Classification for Skin Cancer Diagnostic

Stats 101C Final Project

Manuela Lozano, Gwen Sanders, Christopher Kusmana, Mizuki Shitomi, Shaina Dulles, Naman Satija

Abstract

This project investigates the use of statistical methods to classify skin cancer as benign or malignant using a large tabular dataset containing demographic, behavioral, environmental, and lesion-related predictors. The training data consist of 50,000 observations with 49 predictors and an approximately balanced response variable. A key challenge of the dataset is widespread missingness across predictors, affecting roughly 8% of values, which motivated the use of median imputation for numerical variables and mode imputation for categorical variables. Exploratory analysis identified age, ultraviolet exposure, lesion characteristics, family history, skin tone, and immunosuppression as the most informative predictors. Several classification models were evaluated, including ridge-penalized logistic regression, Linear Discriminant Analysis (LDA), and nonlinear alternatives. Among these, LDA achieved the strongest performance with a Kaggle test accuracy of approximately 0.6048, suggestireceng that the relationship between predictors and cancer status is largely linear. Overall, the results demonstrate that classical, interpretable statistical models can extract meaningful signal from structured skin cancer data, while highlighting the limitations of structured, non-imaging predictors.

1 Introduction

1.1 Context/background

Skin cancer is one of the most common cancers globally, and its incidence continues to rise due to factors such as ultraviolet (UV) exposure, genetic susceptibility, and behaviors like tanning and inadequate sun protection. These determinants are widely documented in dermatology research, which emphasizes that pigmentation traits, family history, and immune status strongly influence melanoma risk, while UV exposure remains one of the most significant environmental drives of malignancy.

Early diagnosis is crucial in improving patient outcomes. When melanoma is detected at an early stage, survival rates are significantly higher. However, distinguishing benign from malignant lesions can be challenging without specialized expertise or advanced imaging. Recent research in dermatological imaging demonstrates that characteristics of lesions, including asymmetry, order irregularity, color variation, and size, provide informative signal for identifying cancerous growths. These clinical findings highlight the potential value of predictive modeling approaches that can help support early assessment when traditional diagnostic resources are less applicable.

1.2 Motivation

The motivation for this project is rooted in both clinical relevance and analytical opportunity. Understanding which factors most strongly predict skin cancer can guide public health and support skin stratification, while also serving as a valuable case study for applying statistical learning methods to a medically meaningful problem. The goal of this project is not clinical diagnosis, but rather statistical prediction and variable importance analysis using structured data.

1.3 Literature Review

The literature highlights the central role of pigmentation traits, ultraviolet exposure, sunscreen habits, and lesion features in determining skin cancer risk (Armstrong & Kricker, 2001; Gandini et al., 2005). While these relationships are well established clinically, fewer studies focus on comparing their predictive contributions using modern statistical learning techniques (Hastie et al., 2009).

Ultimately, this project aims to connect established dermatological research with predictive modeling methods to better understand the key determinants of skin cancer and explore how statistical tools may contribute to earlier and more informed detection.

2 Data Analysis

2.1 Data Structure

The training dataset consists of **50,000 observations** and **50 columns** or **49 predictors** excluding the response variable and the test dataset contains **20,000 observations** and also the same **49 predictors**. The predictors comprise a mix of **20 numerical** and **29 categorical** variables, capturing demographic, behavioral, environmental, and dermatological factors relevant to skin cancer risk.

The numerical variables represent continuous or ordinal measurements such as *age*, *BMI*, *lesion_size_mm*, and *avg_daily_uv* many of which are biologically or environmentally interpretable risk factors. The categorical variables encode discrete attributes, including demographic descriptors (e.g. *gender*), preventive behaviors (e.g. *sunscreen_freq*), medical history (e.g. *family_history*), and lesion characteristics (e.g. *lesion_color*).

The target variable, **Cancer**, is approximately balanced, with **47.7% Benign** and **52.3% Malignant** cases. This near balance implies that accuracy remains informative. All features are stored in appropriate numeric or categorical formats, requiring only standard preprocessing like missing value imputation and categorical variable encoding prior to model development. The large sample size and mixed variable types make this dataset well-suited for comparing classical linear classifiers with more flexible machine learning approaches.

2.2 Missing Data Detection and Cleaning

To better understand the structure of missing data across predictors, we computed the proportion of missing values for every feature in both the training and test sets. Figure 1 displays the per-feature missingness percentages using a horizontal point–line visualization.

Observe that nearly all predictors exhibit approximately **8% missingness**, and this structure is consistent across both datasets. This indicates that the missingness mechanism is primarily *row-wise* rather than feature-specific, with many observations containing missing values in multiple columns simultaneously. As a consequence, omitting all observations with a missing value will shrink the sample from **50,000 training observations to only 334 complete cases**, a loss exceeding **99%**. Such a reduction would severely compromise model stability and generalizability.

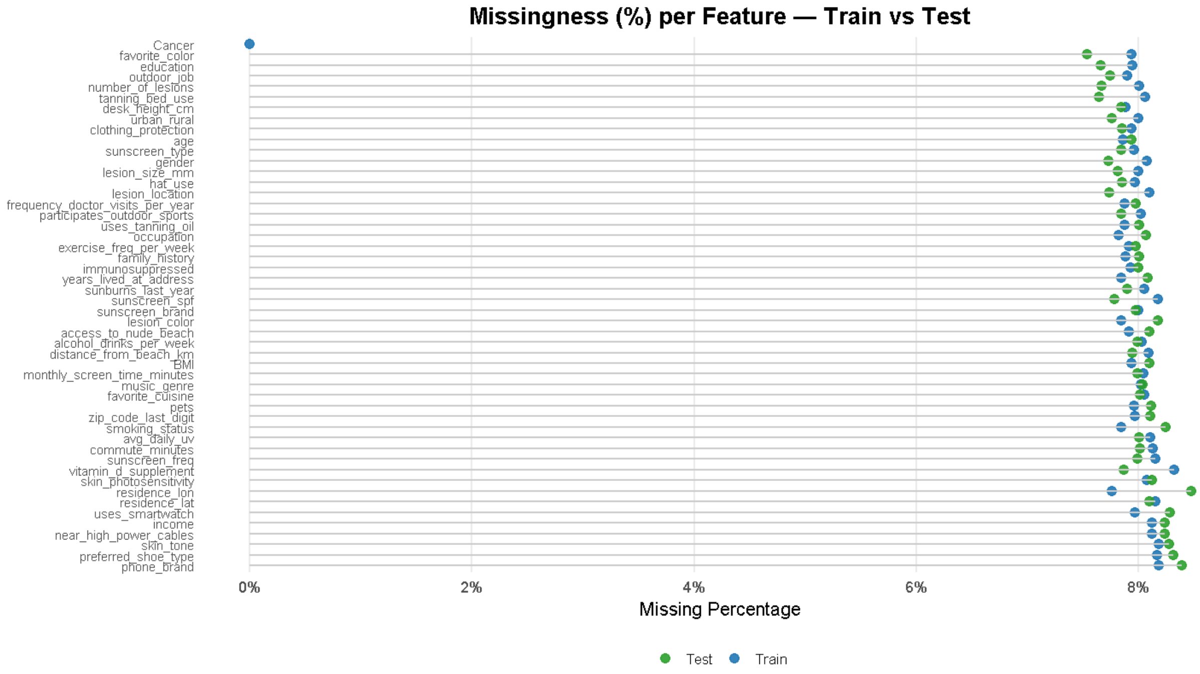

Figure 1: Per-feature missingness percentages in the training and test datasets. Approximately 8% of values are missing across most predictors, resulting in severe row-wise data loss under listwise deletion.

These findings informed our decision to implement both a simple median–mode baseline and a more flexible random-forest-based imputation using `missRanger`.

Through multiple modeling and submissions, it is observed that median and mode imputation outperformed more complex methods such as missRanger, likely due to the large sample size and weak correlation structure among predictors, which limits the benefits of modeling conditional dependencies. Although simple imputation may introduce bias, it substantially reduces variance by preserving sample size and led to improved out-of-sample performance. Consequently, all models discussed in the remainder of this paper are evaluated using median imputation for numerical variables and mode imputation for categorical variables.

2.3 Exploratory Data Analysis

This section summarizes distributional patterns and structural relationships among predictors. The goal is to gauge variable influence before moving on to a more robust selection process, assess assumptions relevant to linear models, and highlight nonlinearities or interactions that help our model selection. These exploratory findings eventually guide subsequent variable selection and inform the suitability of linear versus nonlinear classification methods.

2.3.1 Numerical Variables

Summary statistics indicate that most variables are reasonably centered and symmetric, with only mild skewness. Class-conditional means reveal meaningful separation between benign and malignant groups. Malignant patients are older on average (**52.5** vs. **48.0**), have higher UV exposure (**3.63** vs. **3.41**), and exhibit larger lesion sizes and counts. Behavioral risk factors show similar trends: malignant cases exhibit slightly higher sunburn frequency and alcohol consumption. These patterns suggest that age, UV exposure, and lesion morphology are among the most predictive numerical features.

Correlation analysis shows extremely weak relationships among numerical predictors, with most pairwise correlations close to zero. This low-correlation structure simplifies interpretation and reduces concerns about multicollinearity when fitting generalized linear models.

Figure 2 displays representative boxplots illustrating these class-dependent differences.

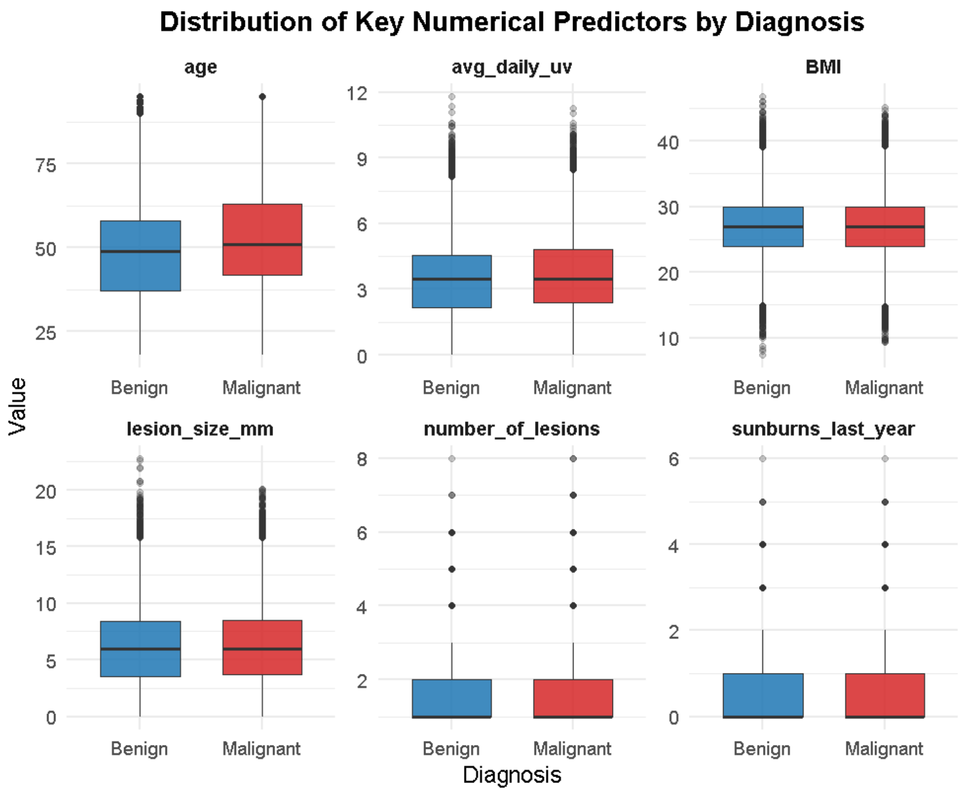

Figure 2: Boxplots of key numerical variables stratified by diagnosis. Age, UV exposure, and lesion-related variables show clear differences between benign and malignant cases.

2.3.2 Categorical Variables

Several categorical features display strong class separation. For example, `clothing_protection` shows substantial variation across classes: the "Medium" protection level is most common in both groups but disproportionately higher among malignant cases (12,795 vs. 11,761). Risk-elevating categories such as `family_history = Yes`, `immunosuppressed = Yes`, and `tanning_bed_use = Yes` also show higher malignant frequencies.

Environmental and behavioral predictors also behaves similarly. For instance, individuals reporting `access_to_nude_beach = Yes` appear slightly more likely to be malignant relative to their benign counterparts. Lesion-based variables such as `lesion_color`, `lesion_location` exhibit substantial within-category differences, making them strong candidates for modeling.

Figure 3 shows representative class-conditional distributions for a subset of highly informative categorical predictors.

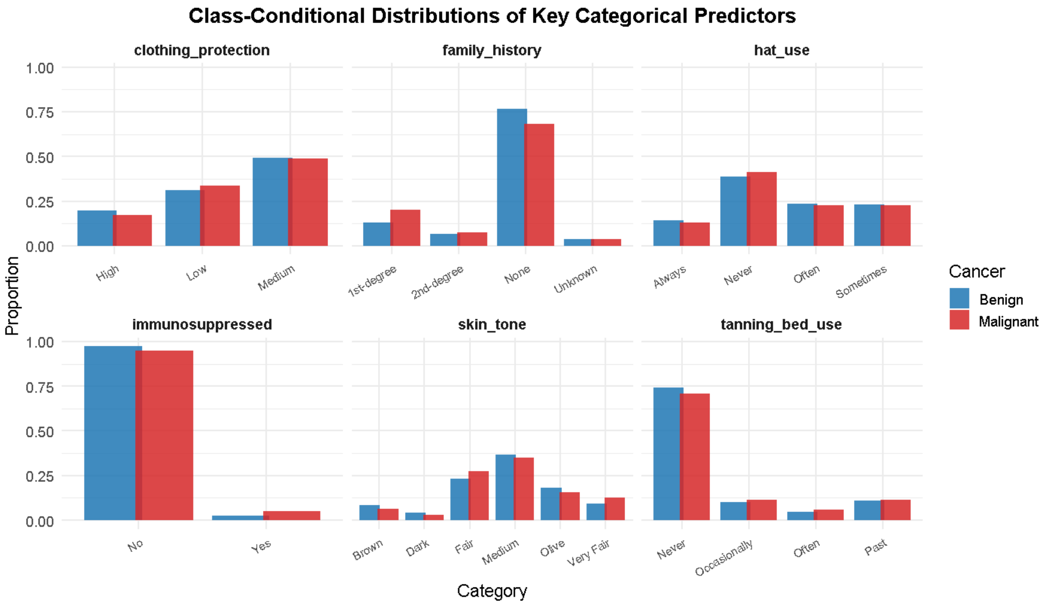

Figure 3: Class-conditional distributions for selected categorical variables. Behavioral and lesion-related features exhibit pronounced differences between benign and malignant groups.

Overall, EDA suggests that meaningful but modest separation exists between benign and malignant cases across both numerical and categorical predictors. Class-conditional shifts in numerical variables such as age, UV exposure, and lesion characteristics appear largely additive, while several categorical variables exhibit consistent differences in class proportions. Additionally, the weak correlation structure among numerical predictors reduces concerns about multicollinearity and supports the stability of linear modeling approaches. Together, these findings motivate the variable selection and modeling strategies pursued in subsequent sections.

3 Variable Selection

Variable selection was performed to improve interpretability, reduce noise, and maintain predictive performance with a smaller subset of features. Using multiple selection methods allows for cross-validation of importance across statistical tests, visual separation, and predictive performance. Hence, this section will go over four methods we used to narrow our variable selection: density plots, stacked bar charts, chi-squared tests for categorical variables, and cross-validation decision trees. As shown in Figure 4, the density plots for variables such as sunscreen SPF and frequency of doctor visits per year showed significant overlap between the probability distributions of Benign and Malignant, occupying largely the same range and peaking near the same value. On the other hand, variables such as age and average daily UV show some separation between the two curves.

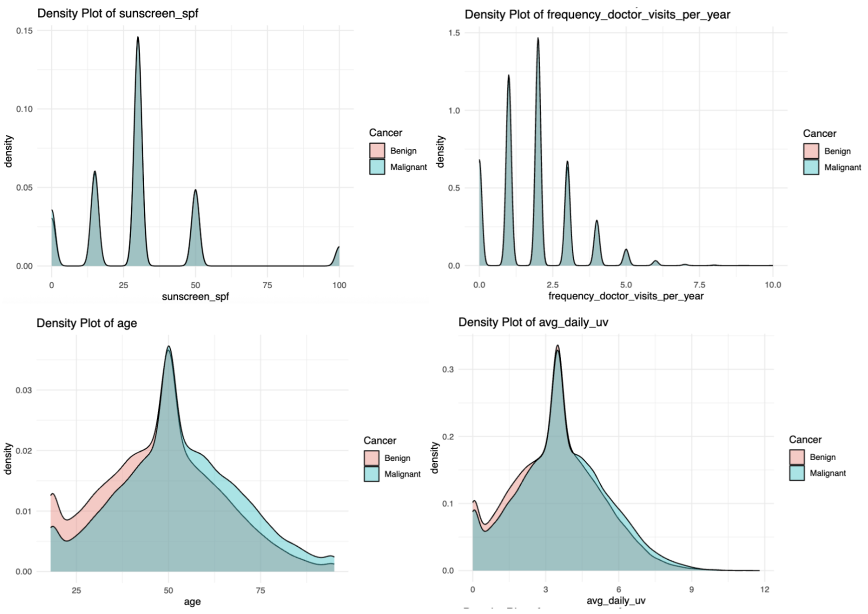

Figure 4: Density Plots

While density plots are useful for predictors with numerical values, to assess the predictive strength of categorical variables, we ran a chi-square test in order to determine which variables are significant predictors. The table below (Table 1) showcases the results of the chi-square test. With a significant threshold of $\alpha = 0.01$, we were able to find 9 categorical variables out of 29 that are statistically significant. These are skin tone, occupation, sunscreen frequency, hat use, clothing protection, tanning bed use, family history, immunosuppressed, and skin photosensitivity. This result is further supported by the stacked bar charts. As reflected in Figure 5, we found that predictors such as education level, living in urban/suburban/rural area, and sunscreen brand do not carry much predictive signal since there are noticeable differences in the proportions of Benign versus Malignant cases across groups. In contrast, the stacked bar charts for predictors such as family history, skin tone, and immunosuppression indicated that these variables exhibit stronger associations with the response.

Variable	**p-value**	**Variable**	**p-value**
gender	0.5672	skin_tone	$< 2.2 \times 10^{-16}$
education	0.4721	urban_rural	0.1112
occupation	$< 2.2 \times 10^{-16}$	sunscreen_freq	$< 2.2 \times 10^{-16}$
sunscreen_type	0.7812	sunscreen_brand	0.3512
hat_use	1.91×10^{-8}	clothing_protection	9.374×10^{-15}
tanning_bed_use	$< 2.2 \times 10^{-16}$	family_history	$< 2.2 \times 10^{-16}$
immunosuppressed	$< 2.2 \times 10^{-16}$	smoking_status	0.4832
access_to_nude_beach	0.9391	near_high_power_cables	0.9995
lesion_color	0.4466	lesion_location	0.4082
pets	0.2827	favorite_color	0.5039
phone_brand	0.3739	music_genre	0.4137
skin_photosensitivity	7.059×10^{-13}	vitamin_d_supplement	0.2836
favorite_cuisine	0.4040	uses_smartwatch	0.2151
participates_outdoor_sports	0.07131	uses_tanning_oil	0.5874
preferred_shoe_type	0.7582		

Table 1: Chi-square test for categorical predictors

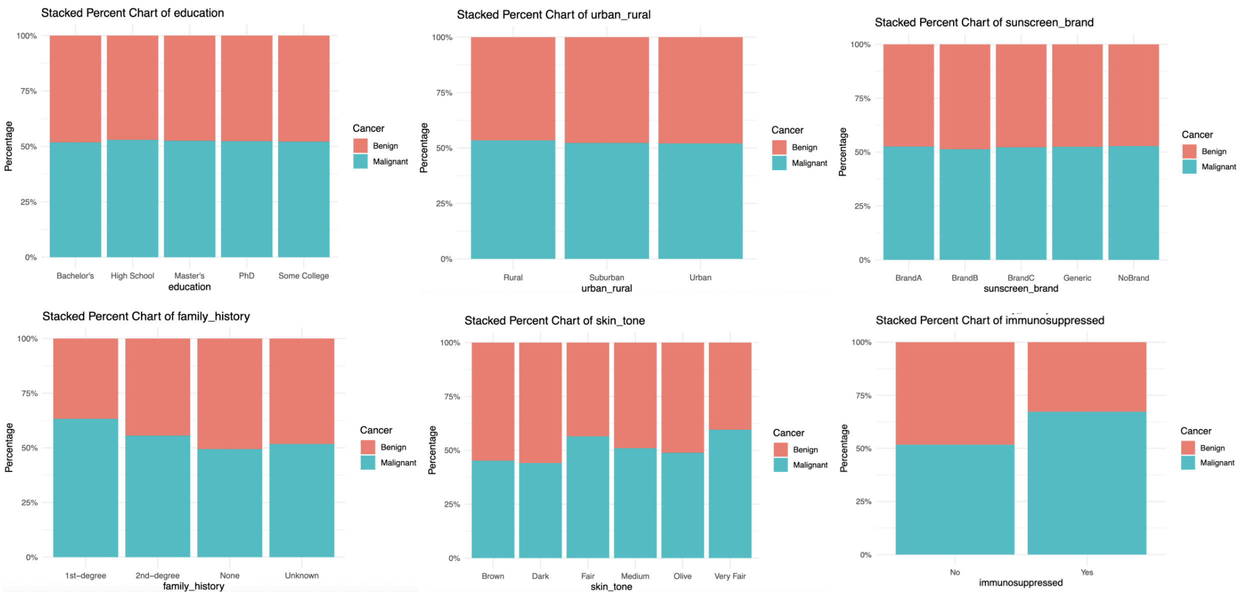

Figure 5: Stacked Bar Charts

Finally, observing the cross-validation decision tree (Figure 6), we were able to identify the most influential variables to be age, family history, skin tone, immunosuppression, sunscreen frequency, average daily UV, and number of lesions, in order of importance. Based on a variable importance analysis, we were able to reduce model complexity from 49 predictors to 7 predictors, which reduced computational cost, ensuring the chosen model is both effective and feasible to deploy. However these benefits comes with a cost of sacrificed some model accuracy. Therefore, the cross-validation decision tree was used primarily to identify influential variables rather than as a final predictive model. The highest-performing model excluded nine predictors which are favorite color, favorite cuisine, phone brand, preferred shoe type, music genre, pets, smartwatch use, desk height, and zip code—due to their minimal relevance to skin cancer prediction.

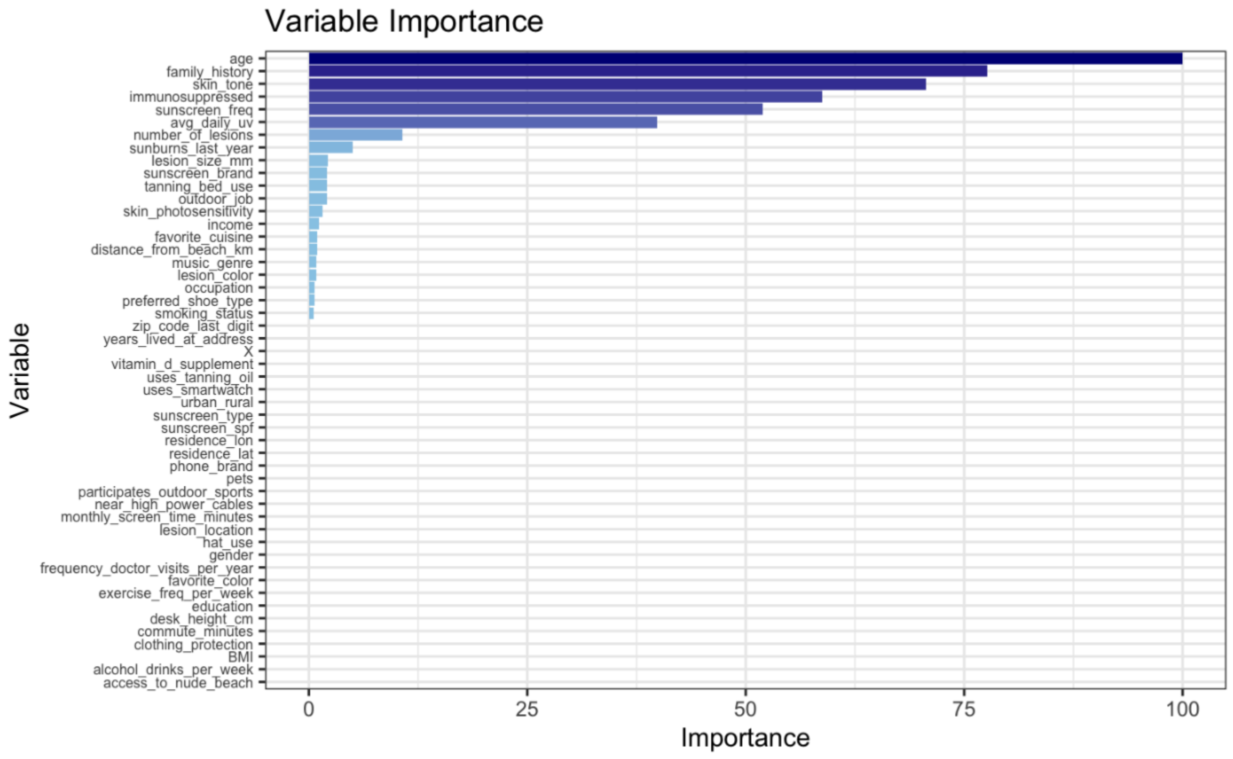

Figure 6: Cross-validation Decision Tree

4 Methods and Models

Our modeling strategy for this project followed a stepwise approach so as to balance interpretability, accuracy, and consistency with learning methods from the course. These modeling choices are informed by the exploratory analysis, variable selection results, and experimental modeling with various statistical models which suggested largely additive effects, weak correlation among predictors, and limited gains from aggressive dimensionality reduction.

Key properties of the data informed our approach. First, the dataset contains a relatively large number of predictors (49), many of which are categorical and expand into multiple dummy variables when encoded. Second, although correlations among numerical predictors were generally weak, the presence of related categorical indicators raises concerns about multicollinearity in unregularized models. Finally, the response variable is approximately balanced between benign and malignant cases, which made classification accuracy an interpretable evaluation metric.

Given these considerations, we began with linear classification models that could offer stability and interpretability, and then moved on to exploring more flexible linear alternatives so as to assess whether additonal complexity could yield meaningful performance gains. This allowed us to benchmark model performance while maintaining clarity on the trade-offs involved.

4.1 Model 1: Ridge Regression

Logistic regression provides a natural starting point for binary classification problems, and can offer a framework for distinguishing between benign and malignant skin lesions. However, in high dimensional settings with many correlated predictors, standard logistic regression can suffer from overfitting and unstable coefficient estimates.

To address these issues, we implemented ridge-penalized logistic regression with all predictors. Ridge regularization shrinks the coefficient estimates towards zero while retaining predictors in the model, thus reducing variance and the effects of multicollinearity. This property is especially well-suited in our dataset, where many predictors may carry weak but complementary information. Unlike lasso, ridge regression avoids aggressive variable elimination, which could discard clinically relevant predictors with small individual effects.

4.1.1 Model Specification

Let $Y_i \in \{0, 1\}$ denote the cancer status of individual i, where $Y_i = 1$ indicates a malignant lesion. The ridge logistic regression model estimates the conditional probability, while minimizing the penalized negative log-likelihood, where $\lambda \geq 0$ controls the strength of the ridge penalty. This penalty discourages large coefficient values, reduces model variance, and thus stabilizes estimation.

4.1.2 Preprocessing and Implementation

All character-valued predictors were converted to factor variables using a model matrix, which automatically generates dummy variables for each categorical level. This enables logistic regression to accommodate the mixture of numerical and categorical predictors present in the data.

Model fitting was performed using the `glmnet` package, with ridge regularization specified by setting the tuning parameter $\alpha = 0$. The regularization strength λ was selected using 10-fold cross-validation via the `cv.glmnet` procedure. The value minimizing the cross-validated deviance ($\lambda_{\min}$) was chosen for the final model, providing a balance between bias and variance while also fully leveraging the large sample size.

4.1.3 Prediction and Classification

The fitted ridge model produces estimated probabilities of malignancy for each observation in the test dataset. These probabilities were converted into class labels using a probability threshold of 0.5, classifying observations with predicted probability greater than or equal to 0.5 as malignant and the remainder as benign. Given the near balance between benign and malignant cases in the dataset, this threshold provides a natural and interpretable decision rule, making accuracy an appropriate evaluation metric.

4.1.4 Performance

The ridge-penalized logistic regression model achieved a Kaggle accuracy of approximately 0.60465. While this accuracy is modest in absolute terms, the model served as a stable baseline that informed subsequent modeling decisions.

Importantly, the ridge model's performance suggests that much of the signal in the data can be captured through linear combinations of predictors, motivating us to further explore linear models. We evaluated additional linear and nonlinear classification methods to assess whether alternative modeling assumptions could improve performance. These included linear discriminant analysis (LDA), as well as nonlinear approaches such as random forests and generalized additive models (GAMs).

LDA was particularly appealing given the approximately linear separability suggested by both exploratory analysis and ridge regression results. In contrast, tree-based methods were explored to test whether nonlinear interactions provided meaningful gains in accuracy. However, these methods did not consistently outperform ridge regression and often exhibited reduced stability or interpretability.

4.2 Model 2: Linear Discriminant Analysis

Model 2 applies Linear Discriminant Analysis (LDA) to predict skin cancer outcomes using a reduced yet informative set of characteristics. LDA was chosen because of its interpretability, computational efficiency in large datasets, and strong theoretical foundation for binary classification problems. This model focuses on balancing predictive performance with simplicity while controlling for overfitting through cross-validation.

4.2.1 Model Specification

The LDA model was trained using 40 predictors, after excluding 9 non-informative or unrelated variables. The response variable, Cancer, was treated as a binary categorical outcome with two classes: Benign and Malignant. The model assumes multivariate normality within each class and equal covariance matrices across classes, consistent with standard LDA assumptions. While these assumptions may not hold flawlessly in practice, LDA is known to be robust in large-sample settings, particularly when class means are well separated as indicated by the EDA.

4.2.2 Preprocessing and Implementation

Missing values were handled using median imputation for numeric variables and mode imputation for categorical variables, computed from the training set only to avoid data leakage. All categorical predictors were explicitly converted to factors, and factor levels in the test set were aligned with those in the training set. After imputation, both training and test datasets contained no missing values.

The model was implemented using the `caret` framework in R, which streamlined training and validation. A 10-fold cross-validation scheme was employed to estimate out-of-sample performance and improve the robustness of the accuracy estimates.

4.2.3 Prediction and Classification

LDA constructs linear combinations of the predictors (discriminant functions) that maximize separation between the benign and malignant classes. Observations were classified based on the discriminant score that yielded the highest posterior probability. Final predictions were generated for both the training set (to assess in-sample performance) and the test set, with predicted class labels used for submission.

4.2.4 Performance

When evaluated on the held-out test set and submitted to Kaggle, the LDA model achieved a public leaderboard score of 0.60480, indicating strong generalization performance. Among all models tested using the same set of 40 predictors, this LDA specification produced the best overall results and was therefore selected as the final model.

4.3 Other Methods

In addition to LDA and Ridge Regression, we experimented with several alternative models, including XGBoost, generalized linear models (GLM), support vector machines (SVM), exponential models, random forests, decision trees, and k-nearest neighbors (KNN). Each model was evaluated using either median/mode imputation or random forest based imputation, where applicable. Despite extensive tuning and validation, none of these approaches produced superior predictive performance compared to LDA when restricted to the same set of 40 predictors, and therefore were not selected as the final model.

5 Discussion and Limitations

An accuracy of approximately 0.6 indicates substantial overlap between benign and malignant feature distributions, highlighting the limited discriminative power of the current sets of predictors. Several data-related limitations contribute to this outcome. Approximately 8% of values were missing across predictors, requiring imputation to preserve sample size. Median imputation for numerical variables and mode imputation for categorical variables proved to be the most stable and computationally efficient approach, though this strategy relies on the assumption that missingness is not informative of cancer status. Violations of this assumption could introduce bias. In addition, the available predictors are limited to survey-style and tabulated variables; the absence of imaging data and clinically assessed dermatological features—such as lesion symmetry, texture, swelling, or laboratory results—likely bottlenecks predictive performance.

Despite these limitations, the results provide meaningful insight into the structure of the data and the suitability of different modeling approaches. The modest accuracy achieved does not reflect a failure of statistical modeling, but rather underscores the inherent difficulty of the task given the available information. In clinical dermatology, where the cost of false negatives is substantially higher than that of false positives, evaluation metrics such as sensitivity or the area under the ROC curve may be more appropriate than overall accuracy. The analysis suggests that predictive signal is distributed across many variables with relatively small individual effects and is largely captured through linear structure. From an applied perspective, this implies that interpretable statistical models may serve as preliminary risk-screening or decision-support tools, while comprehensive diagnosis should continue to rely on clinical expertise and imaging-based methods.

6 Conclusion and Recommendation

To conclude, this analysis demonstrates that structured demographic, behavioral, and lesion-related data contain meaningful but limited signal for distinguishing benign from malignant skin lesions. The most informative predictors align with established dermatological risk factors, and the overall structure of the data suggests that predictive information is distributed across many variables with modest individual effects. As a result, linear statistical models are well-suited to this setting, offering stable performance and interpretability without relying on complex model structure. Future improvements would benefit less from additional model complexity and more from richer clinical features and evaluation metrics that prioritize medically relevant outcomes.

References

[1] Armstrong, B. K., & Kricker, A. (2001). The epidemiology of UV induced skin cancer. *Journal of Photochemistry and Photobiology B: Biology, 63*(1–3), 8–18. https://doi.org/10.1016/S1011-1344(01)00198-1

[2] Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Zanetti, R., Masini, C., Boyle, P., & Melchi, C. F. (2005). Meta-analysis of risk factors for cutaneous melanoma: II. Sun exposure. *European Journal of Cancer, 41*(1), 45–60. https://doi.org/10.1016/j.ejca.2004.10.016

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.