Predicting Skin Cancer Status

Justin Lee, Kimberly Cui, Adya Ganti, Jay Enugula, Djovan Velasco
Professor Almohalwas
STATS 101C

December 17, 2025

Abstract

Skin cancer is one of the most common cancers worldwide, and early detection is critical for improving patient survival. The objective of this project is to develop and evaluate statistical learning models to predict skin cancer status (benign vs. malignant) using a Kaggle dataset containing 50 predictor variables describing tumor characteristics. This report presents the findings of our research, including exploratory data analysis, data cleaning, feature selection, model construction, and overall model accuracy. Multiple classification methods were compared when testing prediction accuracy and model complexity, with the end goal of identifying a model that is both accurate and interpretable for our dataset. The final selected model highlights how statistical learning techniques can aid in predicting skin cancer status, as well as reveals the trade-off between model complexity and predictive performance. Limitations and potential improvements are also discussed.

Our final model is based on logistic regression and uses 19 significant predictors to predict an individual's cancer status. It has a Kaggle score of 0.60585, making it past the 0.60485 accuracy threshold needed.

Table of Contents

1 Introduction

Skin cancer is one of the most common cancers worldwide. It develops when DNA mutations cause skin cells to grow uncontrollably and form tumors. While many cases of skin cancer are highly treatable when detected early, a late or inaccurate diagnosis can significantly reduce survival rates. Therefore, it is important to detect skin cancer as early as possible to improve patient outcomes, which will help physicians identify individuals who may be at a higher risk.

In recent years, data-driven methods have become extremely useful for supporting medical decision-making. By analyzing patterns in large datasets containing relevant information, several statistical learning models can help identify disease outcomes and flag critical risk factors.

In this project, we use a publicly available Kaggle dataset to build prediction models that help predict an individual's skin cancer status as either benign or malignant. The dataset consists of a training set with approximately 50,000 individuals and a testing set with approximately 20,000 individuals. Each observation represents one individual and includes 50 predictor variables describing demographic, biological, environmental, and lifestyle characteristics related to skin cancer risk. These predictors include information such as age, skin tone, sun exposure, sunscreen habits, environmental conditions, and health history. Our goal is to use these variables to accurately predict skin cancer outcomes to help physicians detect high-risk patients quicker, potentially increasing survival rates.

2 Methods

2.1 Dataset

To conduct this analysis, we were provided with two datasets for prediction: a training dataset and a testing dataset. The training dataset contains 50,000 observations with 49 predictor variables and a binary response variable, Cancer, with values "Benign" and "Malignant." The testing dataset consists of 20,000 observations and includes the same 49 predictor variables, but does not contain the response variable.

Each observation represents a single individual and includes demographic, biological, environmental, and lifestyle characteristics related to skin cancer risk. The primary objective of this study is to use the training data to develop a prediction model that can accurately predict cancer status and then apply this model to generate predictions for the testing dataset.

2.2 Data Cleaning

The first step in the data cleaning process was to remove missing NA values and assess potential multicollinearity among the predictor variables. None of the predictors contained more than 80% missing values; in fact, the highest proportion of missing values observed in any predictor was 8.32

To address multicollinearity, we calculated the variance inflation factors (VIF) for the predictor variables. Since none of the predictors exhibited high VIF values, multicollinearity was not considered a concern, and no variables were removed.

Next, predictor selection was performed to remove variables with little to no relevance to skin cancer prediction. This process combined intuition-based reasoning with formal statistical tests. Based on intuition, predictors such as desk height and favorite color were removed, as they seemed unlikely to have a meaningful effect on skin cancer outcomes. For categorical predictors, chi-squared tests were used to assess their association with the response variable, and predictors with insignificant p-values were removed. Furthermore, t-tests were applied to numerical predictors to identify variables with weak relationships to the response.

After this selection process, a total of 18 predictors remained for use in the final model. These predictors were cross-referenced with results from a random forest variable importance analysis, which confirmed that the retained predictors captured the most significant variables in the dataset.

Finally, we imputed the remaining missing values using the imputation method missForest, which was chosen due to its accurate performance on complex datasets with many variables and potential interactions. When compared to alternative imputation techniques, such as MICE and mean or mode imputation, models using missForest imputation consistently achieved higher prediction accuracy, indicating that it was a more effective way to handle missing data.

2.3 Procedure

Step 1: Clean data and remove unnecessary predictors.

Step 2: Use R to create models using the training data to predict cancer: "Benign" or "Malignant". The different models are:

1. Random Forest
2. XGBoost
3. Logistic Regression

Step 3: Use the models to perform predictions on the testing data and fill in the missing responses column in the data.

Step 4: Classify which model gave us the most accurate prediction and report its accuracy.

3 Data Analysis

3.1 Predictor Selection

After completing the data cleaning and imputation, predictor selection was performed to identify the variables most relevant to predicting skin cancer status while reducing unnecessary model complexity. For categorical predictors, chi-squared tests were used. Predictors with insignificant p-values were removed from the data used for the final model. For numeric predictors, two-sample t-tests were used similarly to assess variable significance. Any numeric predictors with insignificant p-values were removed from the final data. This ultimately left us with the 18 predictors listed:

- residence_lon
- hat_use
- education
- exercise_freq_per_week
- skin_photosensitivity
- sunscreen_spf
- lesion_size_mm
- clothing_protection
- tanning_bed_use
- outdoor_job
- sunburns_last_year
- sunscreen_freq
- number_of_lesions
- avg_daily_uv

- skin_tone
- immunosuppressed
- family_history
- age

3.2 Model Testing

In order to run predictions on our selected predictors we used the following three models: Logistic regression, random forest, and XGBoost. All models were training using the same cleaned data to ensure consistency across our predictions. For logistic regression, the binary response variable was coded as Benign = 0 and Malignant = 1. Predicted probabilities were generated and converted into class predictions using a threshold value of 0.5. Random Forest and XGBoost models were initialized using default parameters and then were tuned to maximize accuracy. After training these models they were used to predict upon the test data which generated our prediction submission on kaggle. Overall the highest scoring model was the logistic model, followed by the XGBoost, then random forest.

4 Results

4.1 Model Comparison

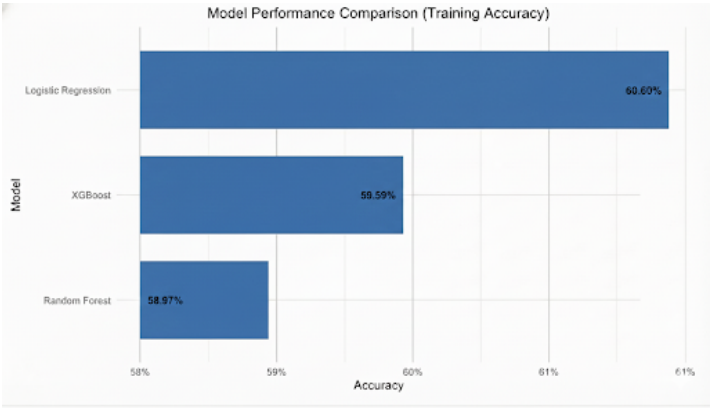

Figure 1: Model Performance Comparison (Training Accuracy)

The performance of the three models, Logistic Regression, XGBoost, and Random Forest, was evaluated based on their training accuracy to determine the most ideal approach for predicting skin cancer status. These results show that Logistic Regression achieved the highest training accuracy at 60.60%.

4.2 Logistic Regression

Logistic regression was an appropriate modeling choice for this analysis because we have a binary response variable. The response variable, "Cancer", was coded so that Benign = 0 and
Malignant = 1, allowing the model to estimate the probability that a tumor is malignant. We ran this model using the selected predictors from our data cleaning and variable selection. After fitting the model, predicted probabilities were generated for the testing dataset. These probabilities were converted into class predictions using a threshold of 0.5, where observations with predicted probabilities greater than 0.5 were classified as Malignant and those below 0.5 were classified as Benign. The model achieved an overall accuracy of 0.60585 on the testing dataset, which was our highest accuracy.

4.3 Confusion Matrix

Confusion Matrix for Training Data

		Benign	Malignant
Predicted Class	Benign	**15689**	**7048**
	Malignant	**8179**	**19084**

Figure 2: Confusion Matrix

The confusion matrix provides a breakdown of the model's predictions versus the actual cancer status, detailing the number of correctly and incorrectly classified instances.

- True Positives (TP): The model correctly predicted 19,084 cases as Malignant.
- True Negatives (TN): The model correctly predicted 15,689 cases as Benign.
- False Negatives (FN): The model incorrectly classified 8,179 malignant cases as Benign. In a medical context, this is a critical type of error (Type II) as it represents missed diagnoses.
- False Positives (FP): The model incorrectly classified 7,048 benign cases as Malignant. This error (Type I) would lead to unnecessary patient anxiety and follow-up procedures.

The overall training accuracy of 60.60% is derived from this matrix, calculating (TP + TN)/Total observations. The confusion matrix allows for a more detailed analysis of the trade-off between sensitivity and specificity, correctly identifying Malignant and Benign Cases.

4.4 Density/Conditional Plots

To gain a better insight into the relationship between the response variable (Cancer Status) and the most important continuous and categorical predictors, density and conditional bar plots were generated. The variable importance analysis identified four key predictors for the model: Age, Average Daily UV, Skin Tone, and Immunosuppressed.

Continuous Predictors(Age and Average Daily UV Exposure):

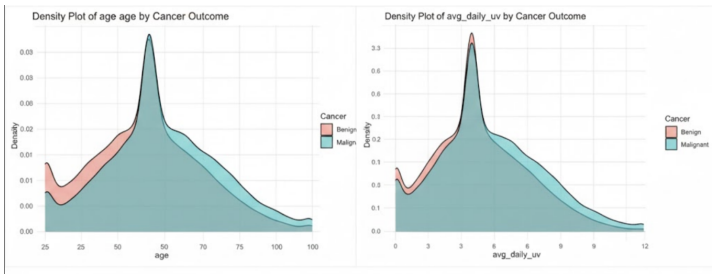

Figure 3: Continuous Density Plot

Age: In comparison to those with benign cancer (red), the distribution of those with malignant cancer (blue) is pushed to the right in the density plot for age. This suggests that older members of the dataset, especially those over 55, have a higher relative risk and a larger density of cancer cases. Average Daily UV Exposure: In a similar comparison, there is a modest change in the distribution of malignant cases toward greater UV exposure values. The figure supports the established link between

long-term UV exposure and the risk of skin cancer by showing that those with an average daily UV index between 5 and 7 had a higher density of malignant cases than benign instances.

Categorical Predictors (Skin Tone and Immunosuppressed Status):

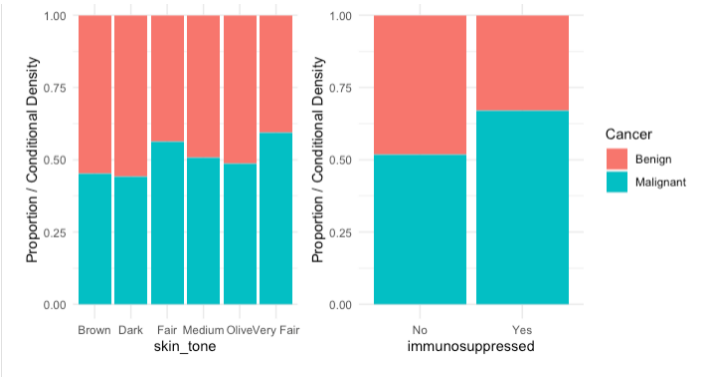

Figure 4: Categorical Density Plot

Skin Tone: The percentage of malignant cases in each category of skin tone is displayed in this graphic. Among all the skin tones, the "Very Fair" group has the largest conditional proportion of malignant cases (more than 50%) compared to benign instances. This is the most notable finding. Given that lighter skin tones are often linked to an increased risk of UV damage and cancer, this validates the variable's significance in the model. Immunosuppressed Status:The plot contrasts the cancer status of

people who are immunosuppressed ("Yes") and those who are not ("No"). The percentage of malignant instances is significantly and significantly greater (about 70%) for those who are immunosuppressed ("Yes") than for people who are not immunosuppressed ("No"). This substantial link emphasizes how important immune function is to the body's defense against the development of cancer and shows why this variable is a key predictor in the model.

5 Discussion

The primary goal of this project was to investigate how different classification models could be used in order to predict the probability that someone has skin cancer given a mix of several numeric and categorical predictors. We tested three separate models (Logistic Regression, Random Forest, and XGBoost) and ultimately settled on the logistic model which gave us our highest kaggle accuracy and provided the best balance between interpretability and performance.

While initially we assumed that for a dataset as complex as this one that models such as Random Forest or XGBoost would have performed the best, in reality it was the opposite. What appeared on the surface to be an incredibly complex data set was really confounded by many unnecessary predictors that made the data seem more complex than it really was. After cleaning it was apparent that the removed predictors in all reality were not useful to the model and thus these models that specialize in complex and non-linear relationships were not that accurate. Instead the remaining predictors favored a linear model such as logistic regression which proved to be the most efficient and accurate model.

One potentially alarming problem we encountered in the prediction was the shocking number of false positives and negatives present in the data. Out of the 50,000 observations we tested about 15,000 or roughly 30

The density and conditional plots provided give valuable insight into the predictors most strongly associated with skin cancer status. Age and average daily UV showed higher malignant risk as the values increased. This aligns with medical research which shows that cumulative sun exposure and aging are both critical factors to increased skin cancer risk. Similarly, categorical predictors such as

skin tone and immunosuppressed status demonstrated strong associations with malignancy. Individuals with very fair skin and those who are immunosuppressed showed higher proportions of malignant cases, which reinforces the strength of these predictors and validates their importance in the model.

Despite the strengths, there are several present limitations to our model. For starters our accuracy was only 60.59

Overall, while our final logistic regression model is not perfect, it provides meaningful insights into key risk factors for skin cancer and serves as a strong foundation for future improvement.

6 Acknowledgement

We would like to give a special thank you to Professor Akram Almohawas for all of the knowledge he shared with us. We greatly appreciate his patience, help, and guidance throughout the quarter.

References

[1] *AI Model Powers Skin Cancer Detection Across Diverse Populations.* https://today.ucsd.edu/story/ai-model-powers-skin-cancer-detection-across-diverse-populations

[2] *Diagnosis of skin cancer by correlation and complexity analyses of damaged DNA* https://pmc.ncbi.nlm.nih.gov/articles/PMC4767458/

[3] *The association between skin characteristics and skin cancer prevention behaviors* https://pmc.ncbi.nlm.nih.gov/articles/PMC2759861/

[4] *Differentiation Between Benign and Malignant Pigmented Skin Tumours Using Bedside Diagnostic Imaging Technologies: A Pilot Study* https://pmc.ncbi.nlm.nih.gov/articles/PMC9631264/