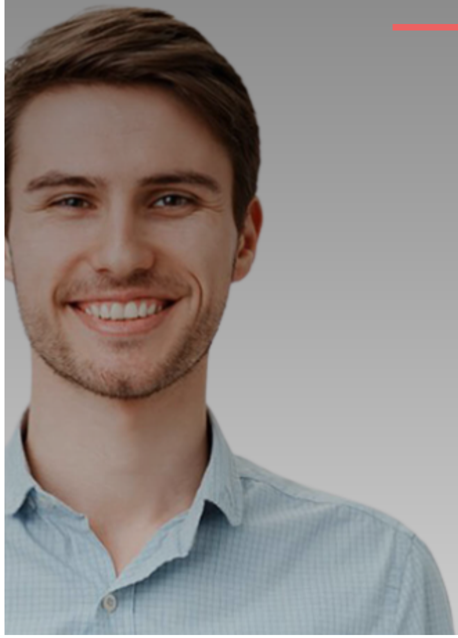
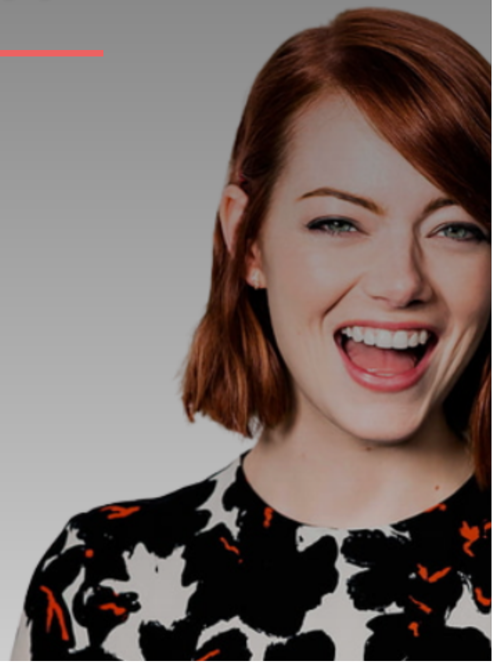
Predicting Dermatology Diseases among Malaysian

By: CPC251_Project_ Part2_Derma_2





In today's digital age, it is easier than ever to access information about dermatology diseases. However, despite the abundance of information available online, many people still have a vague understanding of these conditions. This can lead to misdiagnosis and improper treatment, which can have serious consequences for patients' health.

The Persatuan Dermatology Malaysia (PDM) and the Ministry of Health Malaysia (KKM) are working hard to educate Malaysians about dermatology diseases. However, the use of makeup tools and unregistered medicine is making this matter worse. These practices can lead to the spread of infection and the development of more serious skin conditions.

Aim : To develop a predictive model to help the Ministry of Health Malaysia (KKM) identify and treat dermatology diseases.

COULD OUR PREDICTIVE MODEL BE THE ANSWER TO KKM'S DERMATOLOGY CHALLENGES?

The development of a predictive model to help KKM identify and treat dermatology diseases is a major step forward in the fight against these conditions. This model will help to ensure that patients receive the correct diagnosis and treatment, which will improve their health and quality of life.

DATASET DESCRIPTION

	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	saw-tooth appearance of retes	follicular horn plug	perifollicular parakeratosis	inflammatory monoluclear infiltrate	band-like infiltrate	Age	Class
0	2	2	0	3	0	0	0	0	1	0	...	0	0	3	0	0	0	1	0	55	2
1	3	3	3	2	1	0	0	0	1	1	...	0	0	0	0	0	0	1	0	8	1
2	2	1	2	3	1	3	0	3	0	0	...	0	2	3	2	0	0	2	3	26	3
3	2	2	2	0	0	0	0	0	3	2	...	3	0	0	0	0	0	3	0	40	1
4	2	3	2	2	2	2	0	2	0	0	...	2	3	2	3	0	0	2	3	45	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
361	2	1	1	0	1	0	0	0	0	0	...	0	0	1	0	0	0	2	0	25	4
362	3	2	1	0	1	0	0	0	0	0	...	1	0	1	0	0	0	2	0	36	4
363	3	2	2	2	3	2	0	2	0	0	...	0	3	0	3	0	0	2	3	28	3
364	2	1	3	1	2	3	0	2	0	0	...	0	2	0	1	0	0	2	3	50	3
365	3	2	2	0	0	0	0	0	3	3	...	2	0	0	0	0	0	3	0	35	1

358 rows × 35 columns

Table 1: Dataset Dermatology

The dataset dermatology above represents patients of various ages who are afflicted with dermatology diseases. The dataset has 35 columns, including 11 columns for clinical features, 22 columns for histopathological features, a column for patient age, and a final column that indicates the disease class. Additionally, there are 358 rows in the dataset, each of which represents a patient. Since we want to know which class of diseases the patient has, "Class" is the target column. Psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris are all represented in the class by the numbers 1, 2, 3, 4, 5, and 6, respectively.

DATA ANALYSIS

The horizontal bar graph below displays features with the top ten highest information gain values. The features would be selected by us for the predictive models. These features are stored in an array called sel_features.

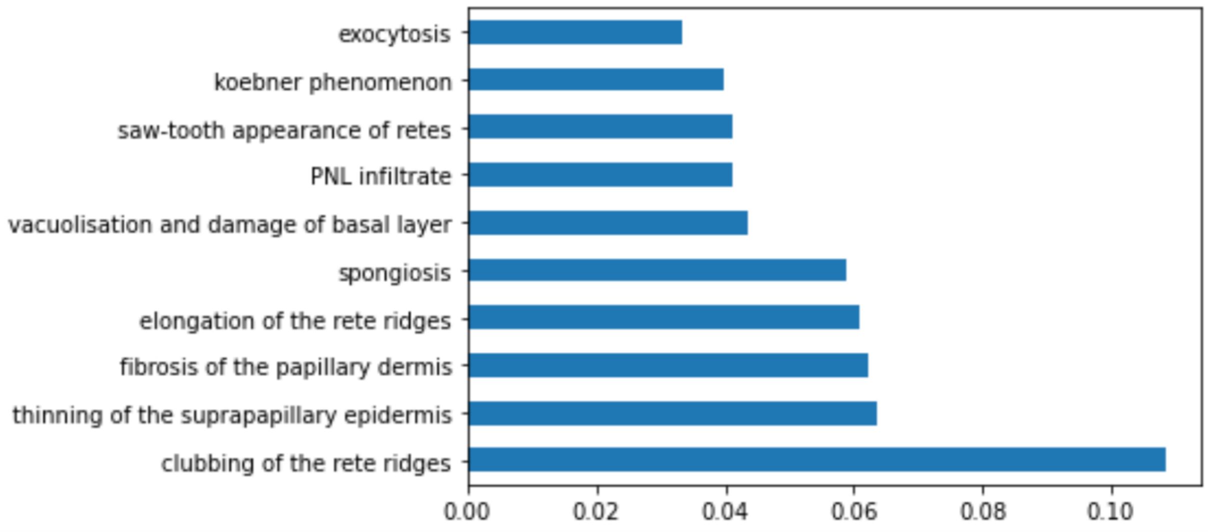


Figure 1: List of Features and Their Information Gain

Feature importance is a measure of how much each feature contributes to the accuracy of a machine learning model. It can be used to understand which features are most important for making predictions, and to identify features that may be irrelevant or noisy. We choose important gain as a method to calculate feature importance. Information gain calculates the amount of information that is gained by adding each feature to the model.

DATA MODELING

Two predictive models are built using KNearestNeighbors (KNN) algorithm and Neural Network. The method we used to divide the dataset is known as a "train-test-validation split," and it entails dividing the dataset into train sets, test sets, and validation sets in the order of 60:20:20. It is a common practice in machine learning to split the data into three sets in order to evaluate the performance of a model more accurately. The table below lists the parameters of the predictive models.

Algorithm	**Value/Statistics**
K-Nearest Neighbor	K = 17
Neural Network	Loss: categorical_crossentropy Optimizer: adam Activation function: relu and softmax
Neural Network (hyperparameter tuning)	Loss: categorical_crossentropy Optimizer: adam Activation function: relu and softmax Learning rate = 0.01, 0.001 or 0.0001
Neural Network and Support Vector Machine	Kernel regularizer = 0.01 Loss: squared_hinge Optimizer: adam Activation function: relu and softmax
Neural Network and Support Vector Machine (hyperparameter tuning)	Optimizer units: (min = 32, max = 512, step = 32) Kernel regularizer = 0.01 Loss: squared_hinge Optimizer: adam Activation function: relu and softmax Learning rate = 0.01, 0.001 or 0.0001

Table 2: Parameters of the predictive models.

The testing accuracy for KNN to choose the most suitable value of k are given below:

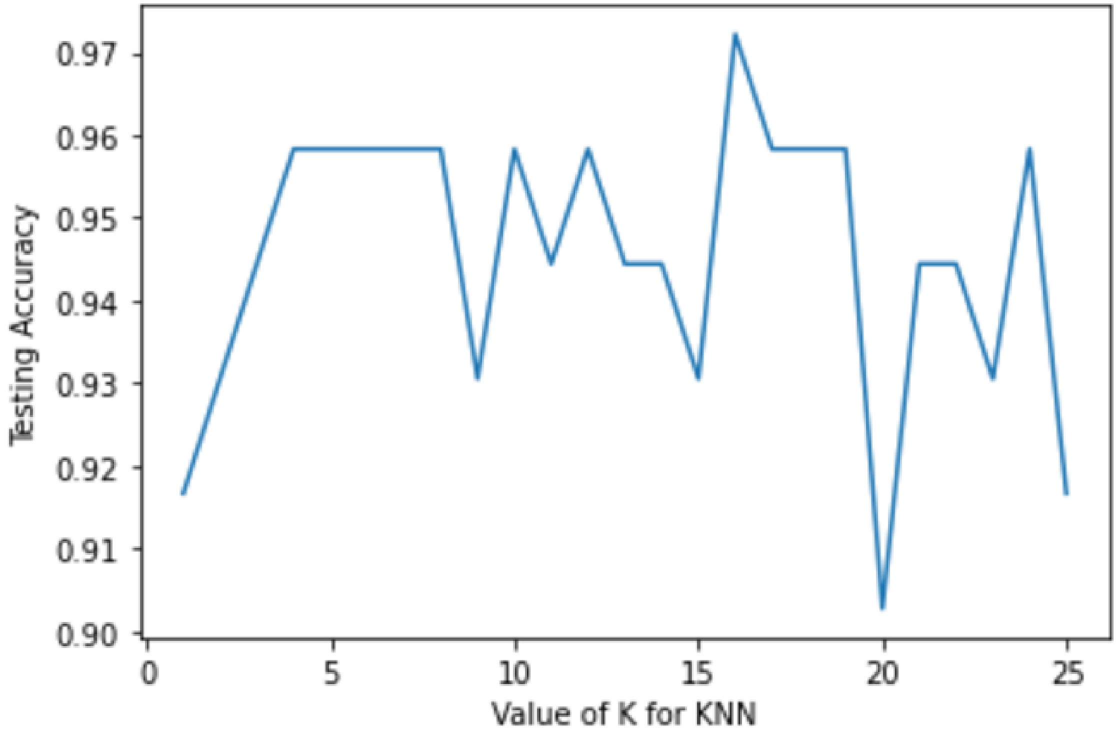


Figure 2: Testing Accuracy for KNN.

RESULT

The outcomes from each of our predicted models are given below:

```
Accuracy: 0.9583333333333334
Recall: 0.9583333333333334
Precision: 0.9642255892255892
F1 Score: 0.9576719576719577
Confusion Matrix:
 [[22  0  0  0  0  0]
 [ 0 10  0  2  0  0]
 [ 0  0 14  0  0  0]
 [ 0  0  0 10  0  0]
 [ 0  0  0  0 10  0]
 [ 0  0  0  0  1  3]]
```

Figure 3: Results of classification using KNN (without hyperparameter tuning)

```
Accuracy: 0.9444444444444444
Recall: 0.9444444444444444
Precision: 0.9478438228438227
F1 Score: 0.9440652557319223
Confusion Matrix:
 [[22  0  0  0  0  0]
 [ 0 11  0  0  0  1]
 [ 0  0 14  0  0  0]
 [ 0  2  0  8  0  0]
 [ 0  0  0  0 10  0]
 [ 0  0  0  0  1  3]]
```

Figure 4: Results of classification using KNN (after hyperparameter tuning)

> Theoretically, KNN with hyperparameter tuning using GridSearch should have better performance than KNN only. However, the result above shows the oppposite result. The performance dropped compared to normal KNN.

```
Accuracy: 0.9027777777777778
Recall: 0.9027777777777778
Precision: 0.9216686795491142
F1 Score: 0.8798029556650246
Confusion Matrix:
 [[22  0  0  0  0  0]
 [ 1 11  0  0  0  0]
 [ 0  0 14  0  0  0]
 [ 0  1  1  8  0  0]
 [ 0  0  0  0 10  0]
 [ 0  4  0  0  0  0]]
```

Figure 5: Results of classification using Neural Network (without hyperparameter tuning)

```
Accuracy: 0.9166666666666666
Recall: 0.9166666666666666
Precision: 0.9328703703703702
F1 Score: 0.8919177977148992
Confusion Matrix:
 [[22  0  0  0  0  0]
 [ 0 12  0  0  0  0]
 [ 0  0 14  0  0  0]
 [ 0  2  0  8  0  0]
 [ 0  0  0  0 10  0]
 [ 2  2  0  0  0  0]]
```

Figure 6: Results of classification using Neural Network (after hyperparameter tuning)

> Results for the predictive model that uses Neural Network with parameter tuning is slightly increased compared to the predictive model that uses Neural Network only. However, results for both predictive models still lower

than the predictive model that using KNearestNeighbors (KNN).

```
Accuracy: 0.9027777777777778
Recall: 0.9027777777777778
Precision: 0.909554482018250３
F1 Score: 0.875881114023884７
Confusion Matrix:
 [[22  0  0  0  0  0]
 [ 0 12  0  0  0  0]
 [ 0  0 14  0  0  0]
 [ 0  2  1  7  0  0]
 [ 0  0  0  0 10  0]
 [ 1  1  0  2  0  0]]
```

Figure 7: Results of classification using hybridization of Neural Network and SVM (without hyperparameter tuning)

```
Accuracy: 0.9305555555555556
Recall: 0.9305555555555556
Precision: 0.9412037037037037
F1 Score: 0.9053309602508688
Confusion Matrix:
 [[22  0  0  0  0  0]
 [ 0 12  0  0  0  0]
 [ 0  0 14  0  0  0]
 [ 0  1  0  9  0  0]
 [ 0  0  0  0 10  0]
 [ 2  2  0  0  0  0]]
```

Figure 8: Results of classification using hybridization of Neural Network and SVM (after hyperparameter tuning)

> “In comparison to other predictive models used in our study, the results from the hybrid Neural Network and SVM predictive model were the lowest. For the predictive model that uses hybridization of Neural Network and SVM with hyperparameter tuning, the result is significantly better than using Neural Network alone or

hybrid with SVM alone. Still, the predictive model with Neural Network cannot compete with KNN.

CHAMPION MODEL

```
The best accuracy for each algorithm is compared to see the best result
for K-Nearest Neighbour:          0.9583333333333334
for Neural Network:               0.9166666666666666
for Neural Network with SVM:      0.9305555555555556
```

Figure 9: Predictive model that become our Champion Model.

Conclusion: KNN without hyperparamater tuning is the champion model because it has the highest values in performance metrics which are accuracy, precision, recall, F1 score and confusion matrix. Hence, KNN without hyperparamater tuning will be used to predict the output in test set.

OUR TEAM

MOHAMAD NAZMI BIN HASHIM (158616)

MUHAMMAD KHAWARIZMI BIN JEFRI (158520)

MIOR MUHAMMAD IRFAN BIN MIOR LATFEE (158450)

MUHAMMAD HAIQAL BIN RAFIQUZZAMAN (158852)