# Predicting Skin Cancer Status

By: Justin Lee, Kimberly Cui, Adya Ganti, Jay Enugula, Djovan Velasco

# Table of **Contents**

# 1) Introduction

# Background & Motivation

**What is skin cancer?**
Skin cancer is one of the most common cancers worldwide. It occurs when DNA mutations cause skin cells to grow uncontrollably and form tumors.

**Why this matters:**
Early detection significantly improves survival. Accurate classification helps physicians identify high-risk patients earlier.

**Project Goals:**

- Analyze the relationship between 50 predictors and cancer outcomes
- Classify skin cancer status (Benign vs. Malignant)
- Compare model accuracy and model complexity
- Choose the simplest, most accurate model

# Dataset Overview

## Training Dataset

| | age | gender | skin_tone | education | income | urban_rural | occupation | avg_daily_uv | sunscreen_freq | sunscreen_spf | sunscreen_type | sunscreen_brand | hat_use | clothing_protection | ta |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 46 | Male | Olive | Bachelor's | 62088 | Urban | Student | 2.44 | Often | 30 | Cream | Generic | Never | Low | N |
| 2 | 78 | Female | Fair | High School | NA | NA | Service | NA | NA | 50 | Spray | Generic | Often | High | N |
| 3 | 54 | Male | NA | Some College | 18635 | Urban | Service | 5.66 | Often | NA | Spray | BrandA | Often | Medium | C |
| 4 | 73 | NA | Medium | High School | 61623 | NA | Office | 5.6 | Rarely | 30 | Cream | Generic | Never | Medium | N |
| 5 | 39 | Female | Medium | Bachelor's | NA | NA | Outdoor | 2.25 | Always | NA | Spray | BrandB | Often | Medium | C |
| 6 | 49 | Male | NA | NA | 22178 | Urban | Outdoor | 6.34 | Always | 15 | Cream | BrandB | Often | High | N |
| 7 | 53 | Male | Brown | NA | NA | Urban | Office | 5.95 | Often | 15 | Cream | NA | NA | Medium | N |
| 8 | 65 | Male | NA | High School | 48868 | Urban | Service | 1.25 | Never | 0 | Gel | NoBrand | Never | High | N |
| 9 | NA | Male | Medium | High School | 39618 | Urban | Office | NA | Often | NA | Cream | Generic | Often | Low | N |
| 10 | 74 | Female | Brown | Some College | 43425 | Urban | Office | 3.44 | Often | 50 | Cream | Generic | Always | Medium | N |
| 11 | 55 | Male | Olive | Bachelor's | 8933 | Suburban | Outdoor | 3.85 | Often | 30 | Spray | NA | Often | Medium | N |
| 12 | 70 | Male | Very Fair | Bachelor's | 8970 | Suburban | NA | 3.94 | Often | 30 | NA | Generic | Never | Medium | N |
| 13 | 49 | Female | Fair | NA | 20209 | Suburban | Healthcare | 3.86 | Sometimes | 15 | Spray | Generic | Always | Medium | N |
| 14 | 62 | Male | Olive | NA | 348729 | Suburban | Service | NA | Sometimes | 100 | Cream | Generic | NA | Medium | N |
| 15 | 50 | Female | Medium | High School | 9230 | Urban | Retired | 4.75 | Often | 30 | Cream | Generic | Never | Medium | N |
| 16 | 33 | Female | Fair | High School | 22389 | Urban | Service | 4.37 | Always | NA | NA | BrandC | Never | Medium | N |
| 17 | 61 | Female | Very Fair | High School | 53267 | Urban | Other | 2 | Sometimes | 15 | NA | BrandA | Often | High | N |
| 18 | 53 | Female | Brown | NA | 43125 | Rural | Outdoor | 2.15 | Rarely | 30 | Cream | BrandA | Never | Low | N |
| 19 | 18 | Male | Medium | Bachelor's | 84552 | Urban | Office | 4.26 | Sometimes | 15 | Cream | NA | Sometimes | Medium | C |
| 20 | 59 | Female | Medium | Some College | 110861 | Rural | Student | 2.01 | Never | 0 | Lotion | BrandB | Often | High | N |
| 21 | 67 | Female | Very Fair | NA | 69147 | Urban | Outdoor | 1.03 | Sometimes | 15 | Lotion | NA | Always | Medium | N |

## Testing Dataset

| | age | gender | skin_tone | education | income | urban_rural | occupation | avg_daily_uv | sunscreen_freq | sunscreen_spf | sunscreen_type | sunscreen_brand | hat_use | clothing_protec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 49 | Male | Fair | PhD | NA | Suburban | Outdoor | 4.07 | Rarely | 30 | Cream | BrandA | Never | Medium |
| 2 | 36 | Female | Fair | NA | 28568 | Urban | Service | 1.46 | Always | 30 | Spray | NA | Often | Medium |
| 3 | 67 | Female | NA | High School | 15521 | Rural | Retired | 7.34 | Rarely | 50 | Lotion | Generic | Always | Medium |
| 4 | 39 | Male | Medium | Some College | 28183 | Urban | Outdoor | 6.24 | Often | 30 | Lotion | Generic | Never | Medium |
| 5 | 29 | Female | Fair | High School | 28238 | Urban | Outdoor | 3.06 | Sometimes | 50 | NA | BrandB | Sometimes | Medium |
| 6 | 34 | NA | Fair | High School | 71492 | Suburban | Service | 3.9 | NA | 0 | Lotion | Generic | Often | Medium |
| 7 | 76 | Female | Very Fair | High School | 49613 | Urban | Healthcare | 4.46 | Sometimes | 15 | Cream | Generic | Sometimes | Low |
| 8 | 43 | Male | Medium | Some College | 30415 | Suburban | Retired | NA | Often | 30 | Lotion | BrandA | Always | High |
| 9 | 23 | Female | Brown | Master's | 27243 | Suburban | NA | 5.59 | Often | 30 | Cream | BrandA | Sometimes | Medium |
| 10 | 42 | Female | NA | High School | 17456 | NA | Service | 2.5 | Sometimes | 15 | Cream | Generic | Always | NA |
| 11 | 65 | Male | Medium | Some College | NA | Suburban | Office | 3.05 | Sometimes | 50 | Spray | BrandC | Always | Medium |
| 12 | 73 | Female | Olive | Some College | 46583 | Rural | NA | 3.58 | Often | 30 | Cream | NoBrand | Sometimes | High |
| 13 | NA | Male | Brown | Bachelor's | 92141 | Rural | Retired | 2.99 | Often | 30 | Cream | Generic | Never | High |
| 14 | 43 | NA | NA | Some College | 10831 | Suburban | Outdoor | 4.51 | Rarely | 30 | Spray | BrandA | Always | Medium |
| 15 | 40 | Female | Dark | High School | 480209 | Suburban | Service | 5.01 | Often | 30 | Spray | NoBrand | Often | Low |
| 16 | NA | Male | Medium | Some College | 10083 | Urban | Outdoor | 5.57 | Rarely | NA | Cream | Generic | Never | Low |
| 17 | 61 | Male | Very Fair | High School | 83554 | Suburban | Retired | 5.11 | Often | 100 | Cream | BrandA | Always | Medium |
| 18 | 54 | Male | Very Fair | Some College | 20021 | Suburban | NA | 2.61 | Sometimes | 30 | Cream | Generic | Always | Low |
| 19 | 54 | Male | Medium | NA | 31212 | Urban | Student | 1.29 | Often | 30 | Cream | NoBrand | Sometimes | Medium |
| 20 | 62 | Male | Brown | High School | 92132 | Urban | Outdoor | 3.9 | Never | 0 | Spray | NoBrand | Sometimes | Low |
| 21 | 65 | Female | Very Fair | Bachelor's | 10482 | Suburban | Outdoor | 5.04 | Often | 15 | Cream | BrandB | Always | High |

**Response variable: Cancer (Benign / Malignant)**

Training set: 50,000– individuals
Test set: 20,000– individuals

**Observations:** Each observation represents an individual with demographic, biological, environmental, and lifestyle predictors related to skin cancer risk.

**Predictors (50):** Detailed information recorded for each person, including age, skin tone, sun exposure, sunscreen habits, environmental factors, health history, and lifestyle predictors.

Each row represents one individual with their associated risk predictors.

# 2)
# Methodology

# Data Cleaning First Step: NA’s and Multicollinearity

Dealing with NA’s

None of the predictors have more than 10% missing values (highest being 8.32%), therefore none of the actions were taken

Dealing with Multicollinearity

None of the predictors shows high correlation with each other, therefore none of the actions were taken

# Data Cleaning Second Step: Unrelated Predictors

We dropped predictors listed on **names(train)** based on how unrelated they are to skin cancer, such as desk height, music genre, and favorite color.

While keeping 100% of the observations (50,000), the amount of predictors have been reduced from 50 to 37.

We then proceeded to the modeling phase, where we aimed to develop the most accurate model for predicting flight cancellations.

# Logistic Regression (pt 1)

**Method:**

- **Logistic regression model** to classify skin cancer status (Benign vs. Malignant)
- **Removing irrelevant predictors** such as lifestyle noise variables (favorite_color, phone_brand, pets, commute_minutes, etc.)

→ After dropping these unrelated predictors, we **fit a logistic regression** using all **remaining variables** on the training dataset to predict cancer status.

```
drop_vars <- c(
  "favorite_color",
  "phone_brand",
  "music_genre",
  "favorite_cuisine",
  "preferred_shoe_type",
  "desk_height_cm",
  "zip_code_last_digit",
  "pets",
  "monthly_screen_time_minutes",
  "years_lived_at_address",
  "access_to_nude_beach",
  "near_high_power_cables",
  "commute_minutes"
)

train_clean <- train[, setdiff(names(train), drop_vars)]
test_clean  <- test[, setdiff(names(test), drop_vars)]

train_clean$Cancer <- factor(train_clean$Cancer, levels=c("Benign","Malignant"))
```

*Snippet of code (variable drop)*

**Predictors dropped (13):** favorite_color, phone_brand, music_genre, favorite_cuisine, preferred_shoe_type, desk_height_cm, zip_code_last_digit, pets, monthly_screen_time_minutes, years_lived_at_address, access_to_nude_beach, near_high_power_cables, commute_minutes

Now... **37** total predictors

**Misclassification Rate: 39.555%**

# Logistic Regression (pt 2)

**Method** ***(Continued from previous)***:

- Further simplified the predictor set by **removing variables with high p-values** in the initial logistic regression ($p > 0.05$).

```
# Statistically insignificant GLM predictors
"genderMale", "genderOther", "skin_toneDark",
"sunscreen_freqOften", "sunscreen_spf",
"frequency_doctor_visits_per_year",
"smoking_statusFormer", "smoking_statusNever",
"alcohol_drinks_per_week", "BMI",
"residence_lat", "residence_lon",
"lesion_colorBrown", "lesion_colorDark Brown",
"lesion_colorLight", "lesion_colorMulti-colored",
"lesion_colorPink",
"vitamin_d_supplementYes", "uses_smartwatchYes"
```

***Snippet of code (variable drop)***

**Predictors dropped (19):** genderMale, genderOther, skin_toneDark, sunscreen_freqOften, sunscreen_spf, frequency_doctor_visits_per_year, smoking_statusFormer, smoking_statusNever, alcohol_drinks_per_week, BMI, residence_lat, residence_lon, lesion_colorBrown, lesion_colorDark Brown, lesion_colorLight, lesion_colorMulti-colored, lesion_colorPink, vitamin_d_supplementYes, uses_smartwatchYes

Now... **18** total predictors

**For this model:**

1. Remove **irrelevant** variables
2. Remove **insignificant** variables
3. **Retrain** logistic regression model on refined predictor set

**Misclassification Rate: 39.545%**

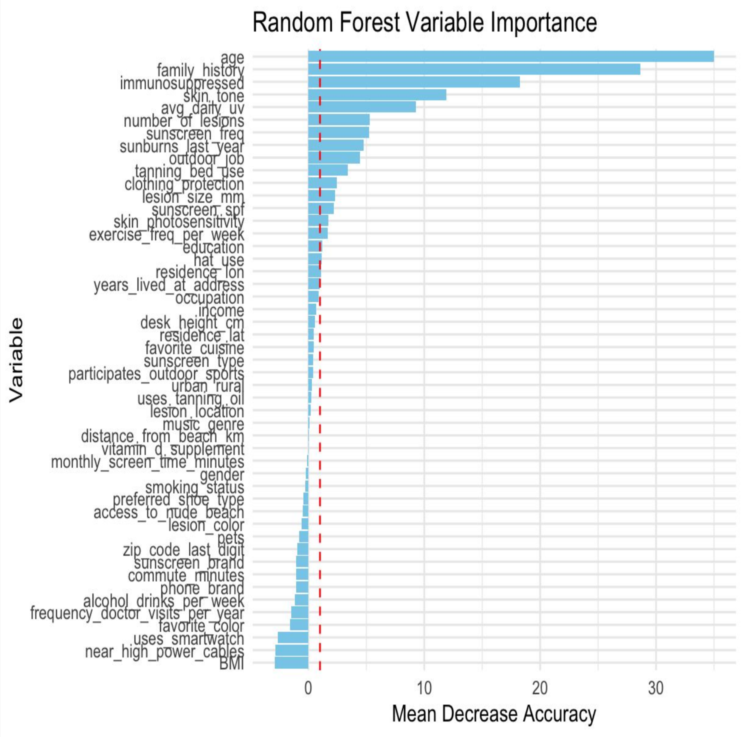
# Random Forest (Full Model)

**Method:** Train a random forest model using all selected predictors. Using mtry as our criteria we found mtry=4 to be the best.

**Key Predictors** (from importance plot):
Age, Family History, Immunosuppressed, Skin Tone, Avg_daily_uv
**MCR**: **0.41035**

**Results:**

- ➔ Many predictors had very little importance
- ➔ Moderate MCR
- ➔ Favors simpler model



# Random Forest (Simplified Model #1)

**Method:** Using variable importance, we removed lowest-importance predictors. Using MDA < 1

**Kept Predictors (18):** 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

**Key Predictors:**

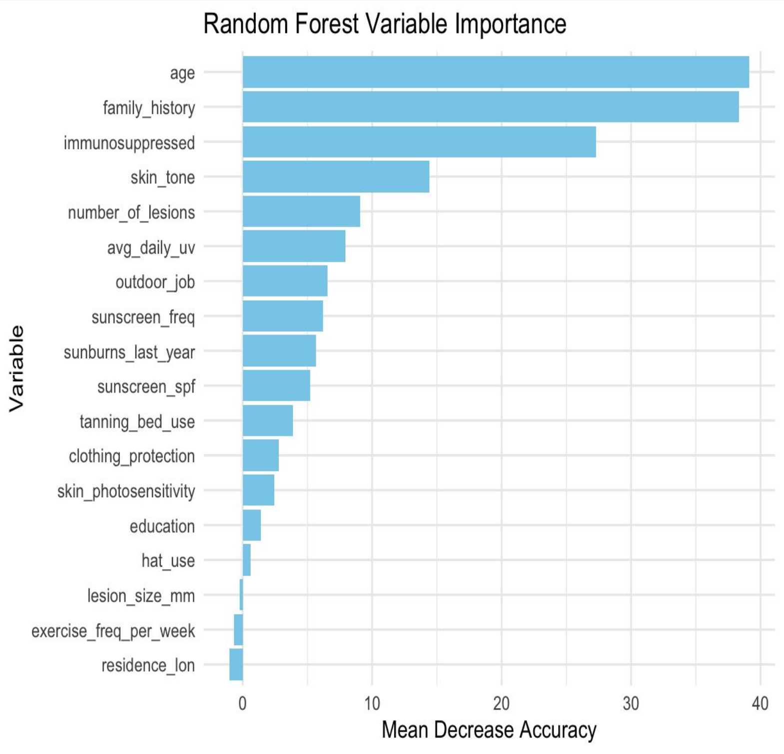
Age, Family History, Immunosuppressed, Skin Tone, number_of_lesions

**Results**:

The simplified model performs better because:

→ Less noise

→ Stronger signal-to-noise ratio




**MCR: 0.4067**

# Boosting Model

**Method:**
Trained a boosting model using the clean data predictors from above.

**MCR: 0.40405**

**Results**:

- ➔ High train acc, low test acc = overfit
- ➔ Better than both RF models

# 3) Results

# Model Comparison Table

| Model | Predictors Used | Misclassification Rate | Notes |
|---|---|---|---|
| **Logistic** | Reduced Predictors. | **39.575%** (Score: 0.60425) | GLM with `family=binomial`. Used `na.roughfix` for missing value imputation. |
| **RF (full)** | All available predictors. | **(Full): 41.035%** (Score: 0.58965) | Used `mtry = 4` and `importance = TRUE` |
| **RF (reduced)** | **(Reduced):** Subset of 17 variables (e.g., `age`, `skin_tone`, `education`). | **(Reduced): 41.545%** (Score: 0.58455) | Variables selected by ”mean decrease accuracy” |
| **Boosting** | "Cleaned" subset. Dropped 13 subjective/noise variables (e.g., `favorite_color`, `phone_brand`, `pets`) | **40.405%** (Score: 0.59595) | XGBoost model. Used `xgb.DMatrix` format. Key parameters: `eta=0.05`, `max_depth=5`, `nrounds=500` . |

# 4) Discussion

# Final Model Selected

We chose: **Logistic Regression model** after **1) removing irrelevant variables unrelated to Skin Cancer** and **2) removing predictors with high p values.**

- 50 → 37 → 18 predictors

**Most Important Predictors:** 1) age, 2) avg_daily_uv, 3) skin_toneVery Fair, 4) immunosuppressedYes

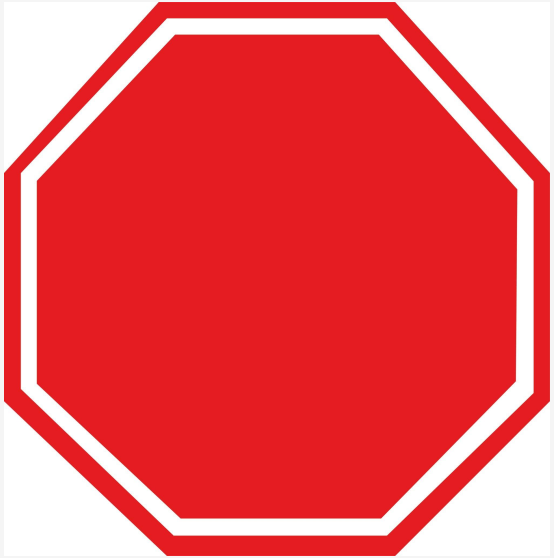
**MCR**: 39.545%

- It had the lowest misclassification rate
- It used a moderate number of predictors
- Strong performance even with noisy predictors removed
- Handles nonlinearities and interactions well

# 5) Limitations

# Limitations

- Behavioral & skincare predictors are self-reported → subject to bias
- Important clinical variables (e.g., lesion imaging) not included
- Some demographic predictors may be proxies
- Random forest lacks interpretability compared to logistic regression
- Correlation ≠ causation



# 6) Conclusion

# Conclusion

- We successfully cleaned and prepared the Kaggle dataset
- Built multiple models and compared their performance
- Identified key predictors of skin cancer risk
- Our final model achieved high accuracy with moderate complexity
- The model could help in early risk screening and decision support

# Thank You!

Special thanks to Prof Almohalwas! :)