Predicting Cancer Risk

Group 2, Project 4

Overview

Overview

This project aims to predict cancer risk by analyzing health metrics such as age, gender, BMI, smoking habits, genetic risk, physical activity, alcohol intake, and cancer history.

The goal is to support early intervention decisions and enable automated risk screening for clinics or mobile health apps, helping to detect cancer earlier and improve patient outcomes.

Data Extraction & Exploratory Data Analysis (EDA)

Extract

- Kaggle API: to extract dataset: https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset/data
- Import CSV containing records into Pandas DataFrame

Exploration

- Checked for null values and duplicate entries
-
- Visualized data distributions
-
-
- Considered distribution of data to determine approach for missing values
-

Preprocessing

- Classified features into binary, continuous, and categorical groups
- Replaced numeric labels with meaningful string categories (e.g., 0 → "No Cancer").

Understanding the Dataset

	Age	Gender	BMI	Smoking	GeneticRisk	PhysicalActivity	AlcoholIntake	CancerHistory	Diagnosis
0	58	1	16.085313	0	1	8.146251	4.148219	1	1
1	71	0	30.828784	0	1	9.361630	3.519683	0	0
2	48	1	38.785084	0	2	5.135179	4.728368	0	1
3	34	0	30.040296	0	0	9.502792	2.044636	0	0
4	62	1	35.479721	0	0	5.356890	3.309849	0	1

- **Age:** Integer (20–80 years)
- **Gender:** 0 = Male, 1 = Female
- **BMI:** Continuous (15–40)
- **Smoking:** 0 = No, 1 = Yes
- **Physical Activity:** Hours per week (0–10)
- **Alcohol Intake:** Units per week (0–5)
- **Genetic Risk:**
  - 0 = Low
  - 1 = Medium
  - 2 = High
- **Cancer History:** 0 = No, 1 = Yes
- **Diagnosis:** 0 = No Cancer, 1 = Cancer

Extract

- Kaggle API: to extract dataset: https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset/data
- Import CSV containing records

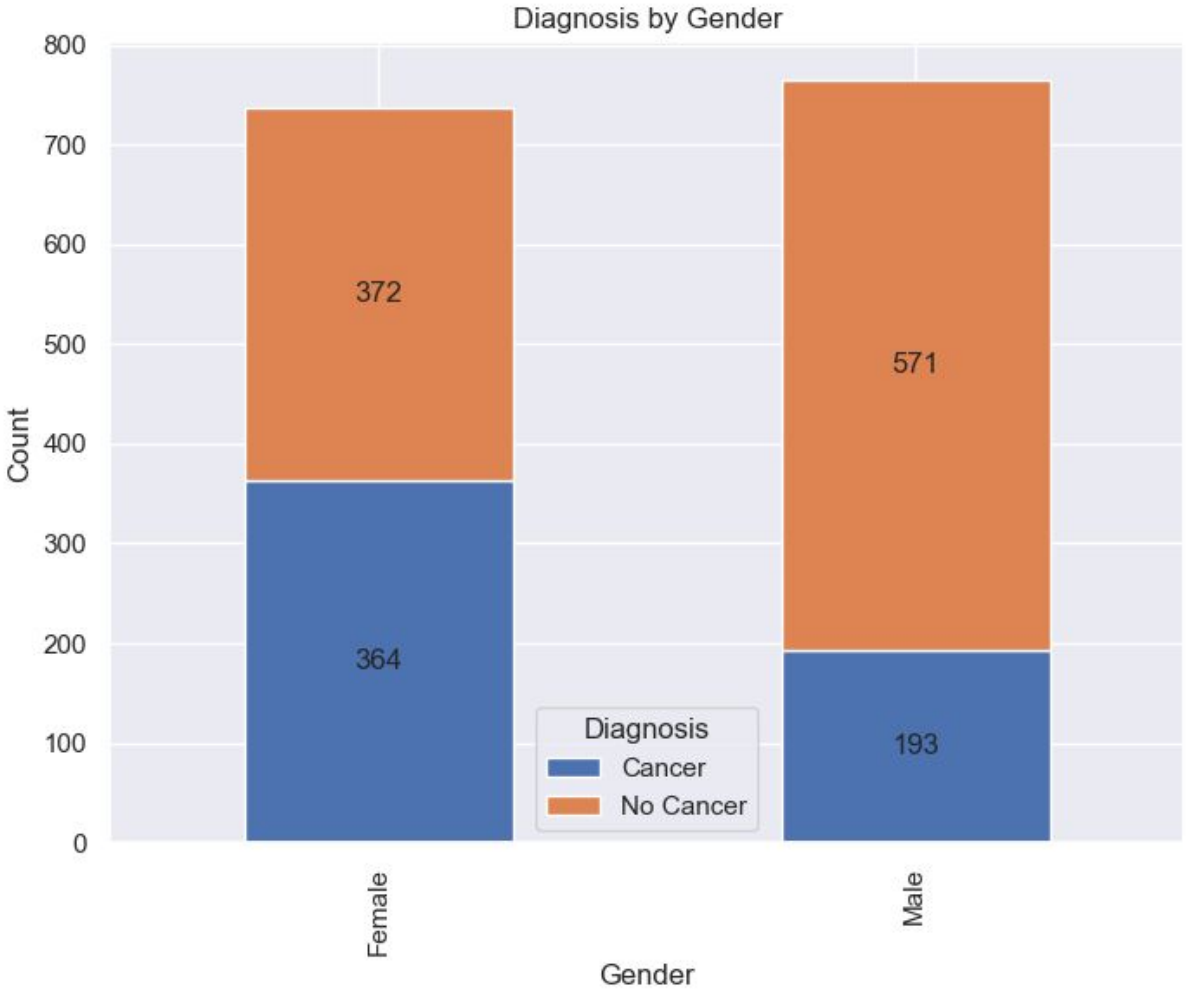
Exploration

- Checked for null values and duplicate entries
- Visualized data distributions
- Considered distribution of data to determine approach for missing values

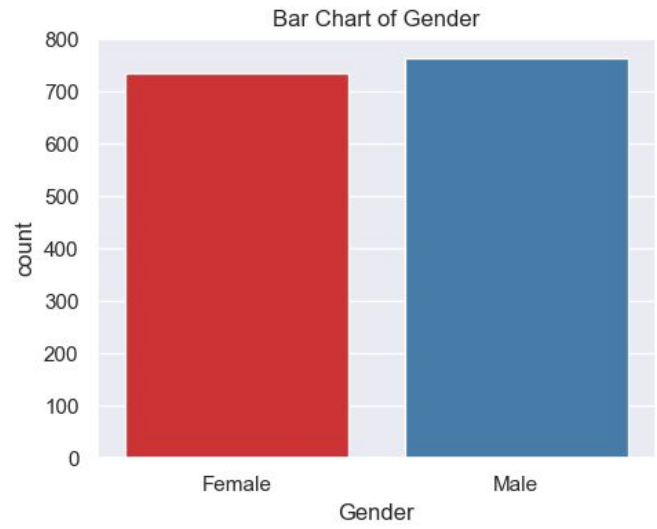
Preprocessing

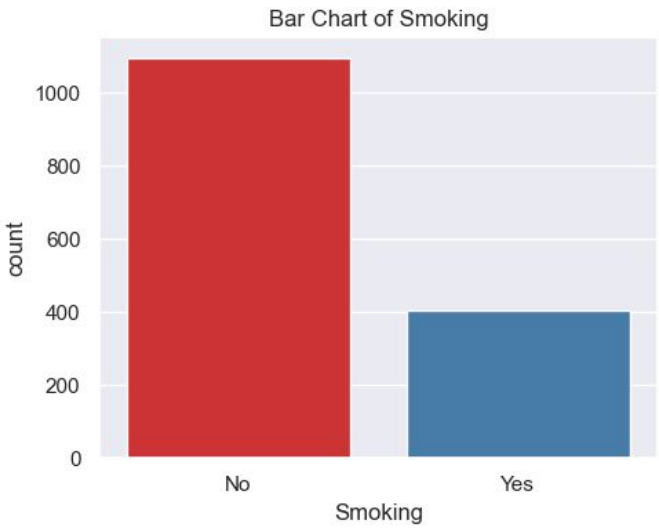
- Classified features into binary, continuous, and categorical groups
- Replaced numeric labels with meaningful string categories (e.g., 0 → "No Cancer").

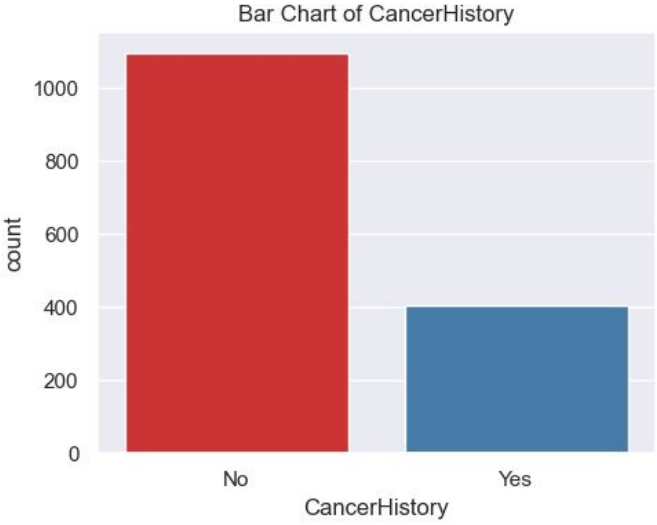
Cancer Diagnosis by Gender

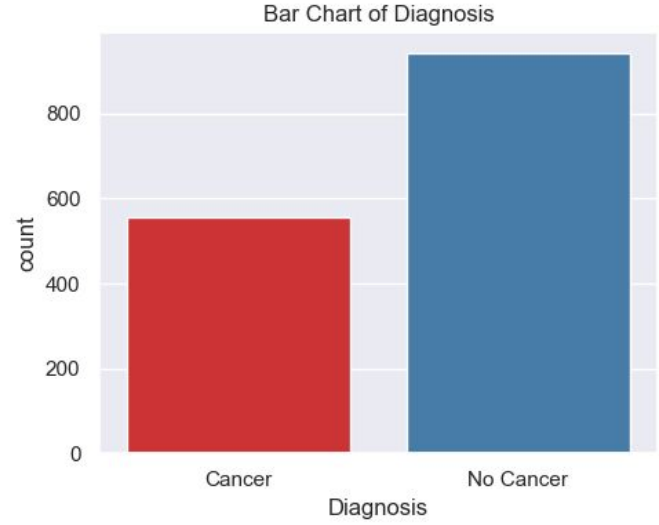


Breakdown of Binary Features

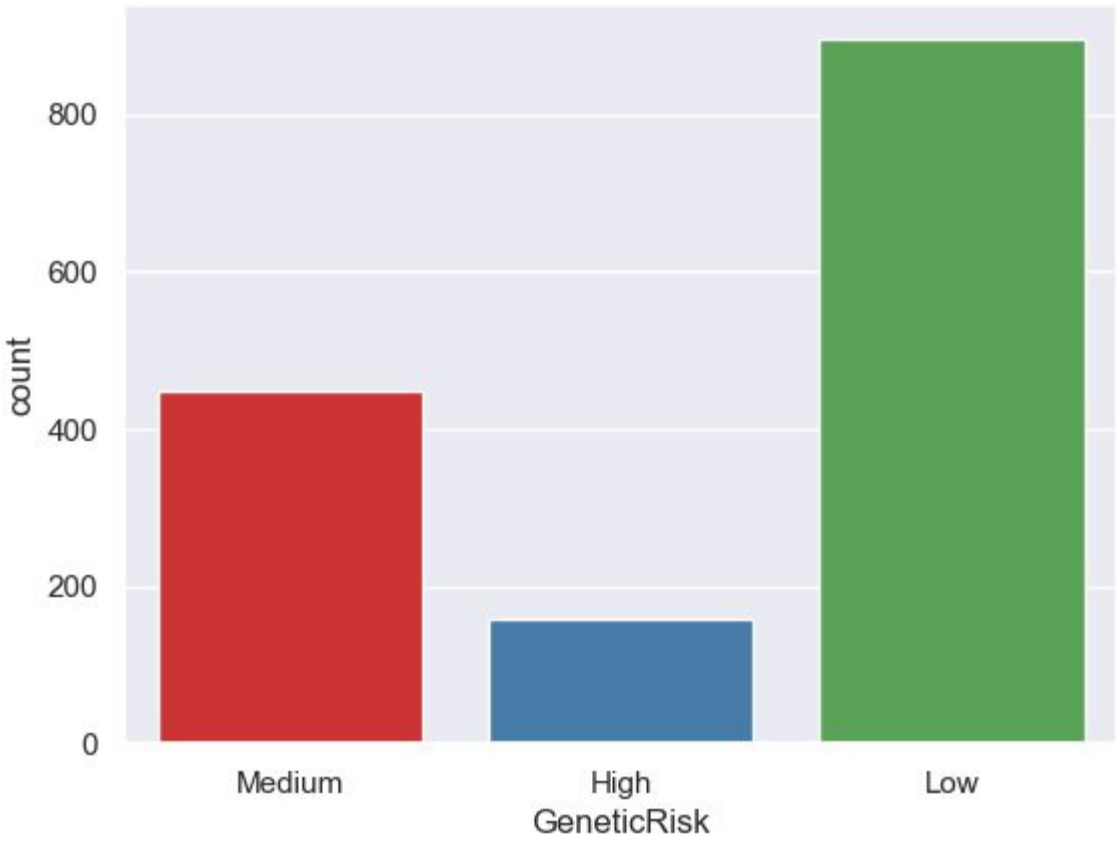








Genetic Risk



Data Preprocessing

Extract

- Kaggle API: to extract dataset: https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset/data
- Import CSV containing records

Exploration

- Checked for null values and duplicate entries
-
- Visualized data distributions
-
-
- Considered distribution of data to determine approach for missing values
-

Preprocessing

- Classified features into binary, continuous, and categorical groups
- Replaced numeric labels with meaningful string categories (e.g., 0 → "No Cancer")
- Assessed class balance in the **Diagnosis** variable.

Feature Encoding & Imputation

- Used **SimpleImputer** to fill missing values in binary and categorical features.
- Applied **OneHotEncoder** to convert categorical variables into numeric format.
- Final dataset assembled using **np.hstack()** to concatenate:
  - Scaled continuous features
  - Imputed binary features
  - One-hot encoded categorical features

Models

Logistic Regression Model

Our first **Logistic Regression model achieved 84% accuracy**, with strong performance in identifying Class 0 and a solid, slightly lower performance for Class 1.

Macro F1-score of 0.83 and **weighted F1-score of 0.84** reflect balanced overall performance, though Class 1 recall suggests a few missed positive cases.

```
Accuracy: 0.84
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.88      0.88       189
           1       0.80      0.77      0.79       111

    accuracy                           0.84       300
   macro avg       0.83      0.83      0.83       300
weighted avg       0.84      0.84      0.84       300
```

Logistic Regression Model

After tuning**, the second iteration of the Logistic Regression model achieved 88% accuracy**, with improved balance across both classes.

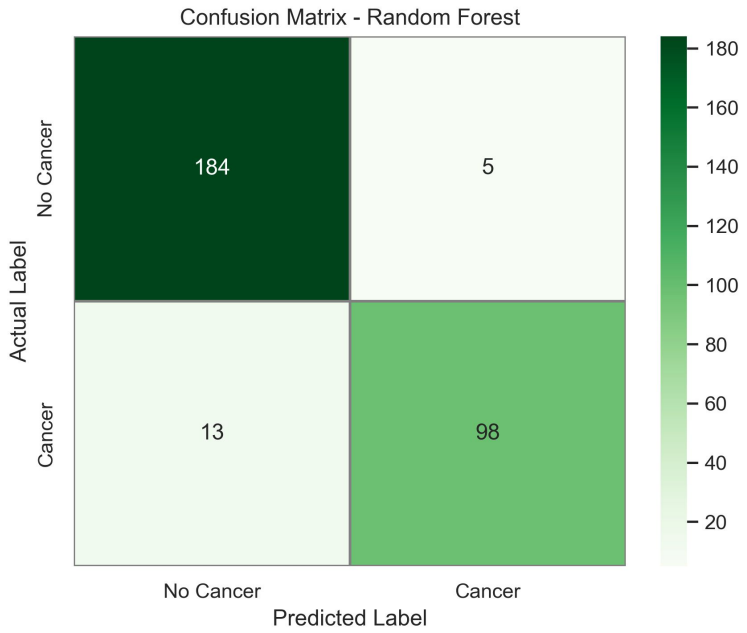
F1-scores of 0.90 (Class 0 = No Cancer) and **0.83 (Class 1 = Cancer)** indicate strong overall performance, with a **macro average F1-score of 0.87**, reflecting effective generalization and fewer missed Class 1 (Cancer) cases.

```
Accuracy: 0.88
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.91      0.90       189
           1       0.84      0.82      0.83       111

    accuracy                           0.88       300
   macro avg       0.87      0.86      0.87       300
weighted avg       0.88      0.88      0.88       300
```

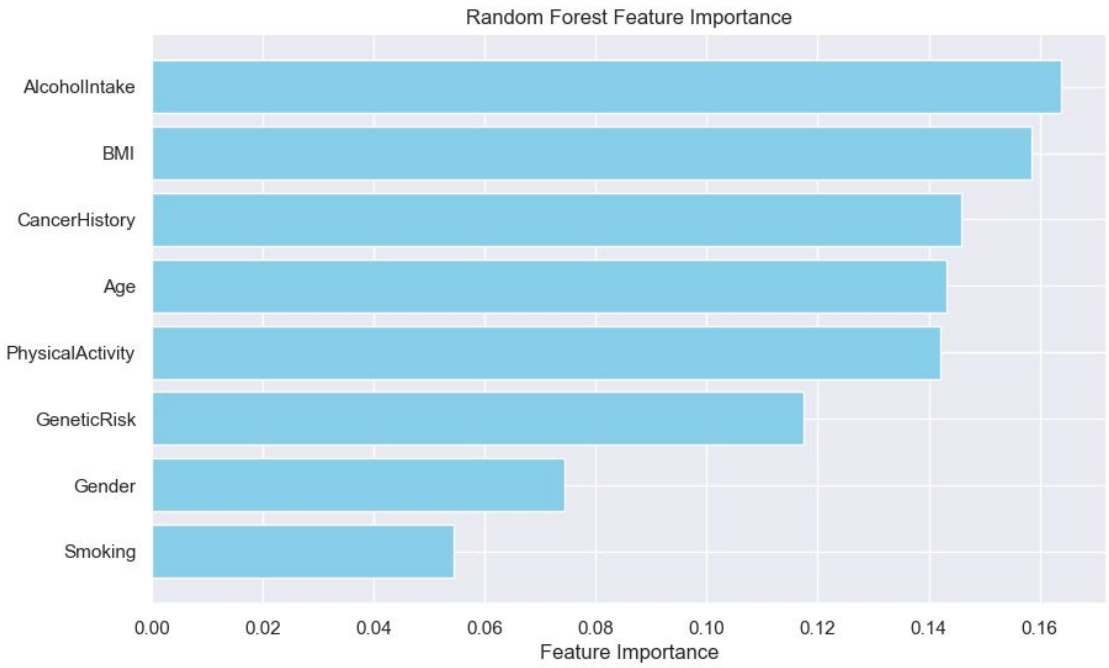
Random Forest Model

Our **Random Forest model achieved 94% accuracy**, showing excellent performance with high precision and recall across both classes.

With a **macro F1-score of 0.93** and **weighted F1-score of 0.94**, the model is highly reliable, though **Class 1 recall (88%)** leaves slight room for improvement in detecting all positive cases.



Random Forest Model



Predictive Insights

Insights

The predictive modeling indicates that lifestyle factors such as smoking, alcohol consumption, and physical inactivity, along with higher BMI and older age, significantly contribute to cancer risk.

The Random Forest model's high accuracy suggests it is a reliable tool for predicting cancer diagnosis based on these variables.

These findings underscore the importance of lifestyle modifications and targeted interventions in high-risk groups to potentially reduce cancer incidence.

Next Steps

Next Steps

- **Feature Selection**
  - Given the varying importance across features, consider prioritizing features with higher importance (e.g., Alcohol Intake, BMI, Cancer History) for a more focused model.
- **Interaction Effects**
  - Explore interactions between features, especially those with moderate to high importance, to capture more complex relationships in the logistic regression model.
- **Cross-validation**
  - Ensure the stability and generalizability of our findings by performing cross-validation on both models.

Thank you!