Predicting Diagnosis of Skin Condition using Images

Sophia Song
Data Science Career Track, 2019

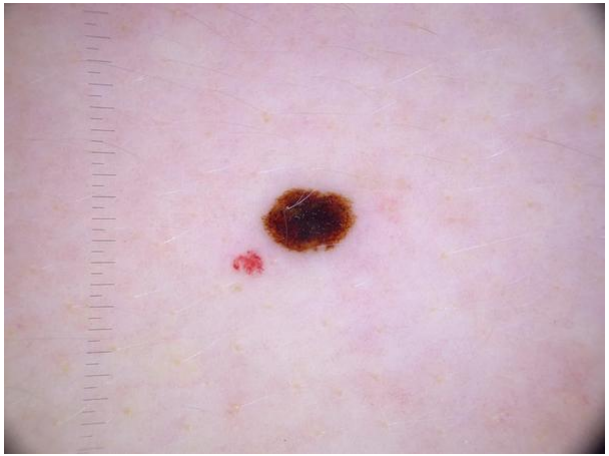
Introduction

Skin cancer is the most prevalent form of cancer in the US, with more than 9,500 cases diagnosed per day.

Early diagnosis is crucial for optimal treatment, however, it is not always possible.

Challenge

Non-Cancerous



Cancerous

Our goal: To implement a classifier using images.

More exactly:

Classify image to out of 7 possible categories (Listed below) using the image features and/or metadata. Diagnoses in bold are cancerous.

- **Melanoma**
- **Basal cell carcinoma**
- Actinic keratoses
- Benign keratosis-like lesions
- Dermatofibroma
- melanocytic naevus
- Vascular lesions

Data Wrangling and EDA

Dataset

The HAM10000 dataset, which contains 10015 images of skin conditions and a .csv metadata.

Metadata columns: lesion_id, image_id, dx, dx_type, age, sex, localization.

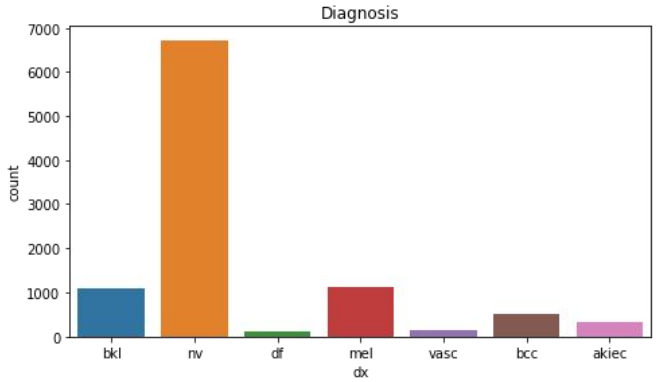
All columns in the metadata were categorical with the exception of age, which is numerical.

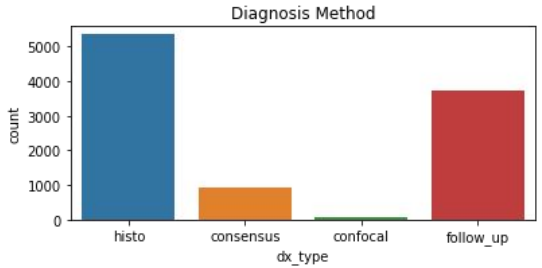
Missing values

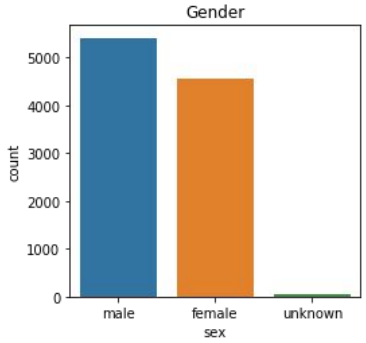
At initial look, the only column which contained any null values was the “age” column. However, there is one string in the gender and location columns, “unknown”, that refers to values not known or missing.

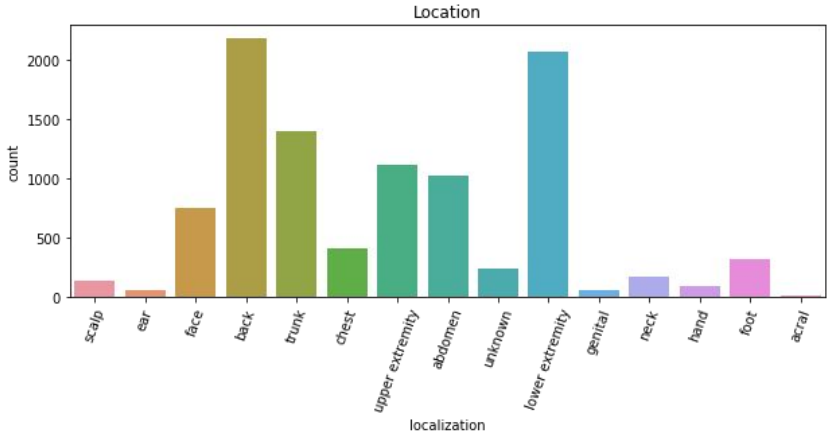
I imputed null numerical and categorical variables using Random Forest Regressor and Random Forest Classifier, respectively.

EDA - categorical variables

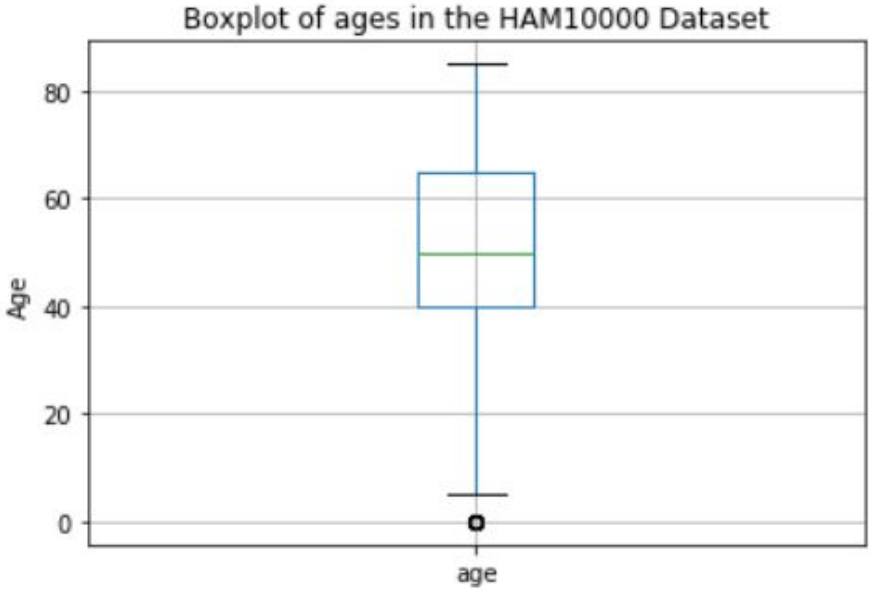


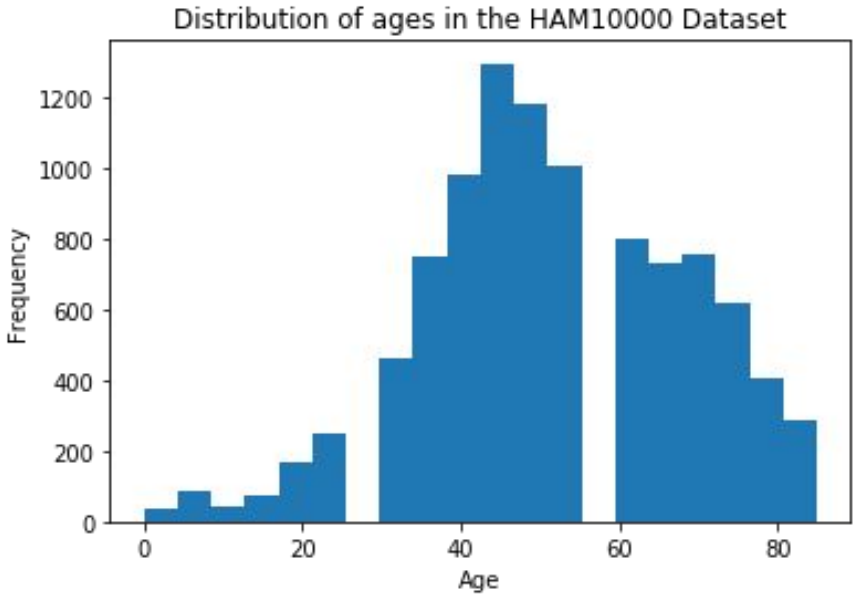






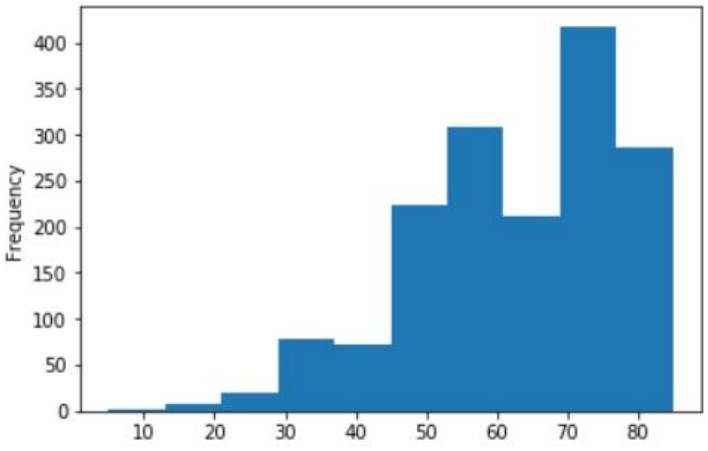
EDA - numerical variable

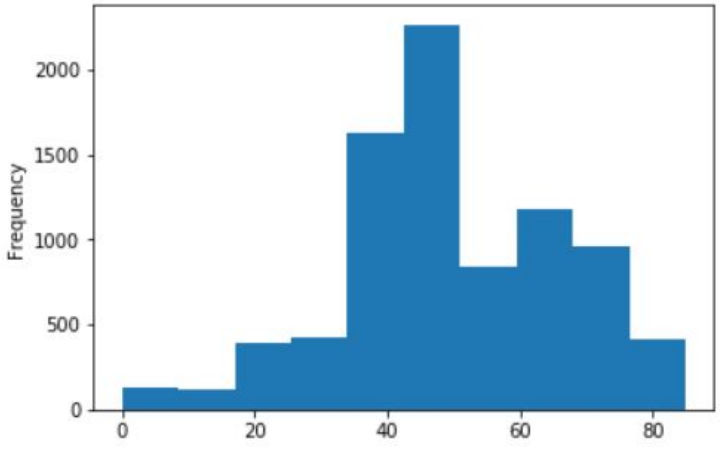




Inferential statistics:

Question: Are patients with cancerous diagnoses older than patients with non-cancerous diagnoses?





There are 1637 cancerous cases and 8388 non-cancerous cases.

Mean age for patients with cancerous conditions: 62.65 +/- 15 years

Mean age for patients with non-cancerous conditions: 49.77 +/- 16.53 years.

With an alpha of 0.05 and the values as listed above, I calculated a t-statistic of 31.11. The p-value is very small, and definitely below our alpha, so we accept the alternative hypothesis: that the age for patients with cancerous conditions are higher than those of patients with non-cancerous conditions.

Machine Learning

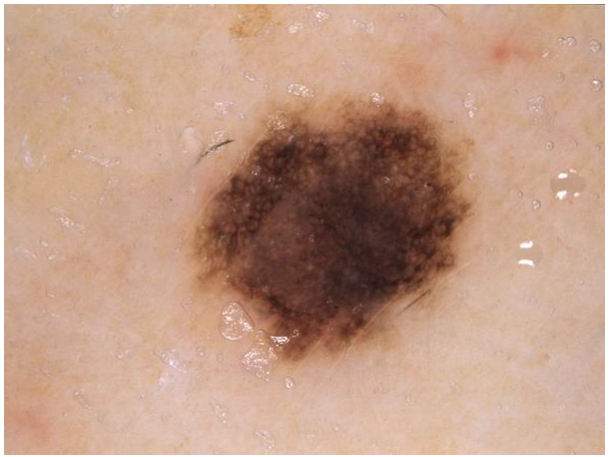
Machine learning steps

1. Using VGG as a feature extractor.
2. Dimension reduction using PCA
3. Fitting of models

VGG

VGG is a convolutional neural network model for image recognition proposed by the Visual Geometry Group from the University of Oxford. It achieved 92.7% top-5 test accuracy in ImageNet, a dataset of over 14 million images belonging to 1000 classes, and has two subtypes, VGG16 and VGG19, that contain 16 and 19 weight layers, respectively.

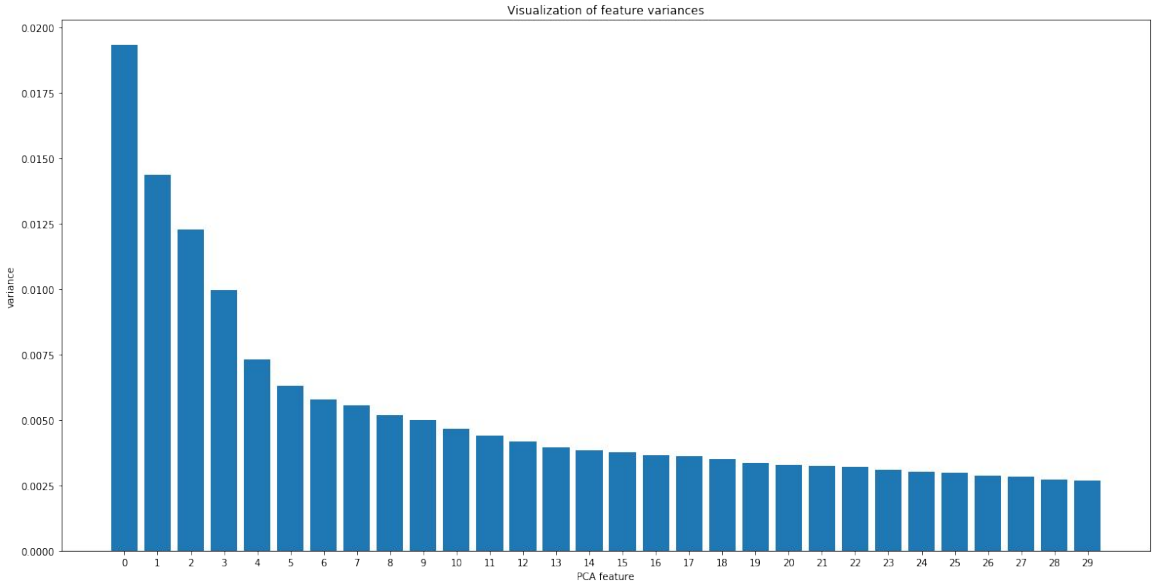
Part of the code was borrowed from:
https://medium.com/@franky07724_57962/using-keras-pre-trained-models-for-feature-extraction-in-image-clustering-a142c6cdf5b1.



	0	1	2	3	4	5	6	7	8	9	...	25080	25081	25082	25083	25084
0	-0.000000	-0.0	-0.000000	-0.0	-0.0	-0.000000	-0.0	-0.0	-0.0	-0.0	...	4.640642	-0.000000	-0.0	-0.000000	-0.000000
1	5.990223	-0.0	-0.000000	-0.0	-0.0	-0.000000	-0.0	-0.0	-0.0	-0.0	...	18.183280	-0.000000	-0.0	-0.000000	1.999651
2	5.352508	-0.0	1.049744	-0.0	-0.0	-0.000000	-0.0	-0.0	-0.0	-0.0	...	26.286820	-0.000000	-0.0	-0.000000	-0.000000
3	-0.000000	-0.0	0.653627	-0.0	-0.0	-0.000000	-0.0	-0.0	-0.0	-0.0	...	4.307326	-0.000000	-0.0	2.370319	-0.000000
4	-0.000000	-0.0	15.096593	-0.0	-0.0	4.645373	-0.0	-0.0	-0.0	-0.0	...	-0.000000	5.191962	-0.0	-0.000000	-0.000000

PCA

2831 features were needed to capture 95% of our variance,



Models used

- SVM (LinearSVC)
- Naive-Bayes
- DecisionTree
- RandomForestClassifier

Each of these models were fitted on 4 combinations of our info:

1. Metadata
2. Unreduced image features
3. Reduced image features
4. Combination of metadata and reduced image features.

Results

F1-scores

	Models				
Dataset	SVM	Naive Bayes	DecisionTree	RandomForestClassifier	Average
Metadata	0.72	0.51	0.73	0.72	0.67
Reduced Image Features	0.69	0.66	0.63	0.70	0.67
Unreduced Image Features	0.77	0.56	0.63	0.70	0.67
Combination of Image and Metadata	0.65	0.53	0.73	0.75	0.67
Average	0.71	0.57	0.68	0.72	

Key observations:

- There is no particular dataset that performed better than the rest.
- Naive Bayes performed significantly poorer on all datasets except the one containing PCA-reduced features.
- RandomForestClassifier performing the best across all four datasets, but SVM produced the model with the highest f1-score using unreduced image features.
- Both DecisionTree and RandomForestClassifier performed better on datasets containing categorical variables.

Conclusion

In this capstone project, I tested multiple classifier models on image features extracted using VGG in combination with tabular metadata.

The f1-scores ranged from 0.51 (Naive Bayes with metadata) to 0.77 (LinearSVC with unreduced image data).

The model that I'd use is RandomForestClassifier (with f1-scores from 0.70 to 0.75), because it performed the best across multiple datasets, is less prone to overfitting, and doesn't require a lot of hyperparameter tuning.