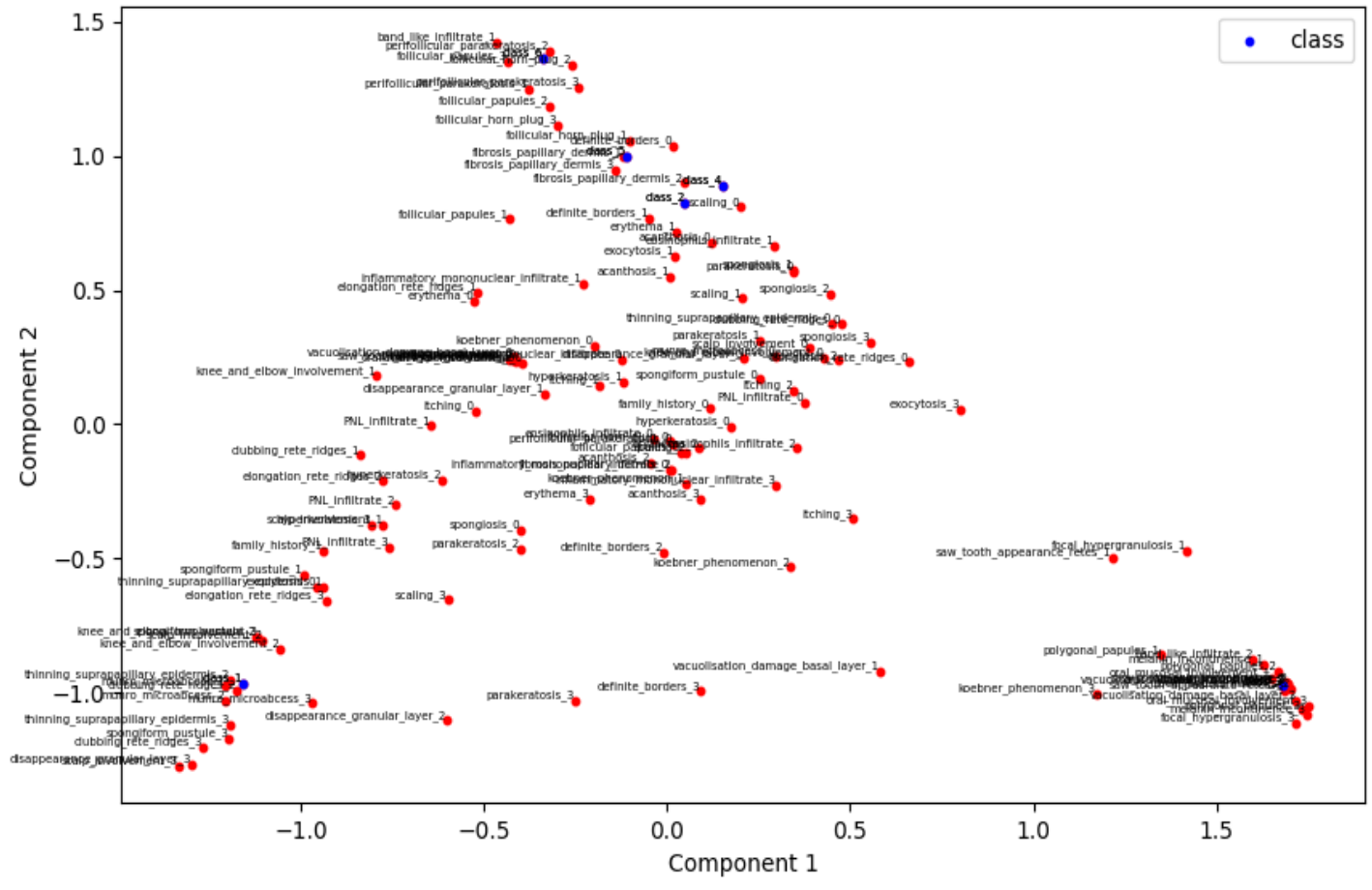

Figure 5.19

The specifically identified associations from the above MCA plot in Figure 5.19 can be summarized as follows.

class 1	clubbing_rete_ridges , thinning_suprapapillary_epidermis, disappearance_granular layer, spongiform_pustule
class 2	spongiosis, parakeratosis, exocytosis
class 3	koeber_phenomen, saw_tooth_apperance_retes, focal hypergranulosis, melanin incontinence
class 4	spongiosis, exocytosis, fibrosis_ papillary_dermis
class 5	fibrosis_ papillary_dermis, follicular_horn_plug
class 6	follicular_horn_plug, follicular_papules, parakeratosis

Table 5.1

Research square article published by Mustafa Necati BOZOK(Düzce University) and Ali ÇALHAN(Düzce University) on "Diagnosis of Erythemato-Squamous Skin Diseases with Machine Learning Algorithms" also states that ***"The histopathological findings of psoriasis are regular elongation of the rete ridges, elongation of the dermal papillae, edema of the dermal papillae, dilated blood vessels, thinning of the suprapapillary plate, intermittent parakeratosis, absence of a granular layer, perivascular and dermal infiltrates of lymphocytes (Kim et al., 2015). The histopathological findings of seboreic dermatitis are parakeratosis, epidermal spongiosis, lymphocytic exocytosis, dermal inflammatory cell infiltration (Park et al., 2016). The histopathological findings of lichen planus are saw-tooth rete ridges, atrophy, acanthosis, hyperorthokeratosis, melanin incontinence (Cheng et al.,2016). The histopathological findings of pityriasis rosea are spongiosis, exocytosis (Özyürek et al., 2014). The histopathological findings of cronic dermatitis are elongation of the rete ridges, prominent hyperkeratosis, minimal spongiosis (Bieber, 2010). The histopathological findings of pityriasis rubra pilaris are psoriasiform epidermis with areas of parakeratosis, plugs of the follicular infundibulum (Katherine L, 2003)."***

Factor Analysis for Mixed Data (FAMD)

In the differential diagnosis of Erythemato-Squamous disease age of the patients also play an important role because some histopathological features develop with aging which hints about the specific category of the ESD disease. Following is the results of FAMD analysis conducted in examining the relationships of the variables with age.

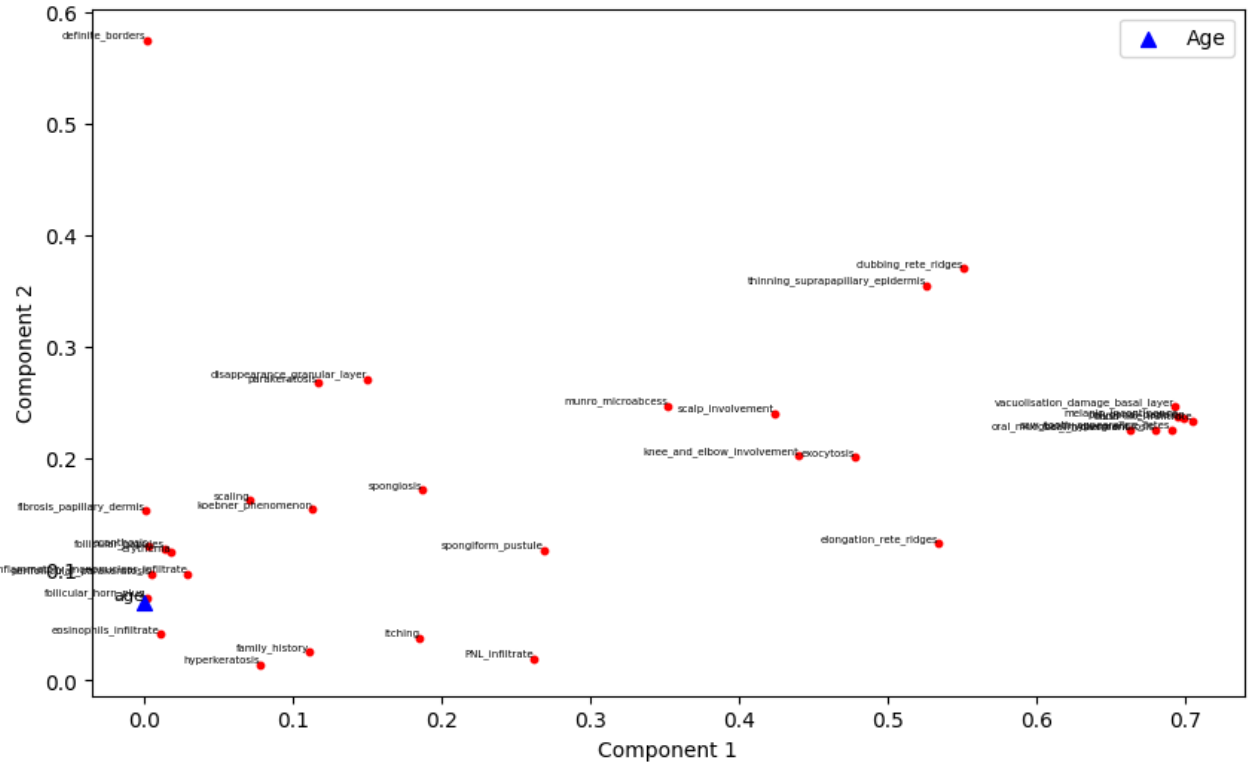

Figure 5.20

FAMD plot in Figure 5.20 results that the features *follicular_horn_plug*, *fibrosis_papillary_dermis*, *follicular_papules*, *eosinophils_infiltrate*, *inflammatory_mononuclear_infiltrate*, *hyperkeratosis* depict specific association with age considering their relative location with *age* variable suggesting an impact on those variables with aging, while features such as *vacuolization_ damage_ basal_layer*, *melanin_incontinence*, *oral_mucosal_involvement*, *clubbing_rete_ridges* depict a distant association with age which suggests a less impact of age on them.
Most of the features which are identified have an association with age are critical symptoms in identifying Pityriasis rubra pilaris(PRP), particularly features such as *follicular_horn_plug* and *follicular_papules* where article on "Diagnosis of Erythemato-Squamous Skin Diseases by Machine Learning Algorithms" in Journal of Clinical & Experimental Dermatology Research by Mustafa Necati Bozok1*, Ali Çalhan21Department of Electical-Electronik and Computer Engineering, Düzce University, Düzce, Turkey; Department of Computer Engineering, Düzce University, Düzce, Turkey states that ***"The patients in the last group are small aged children, but the disease progresses chronically"*** which approves the association of age as they are more probable to come out chronically at a small age.

6. Suggestions for Advanced Analysis

- By observing the distribution of the records over the categories of the variable *class*, 6th category is identified as the minority class with considerably less number of patient records. Hence, techniques such as SMOTE will be beneficial in making the dataset balanced in order to avoid the model fittings being biased towards the majority classes by generating synthetic samples of records.

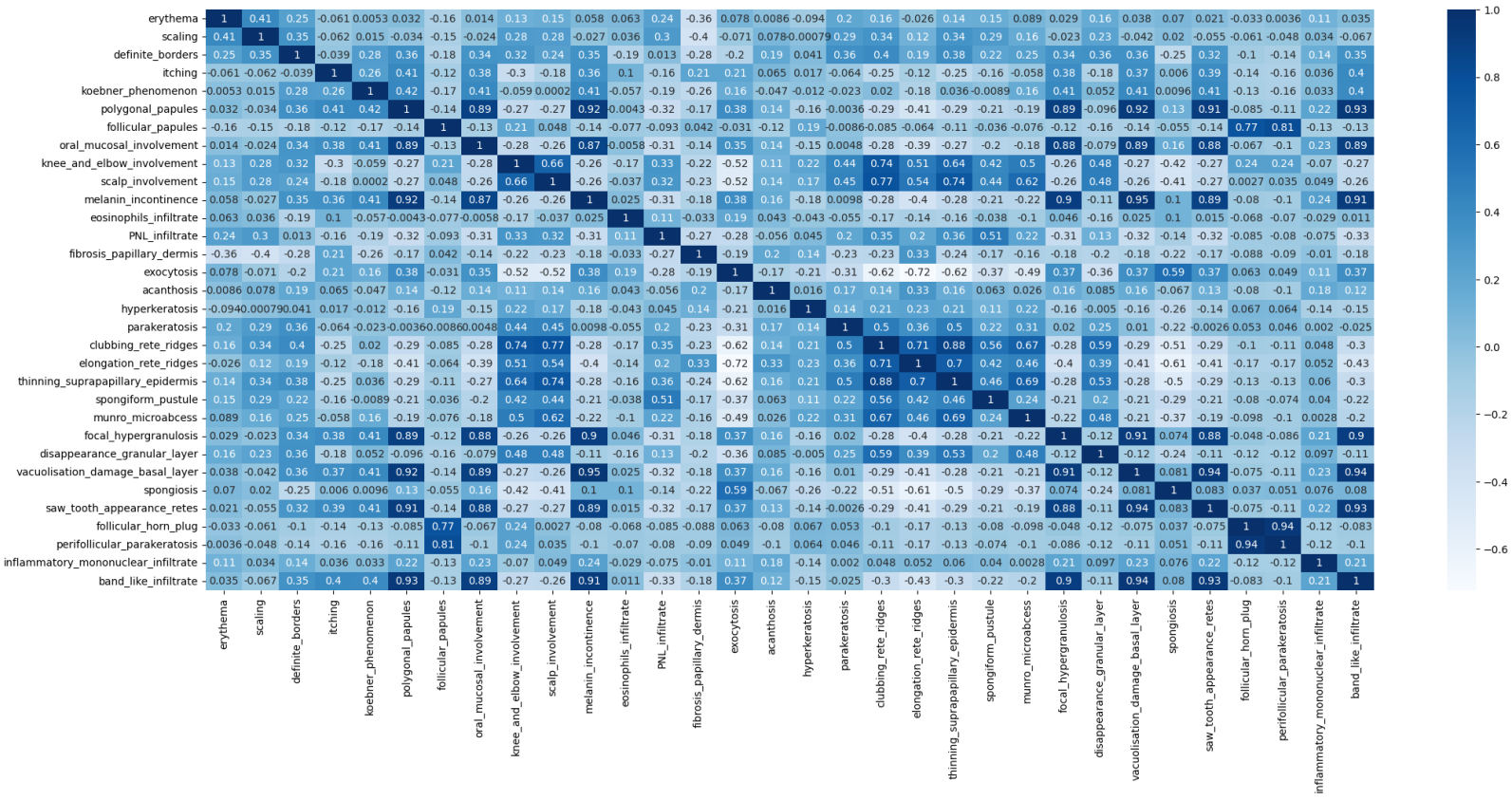

Figure 6.1

- The above spearman correlation plot depicts that there is considerable amount of significant positive and negative monotonic relationships among the ordinal variables, particularly between variables such as *polygonal_papules – focal_hypergranulosis*, *melanin_incontinence - band_like_infiltrate*, *vacuolization_damage_basal_layer – saw_tooth_appearance_retes* etc.

Despite the above identified monotic relationships, multinomial logistic regression model can be used as a base step in modelling the differential diagnosis of the dataset.

- In addition, methods such as k-NN, Naïve Bayes classification will also be appropriate as they also have resulted significantly high accuracies in dermatology classification during the past studies.

- Moreover, the tree-based methods such as Decision trees, Random Forest and Gradient boosting methods will be appropriate in this scenario as our main objective is identifying the variable importance in the differential diagnosis of the Erythemato-Squamous disease.
  The analysis done by Ashok K Singh, Chair & Professor, Department of Resorts Gaming & Golf Management, University of Nevada Las Vegas which is published in the Journal of Dermatology Research Reviews & Reports on ***"Classification of Erythemato squamous Dermatosis"*** concludes that ***" the method of RF classifier is able to classify Erythematosquamous Dermatosis with high accuracy",*** where they have identified *clubbing_rete_ridges, fibrosis_papillary_dermis, thinning_suprapapillary_epidermis, koeber_phenomen, elongation_rete_ridges, spongiosis, band_like_infiltrate, vacuolization_ damage_ basal_layer, knee_elbow_involvemnet, saw_tooth_apperance_retes, polygonal_papules, melanin_incontinence* as the prominent variables in this diagnosis.

- Multi class SVM (Support Vector Machines) will also be appropriate in this scenario as the target variable has more than 2 categories. Here, determination of the kernel function which can be used are the tradeoff parameter c and gamma on Radial Basis Function (RBF) kernel. Some experiments show that the RBF kernel produces a small classification error rate and accelerate computation compared to linear and polynomial kernel function. In addition, it is also crucial to find optimal parameter of SVM kernel function to construct a good model for diagnosis that influences the accuracy. Grid search method can be used to find the optimal parameters of SVM kernel function.

7. Appendix of the code

01.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

```
#Loading the dataset
df = pd.read_csv('dermatology_database.csv')
```

```
df.head()
```

```
df.shape
```

```
#class counts in the target variable
df['class'].value_counts()
```

```
df.info()
```

```
# NaN count
df.isna().sum()
```

Data Pre-Processing

```
age_values_with_question_mark = df[df['age'] == '?']
age_values_with_question_mark
```

```
df['age'] = df['age'].replace('?', np.nan)
```

```
age_values_with_question_mark = df[df['age'] == '?']
age_values_with_question_mark
```

```
df['class'].value_counts()
```

02.

```
df.dtypes
```

```
#Data type conversion of age
df['age'] = df['age'].astype(float)
df.dtypes
```

```
for column in df.columns:
    unique_values = df[column].unique()
    print(f"Unique values in column '{column}':")
    print(unique_values)
    print()
```

```
#Coversion of the ordinal variables accordingly for further analysis
columns_to_exclude = ['age', 'family_history','class','eosinophils_infiltrate']
columns_to_convert = [col for col in df.columns if col not in columns_to_exclude]

# Iterate over the columns and convert them to categorical
for col in columns_to_convert:
    df[col] = pd.Categorical(df[col], categories=[0, 1, 2, 3], ordered=True)
```

```
df['family_history'] = pd.Categorical(df['family_history'], categories=[0, 1])
df['eosinophils_infiltrate'] = pd.Categorical(df['eosinophils_infiltrate'], categories=[0,1,2], ordered=True)
df['class'] = pd.Categorical(df['class'], categories=[1,2,3,4,5,6])
```

```
# Get a list of column names to exclude 'age'
columns_to_convert = df.columns[df.columns != 'age']

# Convert the columns to categorical data type
df[columns_to_convert] = df[columns_to_convert].astype('category')
df.dtypes
```

03.

Removing the outlier observation

```
df = df[~((df['class'] == 2) & (df['koebner_phenomenon'] == 2))]
```

Splitting the dataset

```
from sklearn.model_selection import train_test_split

# Split the data into training and test sets
trainset, testset = train_test_split(df, test_size=0.2, random_state=42)

# Print the shapes of the resulting datasets
print("Training set shape:", trainset.shape)
print("Test set shape:",testset.shape)
```

```
trainset.isna().sum()
```

```
# Plot a histogram of the 'age' column
plt.hist(trainset['age'].dropna(), bins=10)
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Age Distribution')
plt.show()
```

04.

```
#Imputing NAs in age with mean
age_mean = trainset['age'].mean()

# Impute the mean to the NaN values in the 'age' column
trainset['age'].fillna(age_mean, inplace=True)
testset['age'].fillna(age_mean, inplace=True)
```

```
trainset.isna().sum()
```

```
testset.isna().sum()
```

```
plt.figure(figsize=(10, 5))
plt.title('Age Distribution')
sns.distplot(trainset['age'])
plt.show()
```

```
trainset['class'].value_counts()
```

```
# Plot the pie chart of 'class' variable
fig, axs = plt.subplots(1, 2, figsize=(18, 6))

# Pie chart for training set
wedges1, texts1, autotexts1 = axs[0].pie(trainset['class'].value_counts(),
                                         autopct='%1.1f%%',
                                         explode=(0.1, 0.1, 0.1, 0.1, 0.1, 0.1),
                                         labels=trainset['class'].value_counts().index,
                                         shadow=True,
                                         startangle=90,
                                         wedgeprops={'linewidth': 1, 'edgecolor': "black"},
                                         textprops=dict(color="black"))
```

05.

```
axs[0].set_title("Training Set")

# Pie chart for test set
wedges2, texts2, autotexts2 = axs[1].pie(testset['class'].value_counts(),
                                         autopct='%1.1f%%',
                                         explode=(0.1, 0.1, 0.1, 0.1, 0.1, 0.1),
                                         labels=testset['class'].value_counts().index,
                                         shadow=True,
                                         startangle=90,
                                         wedgeprops={'linewidth': 1, 'edgecolor': "black"},
                                         textprops=dict(color="black"))
axs[1].set_title("Test Set")

# Adding Legends
axs[0].legend(wedges1, trainset['class'].value_counts().index, title='Class', loc="center left", bbox_to_anchor=(1, 0, 0.5, 1))
axs[1].legend(wedges2,testset['class'].value_counts().index, title='Class', loc="center left", bbox_to_anchor=(1, 0, 0.5, 1))

# Set aspect ratio to be equal to draw perfect circles
axs[0].set_aspect('equal')
axs[1].set_aspect('equal')
```

06.

```
plt.show()
```

Grouped Boxplots

```
#Age by the disease classes
plt.figure(figsize=(12, 6))  # Set the size of the plot

# Adjust the 'width' parameter to control the width of the boxes
sns.boxplot(x='class', y='age', data=trainset, hue='class', width=1)

plt.ylim(0, 100)
plt.legend(loc='upper right')  # Show the legend
plt.show()
```

Stacked bar charts

```
# Create a contingency table with the counts of occurrences for each combination
contingency_table = trainset.groupby(['class', 'erythema']).size().unstack(fill_value=0)

# Calculate row-wise percentages
percentage_table = contingency_table.div(contingency_table.sum(axis=1), axis=0) * 100

# Reset the index to have 'class' as a column
percentage_table.reset_index(inplace=True)
```

07.

```
# Set up the custom color palette
custom_palette = ['lightblue', 'lightgreen', 'lightcoral', 'lightsalmon']

# Set the custom color palette
sns.set_palette(custom_palette)

# Set up the figure and plot
plt.figure(figsize=(15, 6))

# Manually stack the bars for each "class" and "erythema" category
bottom = np.zeros(len(percentage_table['class']))

for i in range(4):
    ax = plt.bar(percentage_table['class'], percentage_table[i], bottom=bottom, edgecolor='black')
    plt.bar_label(ax, label_type='center', fmt='%.1f')
    bottom += percentage_table[i]

plt.ylabel('Percentage')


# Set custom tick labels for the x-axis
class_labels = ['1', '2', '3', '4', '5', '6']
#plt.xticks(range(len(class_labels)), class_labels)

# Set x-axis limits with padding on the left side to start from 1
plt.xlim(0, len(class_labels)+0.5 )
```

08.

```
# Set explicit legend labels
legend_labels = ['0', '1', '2', '3']
plt.legend(title='Erythema', loc='upper left', labels=legend_labels)
plt.xlabel('Class')
plt.show()
```

```
# Create a contingency table with the counts of occurrences for each combination
contingency_table = trainset.groupby(['class', 'band_like_infiltrate']).size().unstack(fill_value=0)

# Calculate row-wise percentages
percentage_table = contingency_table.div(contingency_table.sum(axis=1), axis=0) * 100

# Reset the index to have 'class' as a column
percentage_table.reset_index(inplace=True)

# Set up the custom color palette
custom_palette = ['lightblue', 'lightgreen', 'lightcoral', 'lightsalmon']

# Set the custom color palette
sns.set_palette(custom_palette)

# Set up the figure and plot
plt.figure(figsize=(15, 6))

# Manually stack the bars for each "class" and "erythema" category
bottom = np.zeros(len(percentage_table['class']))

for i in range(4):
    ax = plt.bar(percentage_table['class'], percentage_table[i], bottom=bottom, edgecolor='black')
    plt.bar_label(ax, label_type='center', fmt='%.1f')
    bottom += percentage_table[i]

plt.ylabel('Percentage')
```

09.

```
# Set custom tick labels for the x-axis
class_labels = ['1', '2', '3', '4', '5', '6']
#plt.xticks(range(len(class_labels)), class_labels)

# Set x-axis limits with padding on the left side to start from 1
plt.xlim(0, len(class_labels)+0.5 )

# Set explicit legend labels
legend_labels = ['0', '1', '2', '3']
plt.legend(title='band_like_infiltrate', loc='upper left', labels=legend_labels)
plt.xlabel('Class')
plt.show()
```

```
# Create a contingency table with the counts of occurrences for each combination
contingency_table = trainset.groupby(['class', 'eosinophils_infiltrate']).size().unstack(fill_value=0)

# Calculate row-wise percentages
percentage_table = contingency_table.div(contingency_table.sum(axis=1), axis=0) * 100

# Reset the index to have 'class' as a column
percentage_table.reset_index(inplace=True)

# Set up the custom color palette
custom_palette = ['lightblue', 'lightgreen', 'lightcoral']

# Set the custom color palette
sns.set_palette(custom_palette)

# Set up the figure and plot
plt.figure(figsize=(15, 6))
```

10.

```
# Manually stack the bars for each "class" and "erythema" category
bottom = np.zeros(len(percentage_table['class']))

for i in range(3):
    ax = plt.bar(percentage_table['class'], percentage_table[i], bottom=bottom, edgecolor='black')
    plt.bar_label(ax, label_type='center', fmt='%.1f')
    bottom += percentage_table[i]

plt.ylabel('Percentage')


# Set custom tick labels for the x-axis
class_labels = ['1', '2', '3', '4', '5', '6']
#plt.xticks(range(len(class_labels)), class_labels)

# Set x-axis limits with padding on the left side to start from 1
plt.xlim(0, len(class_labels)+0.5 )

# Set explicit legend labels
legend_labels = ['0', '1', '2']
plt.legend(title='eosinophils_infiltrate', loc='upper left', labels=legend_labels)
plt.xlabel('Class')
plt.show()
```

```
# Create a contingency table with the counts of occurrences for each combination
contingency_table = trainset.groupby(['class', 'family_history']).size().unstack(fill_value=0)

# Calculate row-wise percentages
percentage_table = contingency_table.div(contingency_table.sum(axis=1), axis=0) * 100
```

11.

```
# Reset the index to have 'class' as a column
percentage_table.reset_index(inplace=True)

# Set up the custom color palette
custom_palette = ['lightblue', 'lightgreen']

# Set the custom color palette
sns.set_palette(custom_palette)

# Set up the figure and plot
plt.figure(figsize=(15, 6))

# Manually stack the bars for each "class" and "erythema" category
bottom = np.zeros(len(percentage_table['class']))

for i in range(2):
    ax = plt.bar(percentage_table['class'], percentage_table[i], bottom=bottom, edgecolor='black')
    plt.bar_label(ax, label_type='center', fmt='%.1f')
    bottom += percentage_table[i]

plt.ylabel('Percentage')


# Set custom tick labels for the x-axis
class_labels = ['1', '2', '3', '4', '5', '6']
#plt.xticks(range(len(class_labels)), class_labels)

# Set x-axis limits with padding on the left side to start from 1
plt.xlim(0, len(class_labels)+0.5 )
```

12.

```
# Set explicit legend labels
legend_labels = ['0', '1']
plt.legend(title='family_history', loc='upper left', labels=legend_labels)
plt.xlabel('Class')
plt.show()
```

Multiple Correspondence Analysis

```
from prince import MCA
```

```
from kmodes.kprototypes import KPrototypes
from sklearn.preprocessing import MinMaxScaler
```

```
import prince
import matplotlib.pyplot as plt
```

```
mca_cols = trainset.select_dtypes(['category']).columns
print(len(mca_cols), 'features used for MCA are', mca_cols.tolist())
```

```
trainset[mca_cols].head()
```

```
trainset[mca_cols].isna().sum()
```

```
mca_cols=['erythema', 'scaling', 'definite_borders', 'itching',
       'koebner_phenomenon', 'polygonal_papules', 'follicular_papules',
       'oral_mucosal_involvement', 'knee_and_elbow_involvement',
       'scalp_involvement', 'family_history', 'melanin_incontinence',
       'eosinophils_infiltrate', 'PNL_infiltrate', 'fibrosis_papillary_dermis',
       'exocytosis', 'acanthosis', 'hyperkeratosis', 'parakeratosis',
       'clubbing_rete_ridges', 'elongation_rete_ridges',
       'thinning_suprapapillary_epidermis', 'spongiform_pustule',
       'munro_microabcess', 'focal_hypergranulosis',
       'disappearance_granular_layer', 'vacuolisation_damage_basal_layer',
       'spongiosis', 'saw_tooth_appearance_retes', 'follicular_horn_plug',
       'perifollicular_parakeratosis', 'inflammatory_mononuclear_infiltrate',
       'band_like_infiltrate','class']
```

13.

```
# instantiate MCA class
mca = prince.MCA()
mca_data=trainset[mca_cols]
mca_data.head()
# get principal components
mca1 = mca.fit(mca_data)
```

```
mca.eigenvalues_summary
```

```
mca.row_coordinates(mca_data).head()
```

```
row_coordinates=mca.row_coordinates(mca_data)
row_coordinates
```

```
# Plot column coordinates with labels
plt.figure(figsize=(10, 6))
plt.scatter(row_coordinates[0], row_coordinates[1], marker='o', s=10, color='blue')

# Add labels for each category
for label, x, y in zip(row_coordinates.index, row_coordinates[0], row_coordinates[1]):
    plt.text(x, y, label, fontsize=5, ha='right', va='bottom')

# Add axis labels and title
plt.xlabel('Component 1')
plt.ylabel('Component 2')
```

14.

```
plt.show()
```

```
mca.column_coordinates(mca_data).head()
```

```
column_coordinates=mca.column_coordinates(mca_data)
column_coordinates
```

```
# Plot column coordinates with labels
plt.figure(figsize=(10, 6))
plt.scatter(column_coordinates[0], column_coordinates[1], marker='o', s=10, color='red')

# Add labels for each category (all six categories of the 'class' variable)
for label, x, y in zip(column_coordinates.index, column_coordinates[0], column_coordinates[1]):
    plt.text(x, y, label, fontsize=5, ha='right', va='bottom')

# Highlight the 'class' variable's categories with a different color
classes_to_highlight = ['class_1','class_2', 'class_3', 'class_4', 'class_5', 'class_6']
class_coordinates_to_highlight = column_coordinates.loc[classes_to_highlight]
plt.scatter(class_coordinates_to_highlight[0], class_coordinates_to_highlight[1], marker='o', s=10, color='blue', label='class')

# Add labels for the highlighted 'class' variable's categories
for label, x, y in zip(classes_to_highlight, class_coordinates_to_highlight[0], class_coordinates_to_highlight[1]):
    plt.text(x, y, label, fontsize=5, ha='right', va='bottom')
```

15.

```
classes_to_highlight = ['class_1','class_2', 'class_3', 'class_4', 'class_5', 'class_6']
class_coordinates_to_highlight = column_coordinates.loc[classes_to_highlight]
plt.scatter(class_coordinates_to_highlight[0], class_coordinates_to_highlight[1], marker='o', s=10, color='blue', label='class')

# Add labels for the highlighted 'class' variable's categories
for label, x, y in zip(classes_to_highlight, class_coordinates_to_highlight[0], class_coordinates_to_highlight[1]):
    plt.text(x, y, label, fontsize=5, ha='right', va='bottom')

# Add axis labels and title
plt.xlabel('Component 1')
plt.ylabel('Component 2')

# Show the plot
plt.legend()
plt.show()
```

```
mca.plot(
    mca_data,
    x_component=0,
    y_component=1
)
```

```
mca.row_contributions_.head().style.format('{:.0%}')
```

```
mca.column_contributions_.head().style.format('{:.0%}')
```

```
mca.row_cosine_similarities(mca_data).head()
```

```
mca.column_cosine_similarities(mca_data).head()
```

16.

FAMD

```
trainset_without_class = trainset.drop('class', axis=1)
famd = prince.FAMD(
    n_components=2,
    n_iter=3,
    copy=True,
    check_input=True,
    random_state=42,
    engine="sklearn",
    handle_unknown="error"
)
famd = famd.fit(trainset_without_class)
```

```
famd.eigenvalues_summary
```

```
famd.row_coordinates(trainset_without_class).head()
```

```
famd_row_coordinates=famd.row_coordinates(trainset_without_class)
```

```
famd.column_coordinates_
```

```
famd_column_coordinates=famd.column_coordinates_
```

```
# Separate the 'Age' variable from the other categorical variables
famd_column_coordinates_no_age = famd_column_coordinates.drop(index='age')
class_coordinates = famd_column_coordinates.loc['age']
```

17.

```
# Plot column coordinates without the 'Age' variable
plt.figure(figsize=(10, 6))
plt.scatter(famd_column_coordinates_no_age[0], famd_column_coordinates_no_age[1], marker='o', s=10, color='red')

# Add labels for each category (excluding the 'age' variable)
for label, x, y in zip(famd_column_coordinates_no_age.index, famd_column_coordinates_no_age[0], famd_column_coordinates_no_age[1]):
    plt.text(x, y, label, fontsize=5, ha='right', va='bottom')

# Plot the 'Age' variable with a different color or marker
plt.scatter(class_coordinates[0], class_coordinates[1], marker='^', s=50, color='blue', label='Age')

# Add labels for the 'Age' variable
plt.text(class_coordinates[0], class_coordinates[1], 'age', fontsize=8, ha='right', va='bottom')

# Add axis labels and title
plt.xlabel('Component 1')
plt.ylabel('Component 2')


# Show the plot
plt.legend()
plt.show()
```

```
famd.plot(
   trainset,
    x_component=0,
    y_component=1
)
```

18.

```
(
    famd.row_contributions_
    .sort_values(0, ascending=False)
    .head(5)
    .style.format('{:.3%}')
)
```

```
famd.column_contributions_.style.format('{:.0%}')
```

Spearman Rank Correlation

```
from scipy.stats import spearmanr
```

```
ordinal_vars = ['erythema', 'scaling', 'definite_borders', 'itching',
       'koebner_phenomenon', 'polygonal_papules', 'follicular_papules',
       'oral_mucosal_involvement', 'knee_and_elbow_involvement',
       'scalp_involvement',  'melanin_incontinence',
       'eosinophils_infiltrate', 'PNL_infiltrate', 'fibrosis_papillary_dermis',
       'exocytosis', 'acanthosis', 'hyperkeratosis', 'parakeratosis',
       'clubbing_rete_ridges', 'elongation_rete_ridges',
       'thinning_suprapapillary_epidermis', 'spongiform_pustule',
       'munro_microabcess', 'focal_hypergranulosis',
       'disappearance_granular_layer', 'vacuolisation_damage_basal_layer',
       'spongiosis', 'saw_tooth_appearance_retes', 'follicular_horn_plug',
       'perifollicular_parakeratosis', 'inflammatory_mononuclear_infiltrate',
       'band_like_infiltrate']
```

```
trainset_new = trainset_without_class.drop('age', axis=1)
trainset_new[ordinal_vars] = df[ordinal_vars].astype('int64')
trainset_new.dtypes
```

```
trainset_new.corr(numeric_only=True, method='spearman')
```

```
plt.figure(figsize=(25,10))
sns.heatmap(trainset_new.corr(numeric_only=True), annot=True, cmap='Blues')
```