Description of Dermatology

About Dataset

- The differential diagnosis of **"erythemato-squamous"** diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with minimal differences. The disorders in this group are psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris. Usually, a biopsy is necessary for the diagnosis, but unfortunately, these diseases share many histopathological features as well.
- Patients were first evaluated clinically with 12 features. Afterward, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope

Feature Value Information

In the dataset constructed for this domain, the **family history feature** has the value 1 if any of these diseases has 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, 2 indicate the relative intermediate values.

Exploration Ideas

- **Distribution of each attribute:** Explore the distribution of each attribute (column) in the dataset. You can use histograms or boxplots to visualize the distribution of each attribute and look for any patterns or outliers.
- **Correlation analysis**: Use correlation matrices to explore the relationship between the different attributes in the dataset. This can help identify which attributes are most closely related to each other and may be useful in predicting the class labels.
- **Missing values analysis**: Investigate the missing values in the Age attribute, which are represented with '?' in the dataset. Determine the proportion of missing values and evaluate whether imputation is needed.
- **Class distribution**: Explore the distribution of the class labels in the dataset. You can use bar plots to visualize the number of instances for each class, and determine whether the dataset is balanced or imbalanced.
- **Feature engineering**: Consider creating new features that may be useful in predicting the class labels. For example, you could create a feature that combines the presence of specific clinical attributes or histopathological attributes.

- **Outlier detection**: Explore the presence of any outliers in the dataset. Outliers can skew the distribution of the data and impact the performance of machine learning models. You can use boxplots or scatterplots to visualize the distribution of each attribute and identify any potential outliers.