**SKIN DISORDER PREDICTION**

**Understand the data set:**

It is a skin disorder data set contains 34 ( 33 categorical features values varies from 0 to 4 and Age is discrete Numerical feature) features and based on this features behaviour/distribution we have to predict the class of Skin disease. We have got 6 classes of skin diseases

1- psoriasis, 2- seborrheic dermatitis, 3- lichen planus, 4- pityriasis rosea, 5- chronic dermatitis, 6- pityriasis rubra pilaris

Out of 34 there are 11 Clinical features, 1 is Age and 22 are histopathological features The values of the histopathological features are determined by an analysis of the samples under a microscope

* The family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise
* Clinical Features are
  + 1: erythema ( 0 - 3 )
  + 2: scaling ( 0 -3 )
  + 3: definite borders (0 -3 )
  + 4: itching ( 0 – 3 )
  + 5: koebner phenomenon ( 0 – 3 )
  + 6: polygonal papules ( 0 – 3 )
  + 7: follicular papules ( 0 – 3 )
  + 8: oral mucosal involvement ( 0 – 3 )
  + 9: knee and elbow involvement ( 0 – 3 )
  + 10: scalp involvement ( 0 – 3 )
  + 11: family history, (0 or 1)
* Histopathological features are
  + 12: melanin incontinence ( 0 – 3 )
  + 13: eosinophils in the infiltrate ( 0 – 2 )
  + 14: PNL infiltrate ( 0 – 3 )
  + 15: fibrosis of the papillary dermis( 0 – 3 )
  + 16: exocytosis ( 0 – 3 )
  + 17: acanthosis ( 0 – 3 )
  + 18: hyperkeratosis ( 0 – 3 )
  + 19: parakeratosis ( 0 – 3 )
  + 20: clubbing of the rete ridges ( 0 – 3 )
  + 21: elongation of the rete ridges ( 0 – 3 )
  + 22: thinning of the suprapapillary epidermis ( 0 – 3 )
  + 23: spongiform pustule ( 0 – 3 )
  + 24: munro microabcess ( 0 – 3 )
  + 25: focal hypergranulosis ( 0 – 3 )
  + 26: disappearance of the granular layer ( 0 – 3 )
  + 27: vacuolisation and damage of basal layer ( 0 – 3 )
  + 28: spongiosis ( 0 – 3 )
  + 29: saw-tooth appearance of\_retes ( 0 – 3 )
  + 30: follicular horn plug ( 0 – 3 )
  + 31: perifollicular parakeratosis ( 0 – 3 )
  + 32: inflammatory monoluclear infiltrate ( 0 – 3 )
  + 33: band-like infiltrate ( 0 – 3 )
* 34: Age (linear)
* Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values

**Basic Checks:**

* Data contains 35 columns including the target variable.
* Target is ‘class’ in the dataset
* Data is very small in size with 366 datapoints
* ‘Age’ feature is having ‘?’ and in ‘objective’ type format , this ‘?’ has been replaced with mean values and typecasted to ‘int’ type
* ‘Age’ feature has one 0 value, this also imputed with mean of ‘Age’

**EDA:**

**Univariate Analysis:**

* in 99% cases whatever the skin diseases erythema is present
* even in 98% cases scaling is present in all the skin diseases
* definite border is not present in the 16% cases when diagnosed with skin diseases
* in 32% cases there are no itching symptoms present
* koebner\_phenomenon is not present in 61% cases and only 39% cases it is present
* polygonal\_papules is not present in 81% cases
* follicular\_papules is present in very rare cases i.e 19%, we have to check with target that when it is present
* Evem oral\_mucosal involvement is not present in 82% cases
* knee\_and\_elbow\_involvement is present in 39% cases only
* 72% cases doesn't have any scalp related symptoms
* 87% cases disease not observed in the family history
* melanin\_incontinence is not found in 81% cases and only in 19% cases it is found
* eosinophils\_in\_the\_infiltrate not found in 89% cases
* fibrosis\_of\_the\_papillary\_dermis is not found in 85% cases
* exocytosis found in 68% cases with any one skin disorder
* acanthosis is present in almost all the cases except few cases
* clubbing\_of\_the\_rete\_ridges is not found in 69% cases
* elongation\_of\_the\_rete\_ridges found in half of the cases and not found in half cases
* thinning\_of\_the\_suprapapillary\_epidermis is happening in 70% cases
* spongiform\_pustule,munro\_microabcess,focal\_hypergranulosis symptomsa present in 80% cases
* in 25% cases the\_granular\_layer got disappeared
* basel layer got damaged in 80% cases
* saw-tooth\_appearance\_of\_retes is not appeared in 80% cases
* follicular\_horn\_plug and perifollicular\_parakeratosis not present in 94% cases
* 96% cases having symptoms of inflammatory\_monoluclear\_inflitrate
* Age is normally distributed and it is having high relation with target class
* band\_like\_infiltrate is not present in 79% of cases
* we can see most cases are 'psoriasis' with 30.6%, and then 'lichen planus' with 19.7% and least skin diseases cases are 'pityriasis rubra pilaris'

**Bi**-**Variate** **Analysis**:

* erythema and scaling is present in almost all the cases
* definite border type-3 symptoms are not present when diseases is calss 2/4/6
* itching is not present in 60 class-1(psoriasis) cases, i can say 50%-50% cases it is present
* koebner\_phenomenon mostly not present in calss-2,5 & 6 skin diseases
* polygonal\_papules present in most of the cases when skin disease is class-3, and not present in the rest of the class
* follicular\_papules is present in all cases when disease is class-6, and not present mostly in other class
* oral\_mucosal\_involvement is there in most of the cases when skin disease is class-3, and not present in other class diseases
* in very very less cases knee\_and\_elbow\_invovement there when disease is class-2,3,4& 5, in class-1&6 it is present
* class-4&5 skin disease dont have scalp related symptoms
* when disease is class-6 , 50% cases it is having family history with disease and class-4&5 diseases doesn't have any family history with disease
* melanin\_incontinence is present only in class-3 skin disease only, not present in anyother class
* when disease is class-1&6, there are very very less symptoms of eosinophils\_in\_the\_infiltrate
* PNL\_infiltrate is not present in class-3&5, and in very less cases it is observed when disease is class-4&6.
* fibrosis\_of\_the\_papilary\_dermis is present in all the cases when disease is class-5, not completely not present in class-1,2,4 &6 , in very very less cases it is present when class-3 skin disease
* exocytosis is present in all the cases except in calss-1 it is present in very less cases.
* acanthosis symptom found in all the classes
* hyperkeratosis almost found in all the skin diseases
* when disease is class-1 , parakertosis present in most cases, and remaining classes also it is present
* clubbing of rete\_ridges found only in class- 1,5 &6
* elongation\_of\_the\_rete\_ridges present in all the cases when disease is class-1, and not found when class-3&4
* thinning\_of\_the\_suprapapillary\_epidermis is found in almost all the cases in class-1 disease, and not found mostly in other class disease
* spongiform\_pustule are present in only class-1&2 skin disease
* munro\_microabcess is present in class-1 skin disease only and in class-3&4 very very less cases it is present
* focal\_hypergranulosis is only present in class-3 skin disease
* disappearance\_of\_the\_granular\_layer is not present in class-2,5&6 skin diseases
* vacuolisation\_and\_damage\_of\_basal\_layer range of 2&3 present only in class-3 skin disease
* spongiosis is not present in class-1 skin disease
* saw-tooth\_appearance\_of\_retes range of 2&3 present only in class-3 skin disease and range-1 present only on class-3&4
* follicular\_horn\_plug & perifollicular\_parakeratosis are almost present in all cases of class-6 skin disease only
* inflammatory\_monoluclear\_inflitrate is present in all skin diseases
* band-like\_infiltrate is present in all class-3 skin diseases

**Multi-Variate Analysis:**

* people is having either symptoed with 'perifollicular\_parakeratosis' or 'koebner\_phenomenon' but not both in common
* people is having either symptoed with 'perifollicular\_parakeratosis' or 'oral\_mucosal\_involvement' but not both in common
* people is having either symptoed with 'perifollicular\_parakeratosis' or 'melanin\_incontinence' but not both in common
* people is having either symptoed with 'clubbing\_of\_the\_rete\_ridges' or 'melanin\_incontinence' but not both in common
* people is having either symptoed with 'clubbing\_of\_the\_rete\_ridges' or 'oral\_mucosal\_involvement' but not both in common
* people is having either symptoed with 'follicular\_papules' or 'oral\_mucosal\_involvement' but not both in common
* people is having either symptoed with 'follicular\_papules' or 'melanin\_incontinence' but not both in common

**Data Pre-processing:**

* There are no categorical(objective) features in the data to encode.
* From the box plot analysis, we can see few datapoint out of the upper and lower bound, but these datapoints are essential data points.
* ‘erythema’ is supposed to present in all skin disease cases but few cases it is not present. I have tried imputing but model performance getting dropped so , I haven’t treated these points.

**Feature Selection:**

* I have dropped few features having high correlation with other independent features, these are 'melanin\_incontinence','follicular\_horn\_plug','vacuolisation\_and\_damage\_of\_basal\_layer', 'band-like\_infiltrate','polygonal\_papules'

**Model Selection:**

* Since data is imbalanced, I have done SMOTE data balancing to create genuine predictive model( as without data balancing few models are performing better Logistic Regression, Decision Tree, Random Forest, Naïve Bayes). After data balancing and scaling the data SVM and KNN models performing the best.
* So, I have created models without scaling(Logistic Regression, Decision Tree, Random Forest, Naïve Bayes) and with scaling( SVM & KNeighbors).

**Model Summary:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Data Balancing & Scaling** | | **Only data Balancing not scaled the data** | | | | | | |
| **Metrics** | **SVM** | **KNN** | **LRCV** | **D T** | **RF** | **NB** | **GB** | **XGB** |  |
| **Recall**  **Class-1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |  |
| **Recall**  **Class-2** | **1** | **0.83** | **0.83** | **0.83** | **0.92** | **0.92** | **0.83** | **0.92** |  |
| **Recall**  **Class-3** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |  |
| **Recall**  **Class-4** | **1** | **1** | **0.93** | **0.93** | **1** | **1** | **0.93** | **0.93** |  |
| **Recall**  **Class-5** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |  |
| **Recall**  **Class-6** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |  |
| **Accuracy** | **1** | **0.97** | **0.96** | **0.96** | **0.99** | **0.99** | **0.96** | **0.97** |  |
|  |  |  |  |  |  |  |  |  |  |

**Result:** SVM out performed with 100 accuracy and 100% recall for all classes

**Challenges Faced:**

* Found Missing Attributes in ‘Age’ Independent feature , which are denoted as ‘?’, and the feature is of ‘objective’ type, so I have typecasted to ‘int’ type and replaced ‘?’ with mean of the ‘Age’.
* Data is imbalanced, when I try to do data balancing few models underperforming compared to models without data being balanced. To create genuine predicting model, I have done data balancing and applied the same to all the models.
* With scaling the data SVM & KNN models performing better, so have applied scaling data to only SVM & KNN models remaining models trained with normal data.
* As data contains small size, models performance changing drastically with small changes in the data, even with changes in the random state.