

Classification of Dermatological Diseases

By,
Achint Kumar

Agenda

- About the research
- Objective and Business Gains
- Dataset Description
- Dataset Preprocessing
- Statistical Insights
- Model Building
- Challenges
- Summary



About The Research

- 6 types of diseases to be classified
- 12 clinical attributes and 22 histopathological attributes considered
- Clinical trials - physical traits and symptoms
- Histopathological attributes - structures at cellular level



Problem Statement

- Classification of six types of diseases based upon the records of patients
- Misclassification of diseases leads to various complications on both sides; the patients as well as the doctors



Dataset Description

- 366 records
- Multi class Classification Problem – 6 levels
- 33 features are categorical
- Age feature is numerical
- 32 categorical features have levels as:
 - 0: Not Present in patient
 - 1,2: Intermediate levels
 - 3: Maximum Level of symptoms reached
- Family History is either 0 or 1



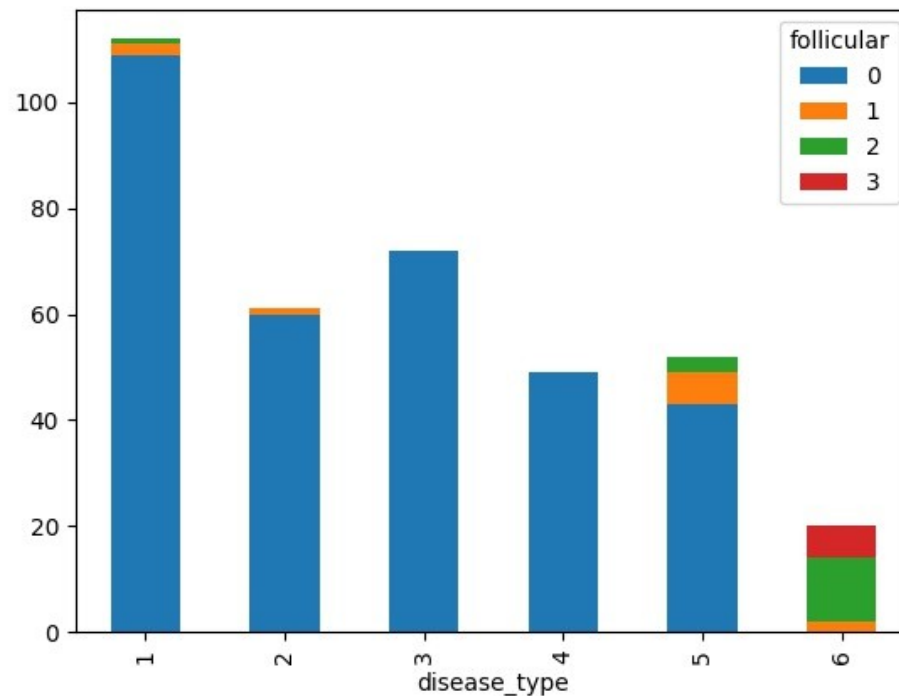
Dataset Preprocessing

- No class imbalance
- All numerical features converted back to categorical
- Age feature converted back to numerical feature
- 8 missing values imputed via mean imputation, KNN imputation and using missForest package
- The missForest package gave the best results for imputation



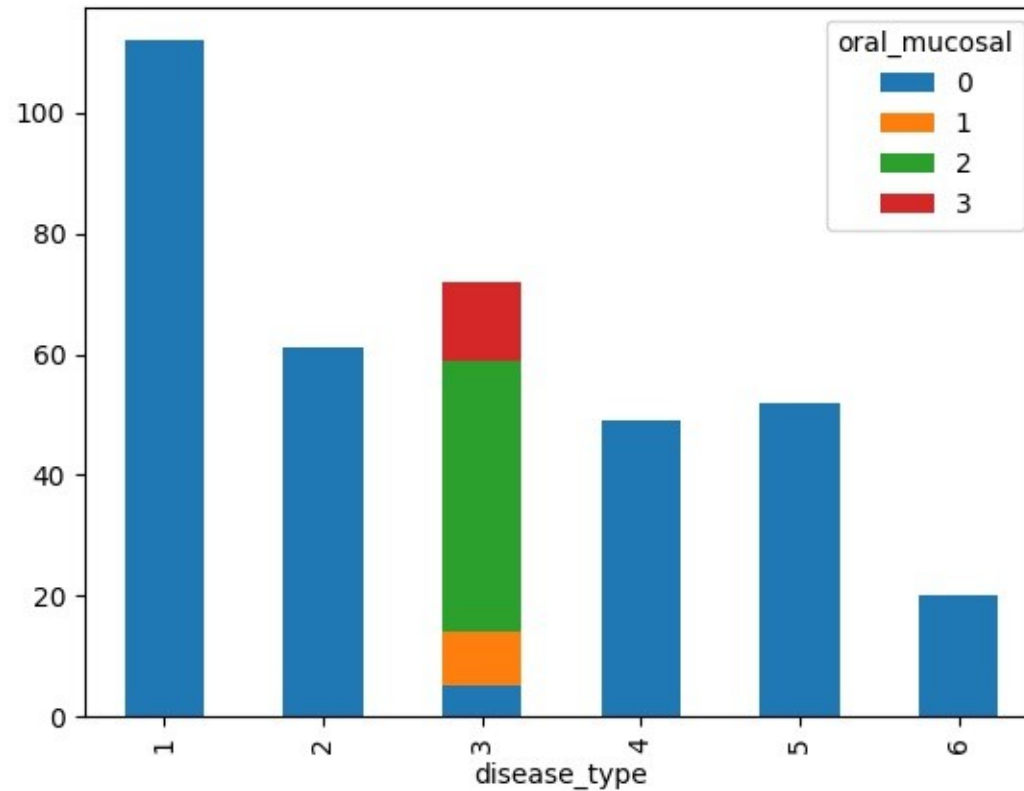
Statistical Insights : Clinical

- Clinical trials :
 - Follicular Papules main cause for Pityriasis Rubra Pilaris



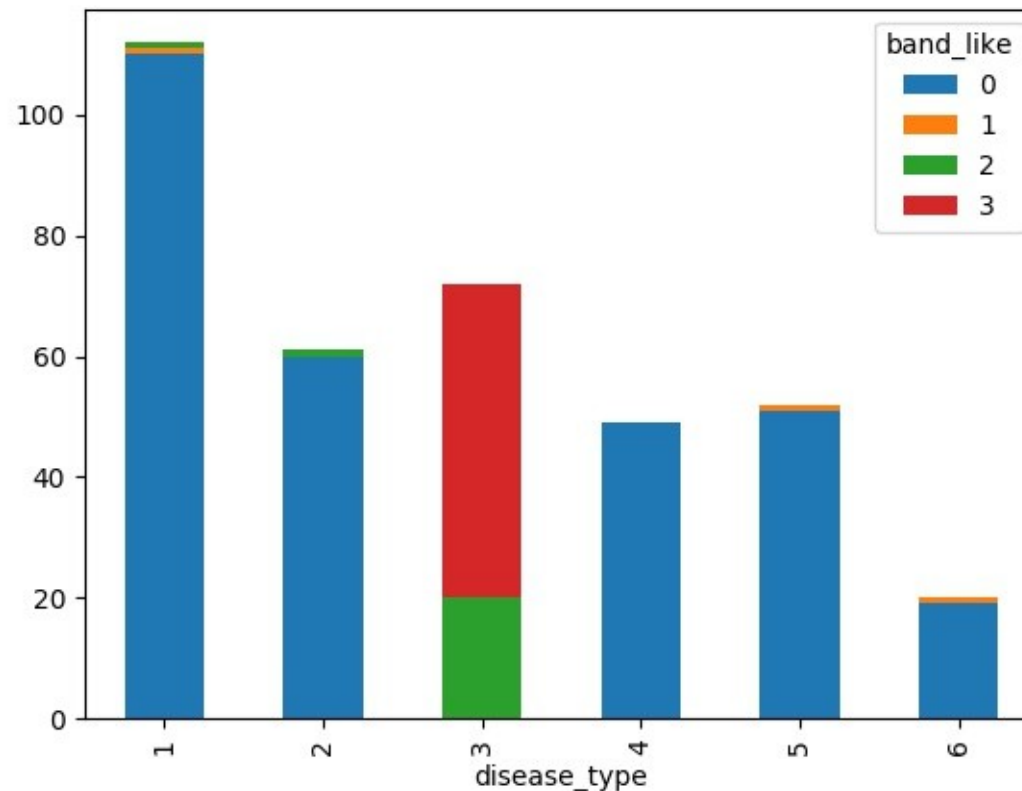
Clinical Trials

→ Oral Mucosal a huge cause for Lichen Planus



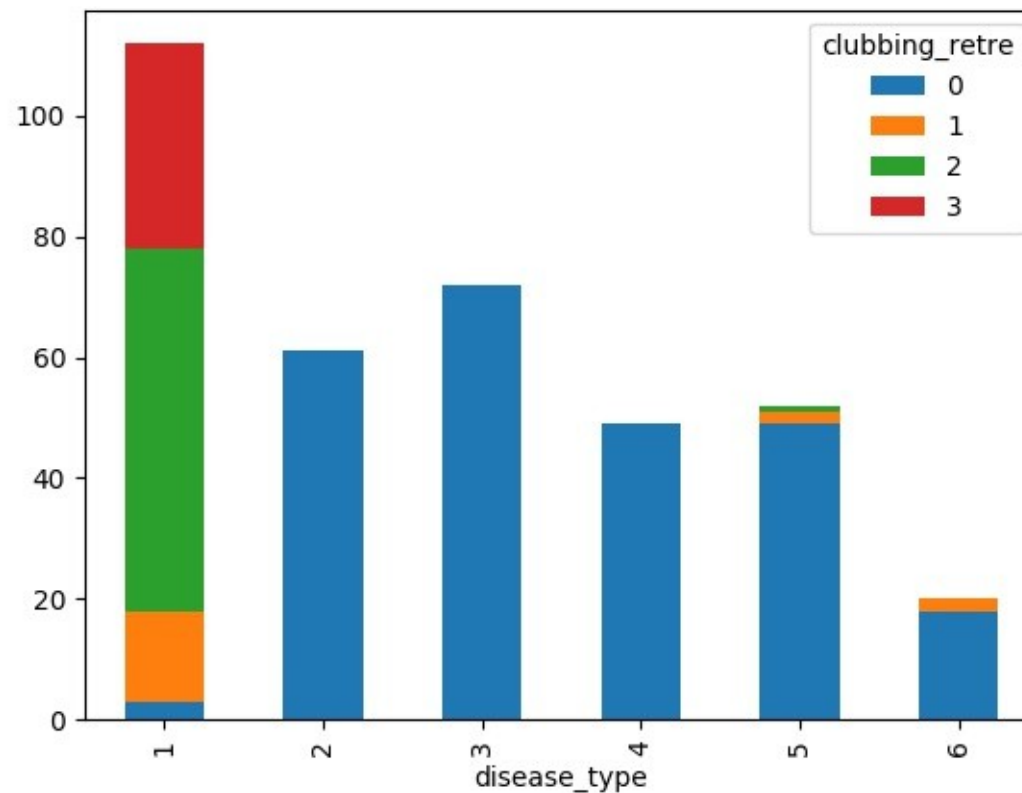
Histopathological Features

- Band like structure of cell is mainly found in patients of lichen planus



Histopathological Features

- Clubbing of the rete ridges mainly found in patients suffering from psoriasis



Histopathological Features

- Fibrosis of the papillary dermis mainly found in Chronic Dermatitis
- Focal Hypergranulosis mainly in Lichen Planus
- Horn Plug like structure found in Pityriasis Rubra Pilaris
- Melanin Inconsistencies, Munro microabscess, perifollicular parakeratosis, Saw-Tooth like structures : All show the same behaviour



Insights : Summary

- Clinical traits generally present in all 6 types of diseases
- Histopathological features are more disease-specific
- The structures and inconsistencies at cellular level more effective at classifying the diseases



Model Building

- Data split into train, validation and test in ratio of 60:20:20
- Models built:
 - Multinomial Logistic
 - C 5.0
 - CART
 - Bagging CART
 - Random Forest
 - Gradient Boost



Multinomial Logistic Model

	Accuracy	Precision	Recall	F1-Score
Class 1	94.23	100	88.46	93.87
Class 2	86.74	85.71	75	80
Class 3	100	100	100	100
Class 4	98.41	84.61	100	91.66
Class 5	99.20	91.66	100	95.65
Class 6	99.27	83.33	100	90.90



C 5.0 Model

	Accuracy	Precision	Recall	F1-Score
Class 1	98.07	100	96.15	98.03
Class 2	93.75	100	87.50	93.33
Class 3	99.18	92.85	100	96.29
Class 4	100	100	100	100
Class 5	100	100	100	100
Class 6	99.27	83.33	100	90.90



CART Model

	Accuracy	Precision	Recall	F1-Score
Class 1	96.15	100	92.30	96
Class 2	91.47	70	87.50	77.78
Class 3	100	100	100	100
Class 4	95.45	100	90.90	95.23
Class 5	100	100	100	100
Class 6	99.27	83.33	100	90.90



Bagging CART Model

	Accuracy	Precision	Recall	F1-Score
Class 1	100	100	100	100
Class 2	98.48	80	100	88.89
Class 3	100	100	100	100
Class 4	90.90	100	81.81	90
Class 5	100	100	100	100
Class 6	100	100	100	100



Random Forest

	Accuracy	Precision	Recall	F1-Score
Class 1	100	100	100	100
Class 2	99.18	92.85	100	96.29
Class 3	100	100	100	100
Class 4	92.85	100	85.71	92.30
Class 5	100	100	100	100
Class 6	100	100	100	100



Gradient Boosting

	Accuracy	Precision	Recall	F1-Score
Class 1	100	100	100	100
Class 2	98.36	86.66	100	92.85
Class 3	100	100	100	100
Class 4	85.71	100	71.42	83.33
Class 5	100	100	100	100
Class 6	100	100	100	100



Best Model

- All the models give a good result on all the error metrics
- However, random forest gives the best results



Challenges

- Method of imputation : Only 8 missing values
 - Introduced missing values at specific intervals in random columns
 - Compared imputed values with original data
- Choice of error metric



Summary

- Objective
- Insight
- Model Selection
- Challenges



Thank You

For Codes, please refer to
www.github.com/AchintKumar/Internship

