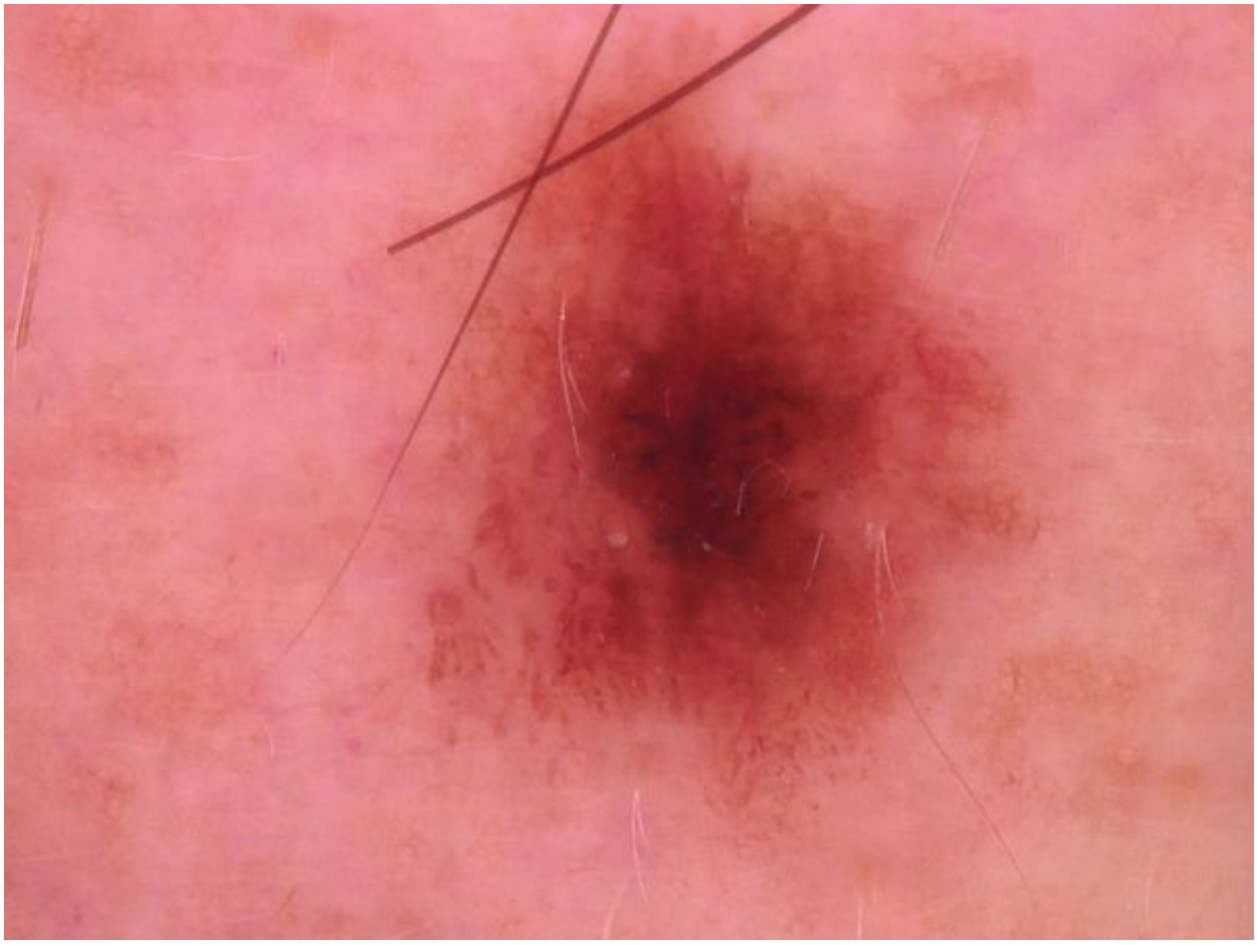


Capstone I

Predicting Skin Condition Based on Image

Shangshang (Sophia) Song

Mentor: Thyago Porpino

Springboard Data Science Career Track 2019

Introduction:

Skin cancer is the most common cancer in America, with more than 9,500 new cases diagnosed every day. As of 2018, the annual cost of treating skin cancer is $8.1 billion, and the Center for Disease Control (CDC) called it “a growing public health issue”.

Skin cancer has a high survival rate when caught early on; however, if left untreated, invasive cancerous cells can migrate to other parts of the body via the blood or lymph supply and become lethal. *Thus, early detection and treatment is crucial for optimal recovery.* The problem is, however, that many benign conditions of the skin appear similar to skin cancer, and for an untrained individual it may be difficult to differentiate between such a harmless condition and something more malign.

In this capstone project, I attempt to build and train a classification model using images from the HAM10000 dataset in combination with patient information obtained from the dataset’s metadata. The model will classify each image as belonging to one of seven possible categories, 2 of which are cancerous.

Dataset Description and Wrangling:

The HAM10000 dataset is from the Medical University of Vienna (available for download at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T). The dataset contains 10015 photos of cancer-like skin conditions and a .csv metadata containing relevant image and patient information.

Refer to Table 1 for a summary of the skin conditions contained in the dataset, Figure 1 for an example image to be analyzed, and Figure 2 for the first ten rows in the metadata.

Table 1: A description of the diagnoses included in the HAM1000 dataset

Abbreviation	Diagnosis	Non/Cancerous	Description	Counts
akiec	Actinic keratoses and intraepithelial carcinoma	Noncancerous	A pre-cancerous area of dry skin caused by UV damage.	327
bcc	Basal cell carcinoma	Cancerous	Manifests as a bump and is usually caused by prolonged exposure to UV rays	514
bkl	Benign keratosis-like lesions	Noncancerous	A benign skin condition on the upper extremity that can be patchy or dry in appearance.	1099
df	Dermatofibroma	Noncancerous	Small modules that form in the upper extremities.	115
mel	Melanoma	Cancerous	Cancerous proliferation of melanocytes. The most common type of skin cancer and manifests as a darkened patch.	1113

nv	Melanocytic nevi	Noncancerous	Noncancerous proliferation of melanocytes. Commonly known as a “mole”.	6705
vasc	Vascular lesions	Noncancerous	Occurs after injury to the skin. Commonly known as a “bruise”.	142

Figure 1. Example of an image to be used for model training.

Figure 2. The first ten rows of the tabular metadata.

	lesion_id	image_id	dx	dx_type	age	sex	localization
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear
5	HAM_0001466	ISIC_0027850	bkl	histo	75.0	male	ear
6	HAM_0002761	ISIC_0029176	bkl	histo	60.0	male	face
7	HAM_0002761	ISIC_0029068	bkl	histo	60.0	male	face
8	HAM_0005132	ISIC_0025837	bkl	histo	70.0	female	back
9	HAM_0005132	ISIC_0025209	bkl	histo	70.0	female	back

The metadata contains 10015 rows, corresponding with the 10015 images in the dataset. The column names are as follows:

- lesion_id - Lesion ID (identification code for the skin condition, may or may not be unique for every picture as multiple pictures may be taken for the same condition
- image_id - Image ID (identification code for the image itself, is unique for every picture
- dx - Diagnosis (as seen in the abbreviation column of Table 1)
- dx_type - How the Diagnosis was made (histology, confocal viewing, at a follow-up, or by consensus)
- age - Age of the patient
- sex - Sex of the patient
- localization - Where the skin condition was found

All columns in the metadata were categorical with the exception of age, which is numerical. Our final model uses the diagnosis column as the target variable and all other columns aside from image and lesion ids, in addition to extracted image features, as predictors.

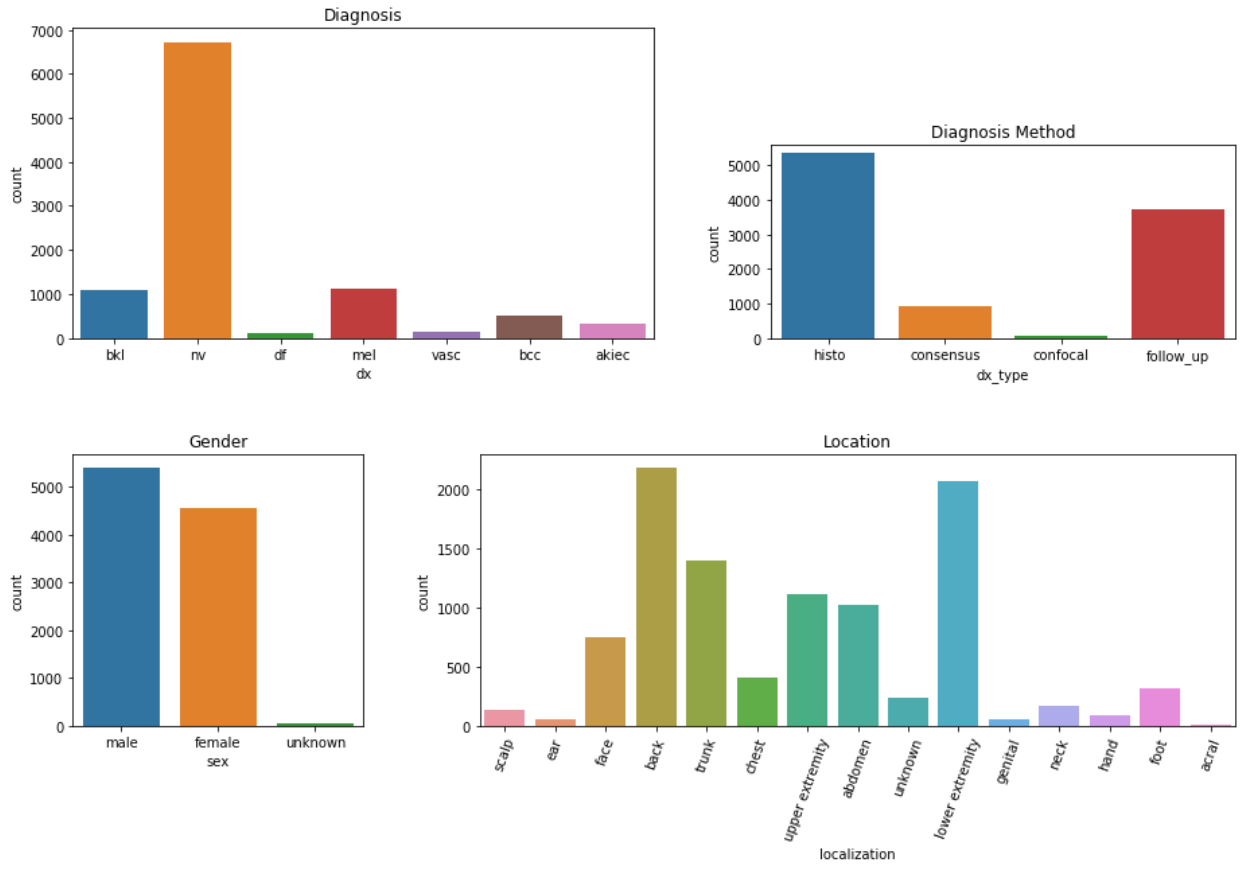
At initial look, the only column which contained any null values was the “age” column with 57 nulls. However, there is one string in the gender and location columns, “unknown”, that refers to values not known or missing.

I imputed null numerical and categorical variables using Random Forest Regressor and Random Forest Classifier, respectively. I chose Random Forest because the algorithm has a high accuracy for large datasets (unlike DecisionTree or kNN), captures relationships between variables (unlike Naive Bayes), and doesn't require a lot of parameter tuning (unlike SVM).

Exploratory Data Analysis:

Figure 3 is a distribution of unique categories for the categorical variables (Diagnosis, Diagnosis Method, Gender, Location) in the HAM10000 metadata. I excluded the image id and lesion id columns because these columns have many values, with most of them unique for every entry, making them both meaningless and tedious to plot.

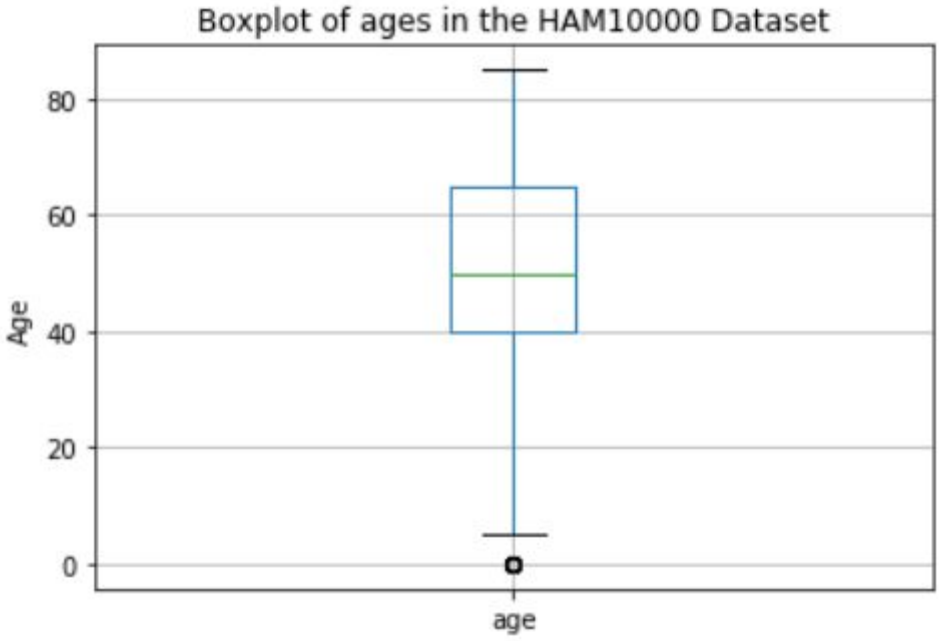
Figure 3. Distribution of data in categorical variables.



One notable observation to be made about the diagnosis column is that the data is very imbalanced, with one category (“nv”) occupying more than 6000 of the 10015 entries. This could have effects on our model testing later on.

In figure 4, we have a box-and-whisker plot that displays the spread of patient age.

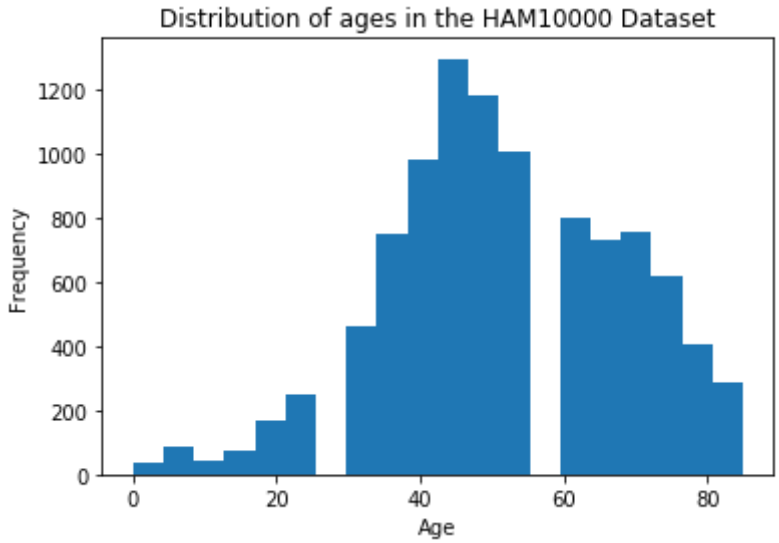
Figure 4. Boxplot of the age column of the HAM10000 metadata.



Patients’ ages range from 0 to 85, with a mean of 52 and a standard deviation of 17 years. There is one outlier at 0, referring to newborn patients, that seems suspicious and warrants further examination. Upon examining the 39 entries where the patient age is 0, the diagnoses are either "nv" (nevi, or a mole/birthmark), "vasc" (vascular lesion, or a bruise), or "bkl" (benign keratosis-like lesions), none of which are cancerous. This information makes sense, could be useful, and doesn't need to be dropped.

Figure 5 further illustrates the distribution of the age column. It appears that patient age does not follow a normal distribution, and the data skew to the left with most of the patients being older than 40 years of age.

Figure 5. Distribution of patient ages in the HAM10000 dataset



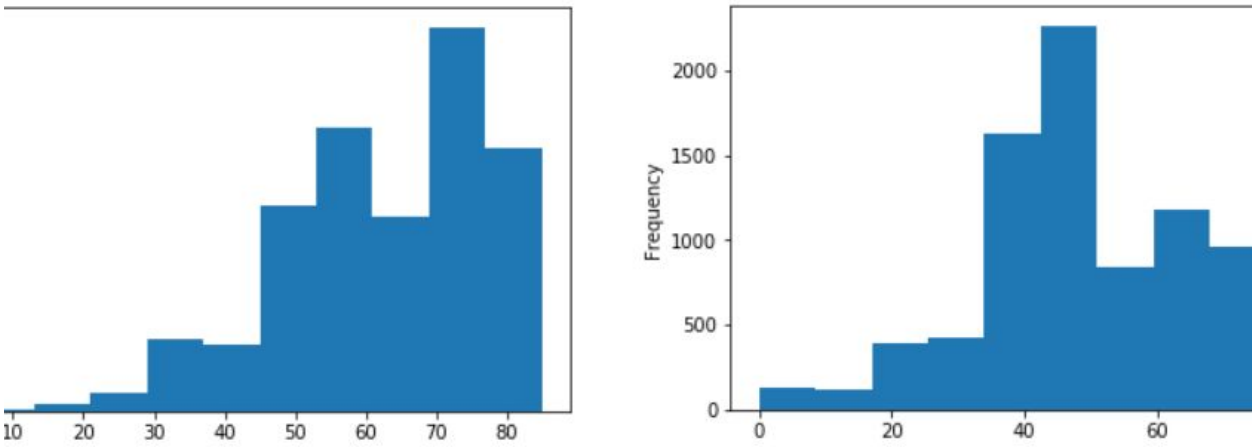
This makes sense, since the risk for all types of cancer, including skin cancer, increases with age and is much higher for people older than 40 years than for those younger. However, I was curious if patients with cancerous diagnoses (Melanoma or Basal cell carcinoma) are actually older than patients with noncancerous diagnoses (Actinic keratoses, Benign keratosis-like lesions, Dermatofibroma, moles, or bruises).

To test this hypothesis, I split the data into cancerous and non-cancerous and compared the distribution of patient age for the two groups.

There are 1637 cancerous cases and 8388 non-cancerous cases. The mean age for patients with cancerous conditions is 62.65 years with a standard deviation of 15 years, and that of patients with non-cancerous conditions is 49.77 years with a standard deviation of 16.53 years.

The distributions can be seen below in Figure 6.

Figure 6. Distribution of ages for cancerous data (left) and noncancerous data (right)

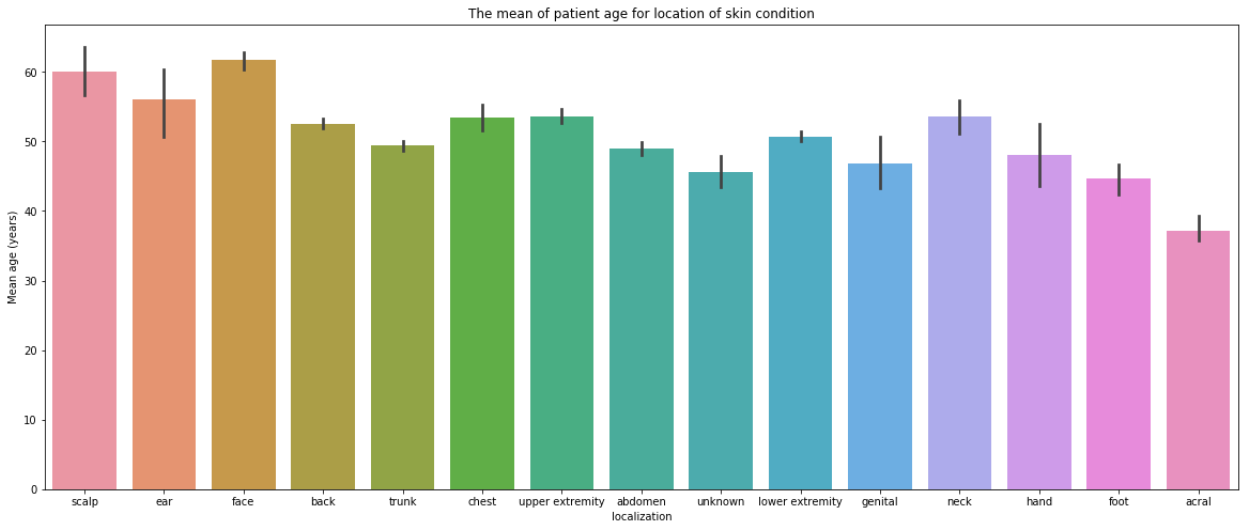


The cancerous data is heavily skewed to the left with the majority of the data concentrated in the higher ages, whereas the non-cancerous data has a more normal distribution and most of the data is centered and symmetrical around the mean.

A formal hypothesis test with a very small p-value further proved that cancerous patients are significantly older than non-cancerous patients.

I also took a look at the mean ages for different locations and genders in figures 7 and 8, respectively.

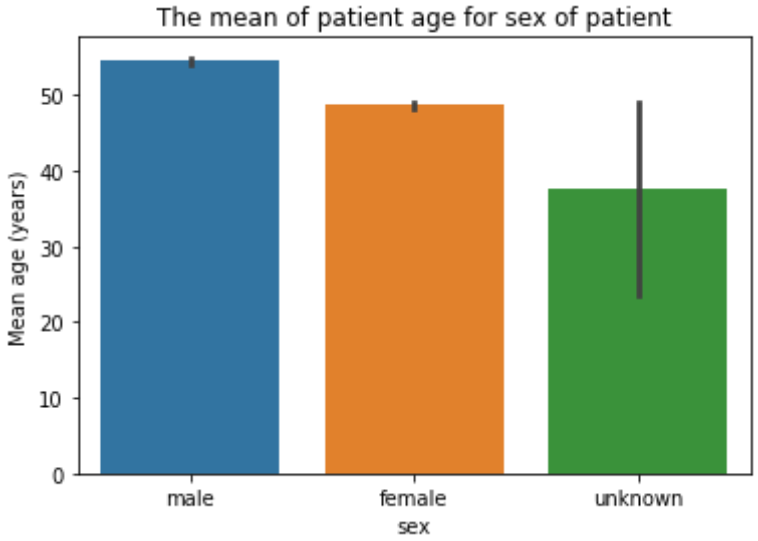
Figure 7. Mean patient ages for different locations of skin condition.



Generally, the mean age for most locations were around 50 years, but patients whose condition was on the scalp and face tended to be older (around 60 years old)

while those with conditions in the foot, unknown, or acral regions tended to be younger (around 45 years old).

Figure 8. Mean patient Ages by gender



It appears that male patients are generally older than female ones, and that those with their sex unknown has the greatest variation.

Model Fitting:

The model fitting process had three parts to it:

1) Use of VGG16 as a feature extractor.
2) Dimensionality reduction using PCA.
3) Fitting classifier models over several combinations of the data.

Feature Extraction:

The feature extraction process was partially based on the method described here: https://medium.com/@franky07724_57962/using-keras-pre-trained-models-for-feature-extraction-in-image-clustering-a142c6cdf5b1.

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.

The input layer accepts a 224 x 224 RGB image, and the output layer is a softmax prediction on 1000 classes on ImageNet. By removing the output layer, VGG can be used as a feature extractor.

I applied this feature extractor over every image in half of the dataset (our training data), converted each image into a numpy array of numerical features, and appended the results into a single pandas dataframe with the identification code of each image (image_id).

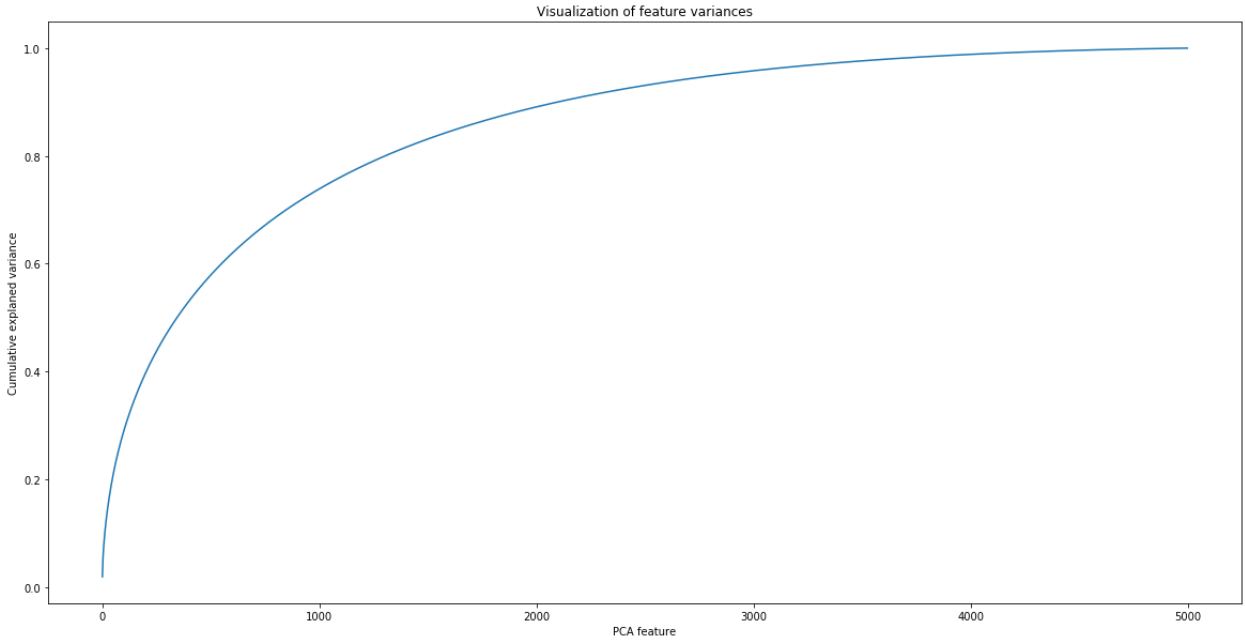
Dimension reduction

Before proceeding with fitting models over the data, I used PCA (Principal Component Analysis) to reduce the number of features in the image features dataframe.

I chose to perform dimension reduction because the image features dataframe contains 25088 features, not including image_id and the diagnosis. This large number of features can induce noise when during the model-fitting process, and it's also computationally expensive.

Before applying PCA, I standardized the image features so that scale would not bias the final feature selection. Then, I selected the number of principal components needed to capture 95% of the variance. In figure 9, the cumulative distribution of percentage of variance captured vs. number of features is shown, Out of 25088 features, only 2831 features were needed to capture 95% of the variance. By shortening the number of features, I significantly reduced the size of the dataframe.

Figure 9. The explained variance of PCA features



When I fit the classifiers over the data, I'll compare the prediction efficacy between the PCA features and the unreduced image features to see if the dimension reduction was effective.

Fitting classifier models

In this project, I used the following models:

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

To see whether the metadata, unreduced image feature data, PCA feature data, or a combination of image and metadata yields the best results, I fitted each of the aforementioned models with the best-performing parameters on each dataset and compared their f1-scores in Table 2.

I chose f1-score as the measure of efficacy because 1) the data is imbalanced, so accuracy is not a robust measurement 2) it takes into account both precision and recall, providing a more effective measurement of the classifier's accuracy.

For parameter optimization, I used 70% of the training data to train the parameters and the remaining to test. The parameters were trained using GridSearchCV, and f1_score was the scoring method used.

Table 2. F1-scores of models on datasets

	Models				
Dataset	SVM	Naive Bayes	DecisionTree	RandomForest Classifier	Average
Metadata	0.64	0.52	0.73	0.71	0.65
PCA Features	0.65	0.49	0.64	0.70	0.62
Unreduced Image Features	0.75	0.68	0.67	0.69	0.70
Combination of PCA and Metadata	0.68	0.62	0.70	0.69	0.67
Average	0.68	0.58	0.69	0.70	

Judging from table 2, it appears that the unreduced image features performed the best, followed by the combination and metadata, with the PCA features alone performing the worst. It appears that by decorrelating and reducing the number of dimensions, important aspects of the images were lost in the PCA process.

There is a notable difference in results from the different models, with RandomForestClassifier having the highest average f1 score across all four datasets, and SVM producing the model with the highest f1-score with unreduced image features.

Naive Bayes performed significantly poorer on all datasets except unreduced image features, suggesting that there were interactions between columns in the metadata andPCA features that were important for the final classification that the algorithm did not take into account.

Both DecisionTree and RandomForestClassifier performed better on datasets containing categorical variables. However, the difference is more pronounced in

DecisionTree versus in RandomForestClassifier. This is perhaps because RandomForest is an ensemble method featuring a collection of shortened DecisionTrees, so it's less likely to overfit.

To make predictions on a set of unseen image, I'm going to use the metadata with RandomForestClassifier. Even though LinearSVC produced the model with the highest f1-score, it is more prone to overfitting than RandomForest due to the latter's ensemble nature. Thus, the model might not replicate its success in unseen data.

Also, a lot of hyperparameter tuning took place for LinearSVC, which took a lot of time, and RandomForestClassifier did not have this issue.

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 0.49 (Naive Bayes with PCA features) to 0.75 (LinearSVC with unreduced image data), and the resulting best model is RandomForestClassifier (with f1-scores from 0.69 to 0.71), because it performed the best across multiple datasets, is less prone to overfitting, and doesn't require a lot of hyperparameter tuning.

In this project, I only used simple classification models, and a possible next step would be to develop a custom model that's specifically designed to work with this particular dataset, perhaps utilizing deep learning. Another option to explore would be modifying the images so that different lighting, contrast, or size would not bias the results.

Our model can perhaps be implemented in an application of sorts for a non-health professional individual who suspects that he/she may have skin cancer. Ideally, the individual can upload a picture of their skin condition, and the system will return the most likely diagnosis among those included in the dataset. Using this information, the individual can make a more informed decision on whether or not to pursue further treatment.