Predicting skin symptoms on dogs

Project by: Andrea Ternera

Introduction:

According to a survey, done by marketwatch.com, of more than 1,000 dog owners, most spent an average of $40 to $290 per month on their dogs — or an average of $480 to $3,470 annually. But this number will vary depending on a dog's individual needs, size, pet insurance plan's costs, and health status. Pet insurance premiums can range between $360-$720 annually. Sick and emergency veterinary visits can cost an owner between $150-$1,200 per visit. Looking deeper, according to Healthy Paws Insurance 2019 report, potential treatment costs for skin conditions can go up to $4,100.

*https://www.marketwatch.com/guides/insurance-services/cost-of-owning-a-dog ** https://www.thisoldhouse.com/home-finances/reviews/cost-of-owning-a-dog

Recap of problem:

Do environmental factors/living conditions increase the likelihood of a dog developing skin conditions? Are certain breeds (because of their size and other characteristics) more prone to these conditions including canine atopic dermatitis? If so, can we determine if a specific breed or specific dog regardless of breed, given specific environmental exposure, will present skin issues? Can we identify and communicate the risks and preventative living conditions for owners to help their dogs and prevent emergency vet visits and treatment cost?

Obtaining the Data and Cleaning:

In order to complete this project, the following datasets were downloaded to local drive (from):

- Data_Environmentandskinsymptoms.xlsx (figshare.com)
  - 8644 rows/ 26 columns containing medical records of dogs that include skin symptoms, breed, age, gender, vaccination, and other environmental aspects of living area (other animals, other dogs, born in owner family, etc.)
- Dogs Intelligence and Size (kaggle.com)
  - 136 rows/ 5 columns containing specific information for each breed intelligence and classification.
- Best in Show (kaggle.com)
  - 150 rows/5 columns containing basic information for each dog breed such as high AND low height in inches, and high AND low weigh in lbs.

Translating:

Due to Breeds in the Data_Environmentandskinsymptoms being in Finnish, we translated the breeds to English and merged that translated file.

Missing values:

Columns with over 50% missing values in original EnvironementSkin df.

- Dam_vaccinated_prebirth (72% missing values), 1241 records were a Yes, and 1160 were a No. *Given the almost 50-50 diff in values we filled all NaN values in this column with 50-50 0 and 1.*
- Skin symptoms_dam (66% missing values), 2823 records were a No, and 121 a yes.
- Smoking_previously (65% missing values), 2760 record only outside, 192 rarely inside, and 93 mainly inside.
- Dam_dewormed_prebirth (53% missing values), 3903 record Yes and 158 No.

Dropped rows:
Where all values of the 4 columns above were NaN

- only 569 records contain ALL information for these 4 columns.

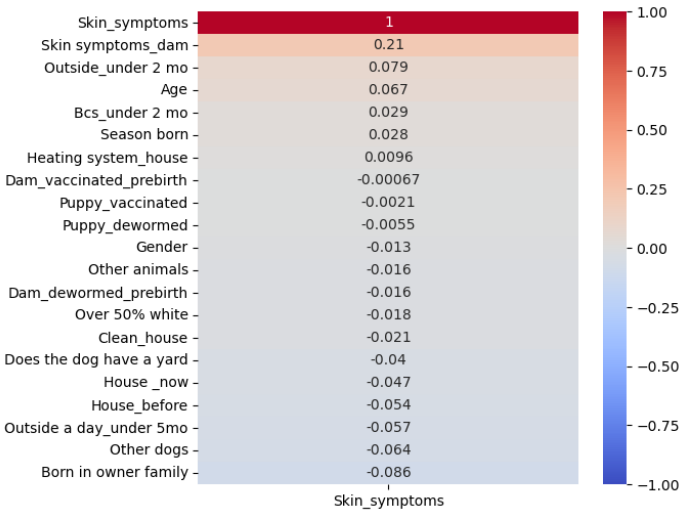
Dropped columns:

- Id and Fci dropped due to being unnecessary for this project.
- Smoking_previously due to:
  - high number of NaN values
  - high discrepance between possible values

Cleaned data contained: 5265 rows / 33 columns.

Exploratory Data Analysis:

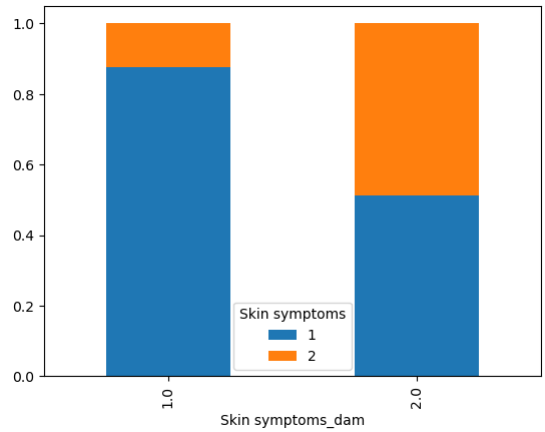
Total rating and all features



Using .groupby and .crosstab. plot we were able to find the following in the features provided:

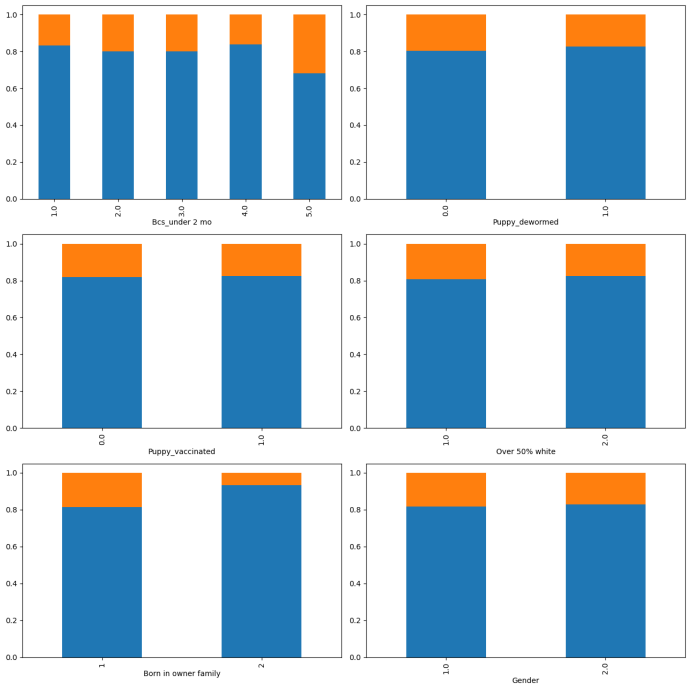
Skin symptoms Dam & skin symptoms

- When Dam presents skin symptoms (right bar), the puppy's likelihood to also have skin symptom almost triples from 22% to 65%



Puppy characteristics & skin symptoms

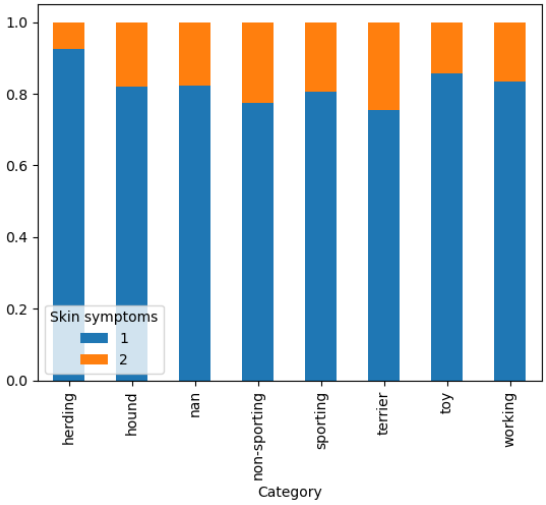
- Puppy being dewormed or vaccinated does not seem to show any relation to them representing skin symptoms.
- Likewise, being over 50% white or the puppy's gender also do not seem to be related.
- Checking on the two features that show some variability we find that:
    - Skin condition based on the puppy's BCS (body condition score) remains steady within the first 4 but has a 25.7% increase when the puppy has a 5 BCS which refers to "Very Slim".
    - Skin condition based on being born in the owner's family show that when the pup is not born in owner's family, they show an 18.7% more cases of skin condition than if they are born in the family.



Outside environment and house environment showed very little correlation with skin condition.

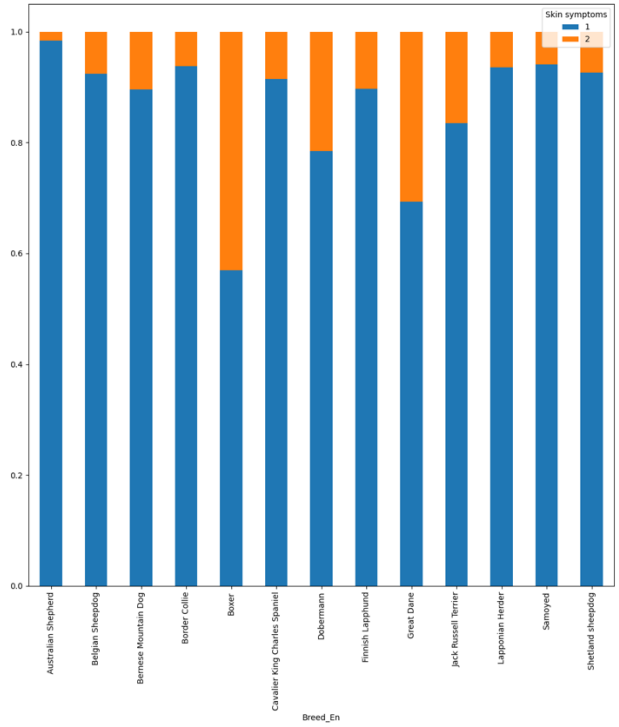
Category & skin symptoms

- Herding has the least skin conditions at only 14% recorded.
- Toy and Working stay right below 29% followed by Hound.
- Terrier and non-sporting have the highest cases at over 35% each.
    - Terrier with the highest from all categories at 39.3%



Breed & skin symptoms

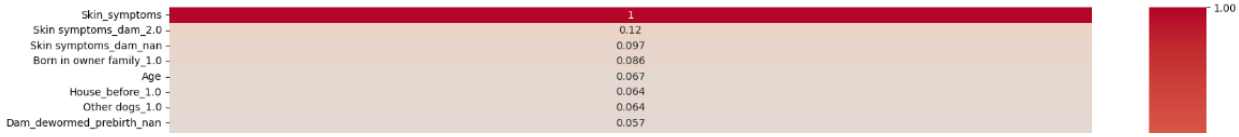
- German Shepperd, Staffordshire Bull Terrier, Boxer, Doberman, Great Dane, Chinese Crested, French Bulldog, and Parson Russellian Terrier all present over 20-25% cases of skin symptoms.
- The Boxer has over 40% of records with skin symptoms.



Model selection:

Feature engineering

- Given that some features are categorical (in this case most features) we transformed these with get_dummies. We also made the decision to maintain NaN values as its own count when hot encoding to avoid overfitting or filling null values with the mode and affecting the overall impact of feature on model.
  - Once the hot encoding was completed, we had a set of 5265 rows and 78 columns.
- We checked the positive correlation of features with skin symptoms again to see if there was any change.



  - As the one done with original data, skin symptoms dam is present as first and second positive feature followed by the puppy not being born in owner's family (this could be due to the increased medical risks of puppies separated from mom too early). Age is also a positive feature observed in original and one hot encoded set.

Train/test split

- The data was split two different ways to ideally further verify which split was best for modeling.
  - The first split was a simple train_test_split with test size .25 and random_state= 2.
  - The second split has the same 2 parameters but with stratify=y added.
    - Given that our classes are disproportionate with more "no skin symptom" cases than positive ones, stratify tries to preserve as much as possible the proportions among that class. *The Age feature also presented 54 null values which we decided to fill with the mean for all splits. *

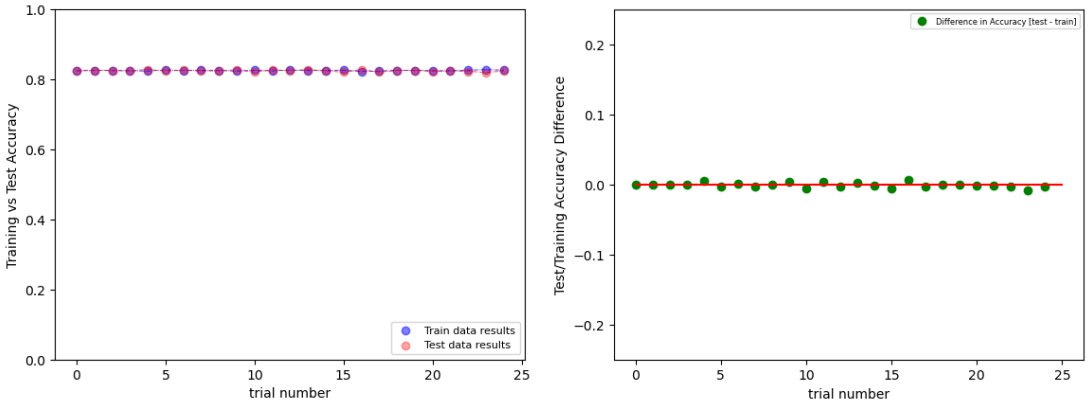
Models

We looked at 3 models:

1. Logistic regression
2. Random forest classifier
3. KNeighborsClassifier.

All 3 models were run with both splits to identify best case/parameter.

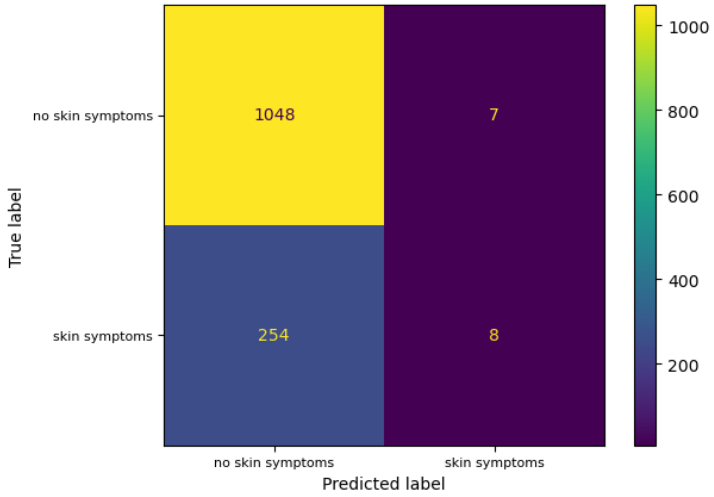
We also evaluated the test size of the split to see if this was a significant indicator of performance.



There was no evidence that accuracy had any significant variation when changing the test size. Therefore, we kept .25 as the test size for split.

1. ***Logistic Regression:***

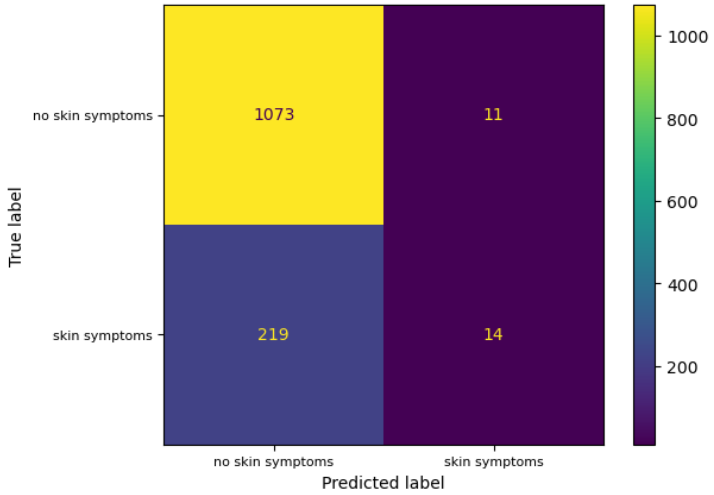
- No stratify.



	precision	recall	f1-score	support
1	0.80	0.99	0.89	1055
2	0.53	0.03	0.06	262
accuracy			0.80	1317
macro avg	0.67	0.51	0.47	1317
weighted avg	0.75	0.80	0.72	1317

- From the confusion matrix we can see that while the overall accuracy was 80%, when we predict skin condition, 47% (7 of 15) of the time we are predicting a false positive, while the false negatives (predicting no skin symptoms when in fact there is skin symptom) is about 20% (254 of 1302).

- With stratify.

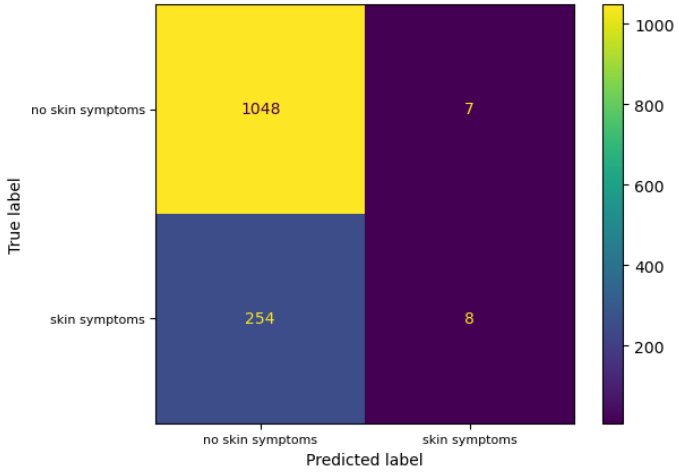


	precision	recall	f1-score	support
1	0.83	0.99	0.90	1084
2	0.56	0.06	0.11	233
accuracy			0.83	1317
macro avg	0.70	0.52	0.51	1317
weighted avg	0.78	0.83	0.76	1317

  - From the confusion matrix we can see that this model is indeed better than using the split data with no stratify.
  - The overall accuracy was 83%
  - When we predict skin condition 44% (11 of 25) of the time we are predicting a false positive which is less than the above 47%.
  - False negatives are about 17% (219 of 1292) which again, is less than the above 20%.

2. ***Random Forest Classifier***

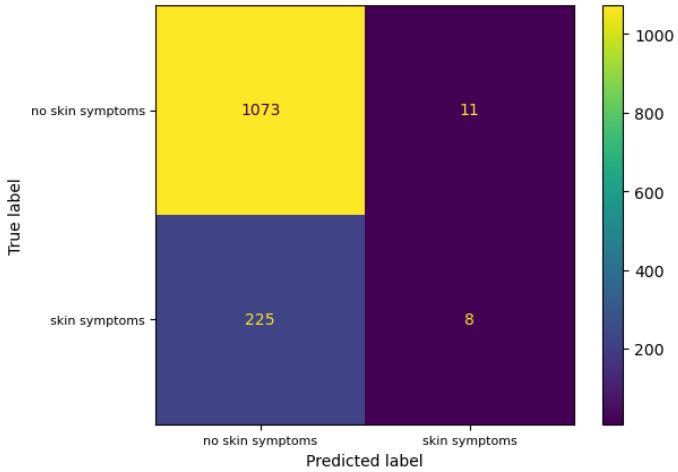
- No stratify.



	precision	recall	f1-score	support
1	0.80	0.99	0.89	1055
2	0.53	0.03	0.06	262
accuracy			0.80	1317
macro avg	0.67	0.51	0.47	1317
weighted avg	0.75	0.80	0.72	1317

  - We see the same limitation as using the logistic regression, the models are producing very high percentage of false positives.

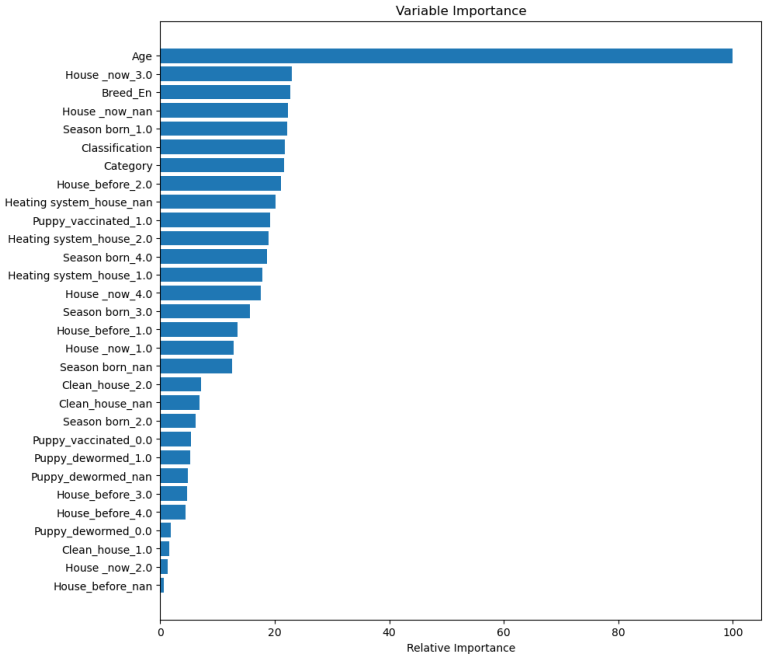
- With stratify.

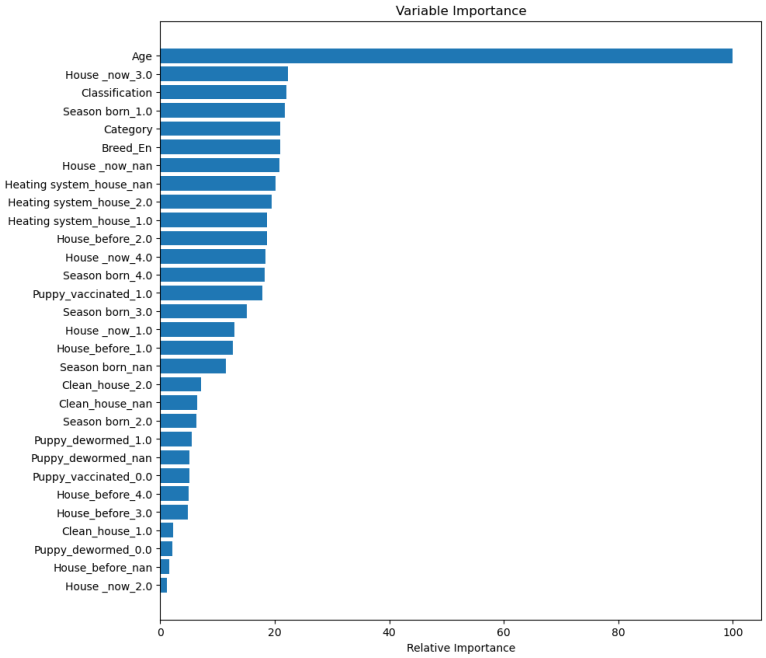


	precision	recall	f1-score	support
1	0.83	0.99	0.90	1084
2	0.42	0.03	0.06	233
accuracy			0.82	1317
macro avg	0.62	0.51	0.48	1317
weighted avg	0.75	0.82	0.75	1317

  - With stratify the precision of no skin symptoms worsens by 11%
  - Overall accuracy improves by 2%
- Although accuracy in between both models is very similar; the precision, recall, and f1-score of specifically the model predicting NO skin symptoms correctly is as good or 1% better in this model than the logistic regression model. However, the avg scores and correct positive skin condition scores for Random Forest classifier are lower than those seen with Logistic regression.

Under this model, the following shows the variable importance with the first image referring to the split with no stratify and the second/last referring to the one with stratify.

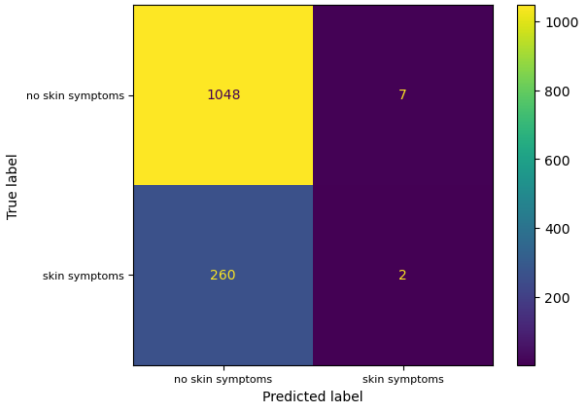


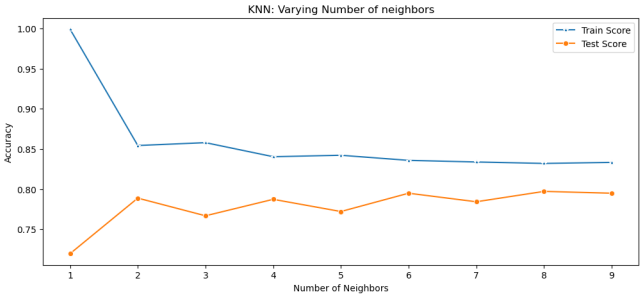


- Age and house now 3 (detached wood) are both the top features of importance for both models, however, looking at #3 we have Breed for plain split and Classification for the stratified split. Also looking into these two variables, they are each represented in 6th place on their opposite relative importance chart.
- Overall, the same top 7-8 variables seem to be the same in both cases just slightly different order for a few.

3. ***KNeighbors Classifier***

- No stratify

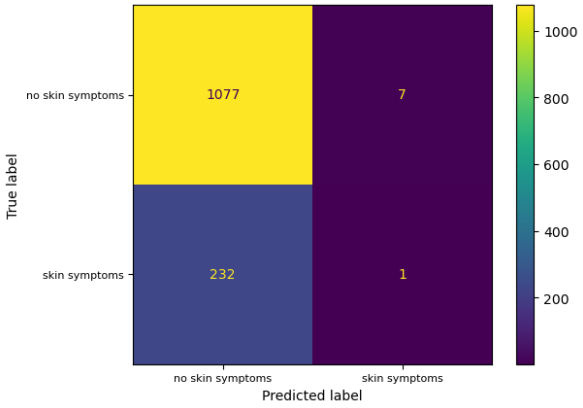


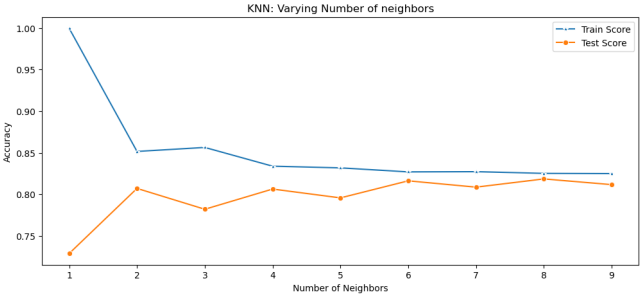


	precision	recall	f1-score	support
1	0.80	0.99	0.89	1055
2	0.22	0.01	0.01	262
accuracy			0.80	1317
macro avg	0.51	0.50	0.45	1317
weighted avg	0.69	0.80	0.71	1317

  - The percentage of false positives is increasingly high with this model. Scores as can be seen are extremely low for the model's ability to correctly predict a positive skin symptom on a puppy.

- With stratify





	precision	recall	f1-score	support
1	0.82	0.99	0.90	1084
2	0.22	0.01	0.02	233
accuracy			0.82	1317
macro avg	0.52	0.50	0.46	1317
weighted avg	0.72	0.82	0.74	1317

  - This is by far the worst performing model. Although in both cases the optimal number of neighbors for test data is 8 and in this stratify split, the accuracy of both is extremely close to one another:
    - We have the highest false positive of 87.5%
    - Similar false negative at 17.8%

Conclusion:

- Each model showed better results when using the split that was stratified.
- All 6 (2 of each) had issues and low performance scores when correctly predicting skin symptoms.
    - Likewise, all of them performed significantly well when correctly predicting no skin symptoms.

This information can be discussed to decide which is more important, reducing false positives or false negatives, assuming overall accuracy is acceptable.

TOP MODEL:

- Based on weighted average scores is:
    - Logistic regression with the stratified split.
        - Precision: 78%
        - Recall: 83%
        - f1-score: 76%
        - False negatives: 17%
        - False positives: 44%

WORST MODEL:

- Based on weighted average scores is:
    - KNeighbors Classifier (with or without stratified split).
        - Precision: 72%
        - Recall: 82%
        - f1-score: 74%
        - False negatives: 17.8%
        - False positives: 87.5%

It is important to notice the following:

- While the overall accuracy of the models ranged between 80- 83%, this measure should not be taken into consideration alone. As shown above, the other scores such as precision, recall, and f1-score in every case are being "improved" in the overall average thanks to the model being able to have such a high performance when predicting no skin symptoms correctly.
- Given the class imbalance in our problem, we might need to gather more information, ideally of such cases where the pup does present skin symptoms or develop more models with less features to identify if some of the information is conflicting for the model to correctly label a positive skin symptom.