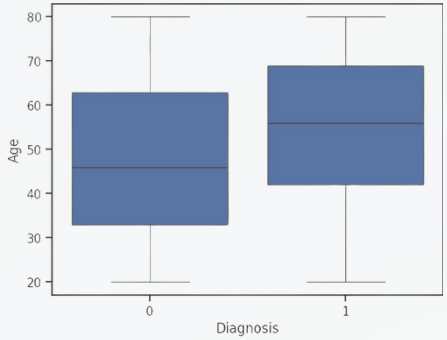
CONTENT

Abstract:

The primary idea behind this project is to develop a predictive model that can identify whether a patient has been **diagnosed with cancer** based on **eight predictors.** To do this, we plan on leveraging different algorithms and comparing their performance to select the best model.

DATA & PROBLEM DESCRIPTION

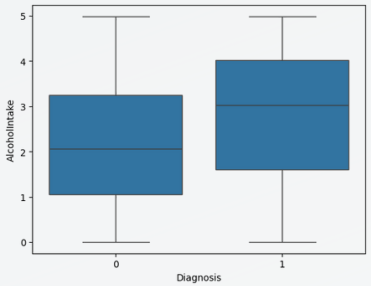
- Patient data with goal of predicting cancer risk
- **Qualitative & quantitative (encoded) variables**
- **Binary target** (no cancer (0) / cancer (1))
- 1500 observations aged from 20 to 80 years
- Slight class imbalance (no cancer: 943 obs. vs. cancer: 557 obs.)
- **8 predictors** + target column (Diagnosis)
- 80-20 ratio (80% training + 20% test)
- Physical activity (hours/week) and Alcohol Intake (alc. units/week)

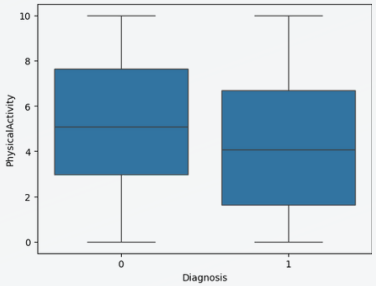


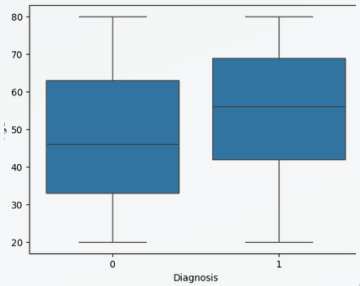
Age	Gender	BMI	Smoking	GeneticRisk	PhysicalActivity	AlcoholIntake	CancerHistory	Diagnosis
58	1	16.08531332	0	1	8.14625056	4.148219027	1	1
71	0	30.82878439	0	1	9.361630416	3.519683335	0	0
48	1	38.78508356	0	2	5.135178667	4.728367685	0	1
34	0	30.0402955	0	0	9.502792236	2.044636179	0	0
62	1	35.47972149	0	0	5.356889705	3.309849197	0	1
27	0	37.10516158	0	1	3.941904946	2.324273904	0	0
80	1	20.7019943	0	0	8.482031286	3.152943487	0	0
40	0	20.301121	1	0	4.929827105	2.24799453	1	0
58	1	30.27452472	0	1	4.719025106	0.9431612321	1	1

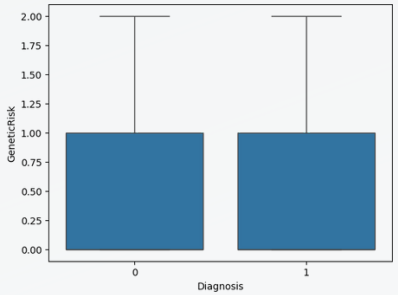
(from Kaggle)

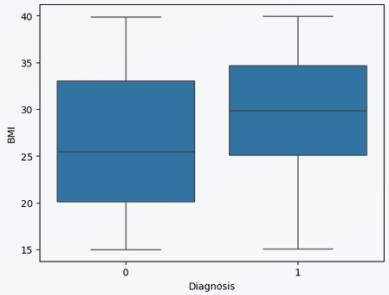
DATA & PROBLEM DESCRIPTION

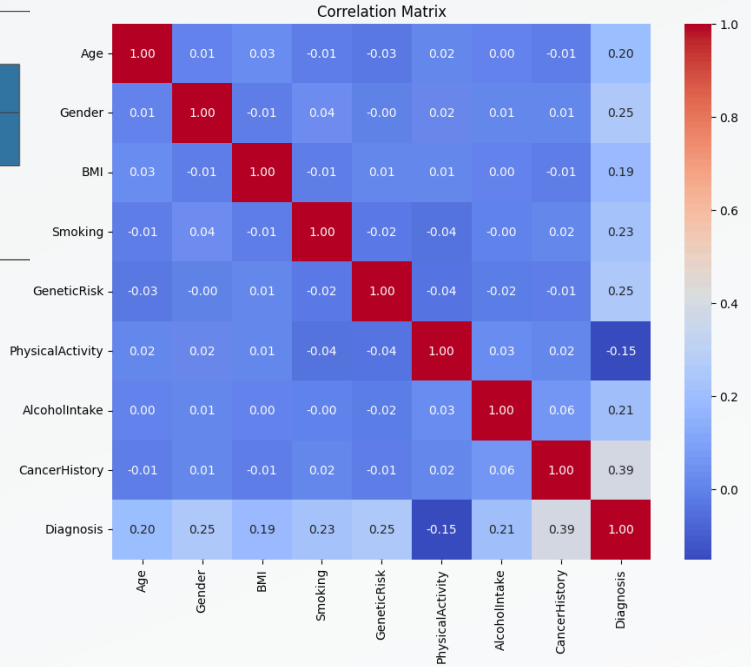












LOGISTIC REGRESSION

- **Parametric model**
- Linear decision boundary
- For **binary classification** problem
- Handles both categorical and numerical variables
- Easy to interpret the coefficients of the model
  - **The coefficients convey the influence of each predictor**
  - these are the log odds of the target variable
- The predictors within the dataset show a low correlation, ensuring stable coefficients estimates

COEFFICIENTS OF TRAINED LOGISTIC REGRESSION

Age	Gender	BMI	Smoking	Gen. Risk	Phys. Act.	Alc. Intake	Cancer Hist.
0.81	**1.78**	0.80	1.77	1.39	**-0.67**	0.77	**3.66**

DISCRIMINANT ANALYSIS

LINEAR

```
accuracy score: 0.86
roc_auc_score: 0.8396739130434783
Accuracy: 0.86
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.93      0.89       184
           1       0.87      0.75      0.81       116

    accuracy                           0.86       300
   macro avg       0.86      0.84      0.85       300
weighted avg       0.86      0.86      0.86       300
```

QUADRATIC

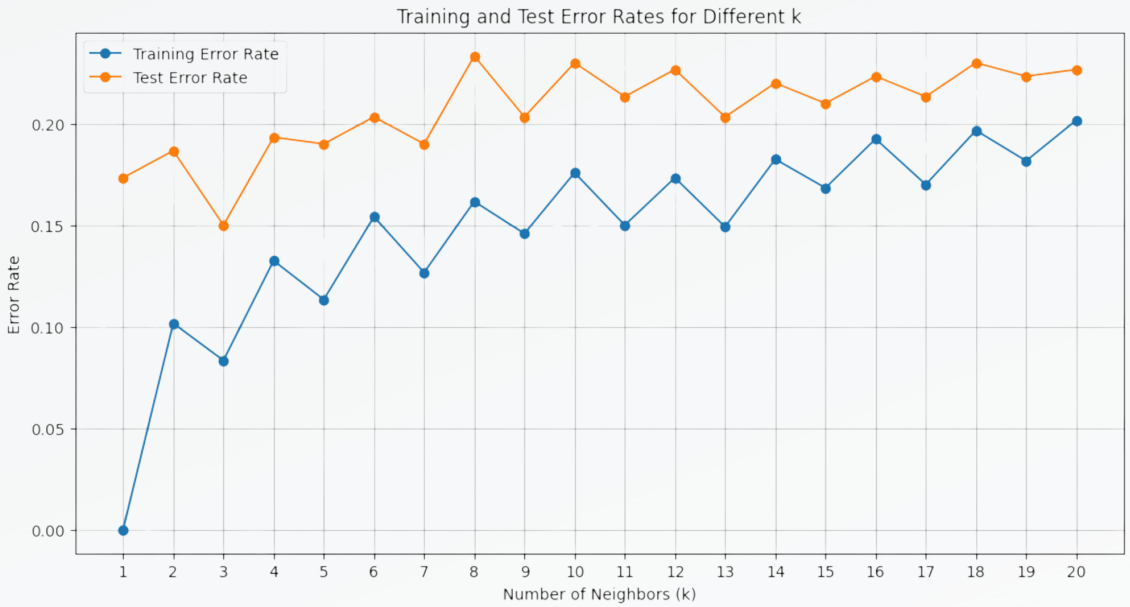
```
QDA Accuracy: 0.8533333333333334
QDA Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.92      0.89       184
           1       0.86      0.74      0.80       116

    accuracy                           0.85       300
   macro avg       0.85      0.83      0.84       300
weighted avg       0.85      0.85      0.85       300
```

K-NEAREST NEIGHBORS

- **Non-parametric**: KNN does not make any assumptions about the data distribution, which makes it suitable for a wide range of problems.
- **Sensitive to Scale**: The algorithm uses distance to make predictions. We standardize the dataset before we use it.
- **Choice of K values**



PRECISION, RECALL & F1

Precision - *What proportion of positive identifications was actually correct? [2]*

$$precision = \frac{tp}{tp + fp}$$

Good for when the cost of **false positives** is high!
(False positive = pred. cancer but has no cancer)

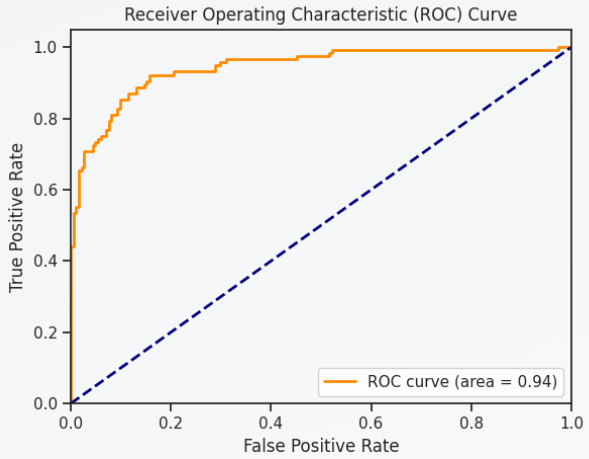
Recall - *What proportion of actual positives was identified correctly? [2]*

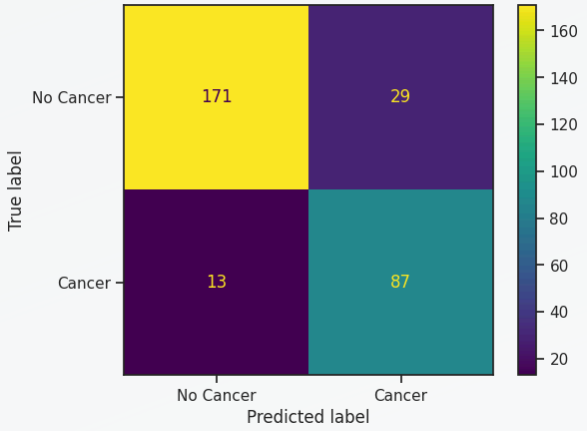
$$recall = \frac{tp}{tp + fn}$$

Good for when the cost of **false negatives** is high!
(False negative = pred. no cancer but has cancer)

RESULTS - LOG. REGRESSION

TRAINING ERROR RATE: 0.153
TEST ERROR RATE: 0.140
ROC AUC: 0.94
ACCURACY: 0.86

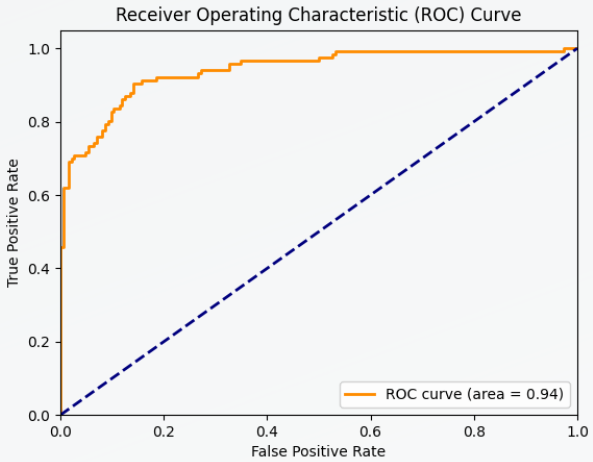


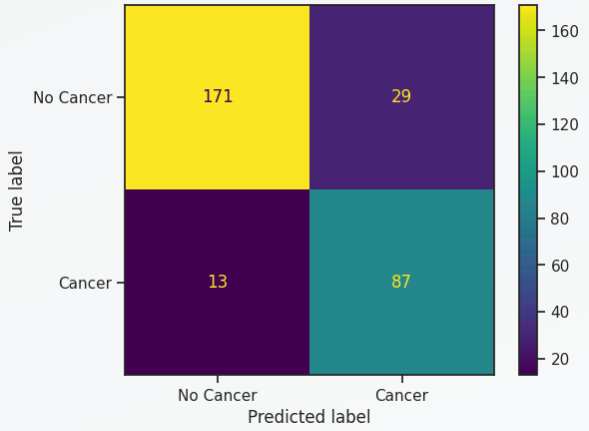


	Precision	Recall	F1	Support
0	0.85	0.93	0.89	184
1	0.87	0.75	0.81	116

RESULTS - LINEAR DISCR. ANALYSIS

TRAINING ERROR RATE: 0.153
TEST ERROR RATE: 0.140
ROC AUC: 0.94
ACCURACY: 0.86

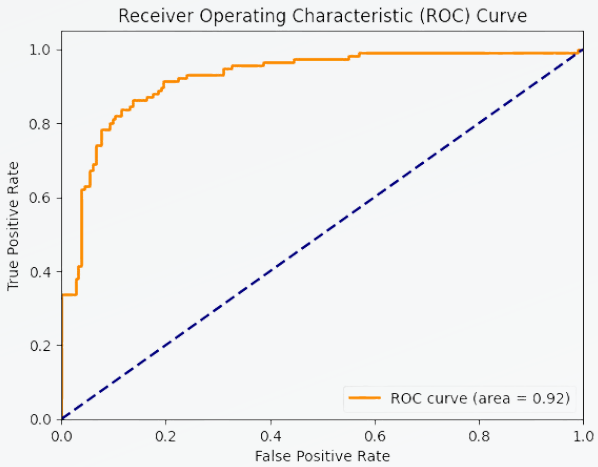


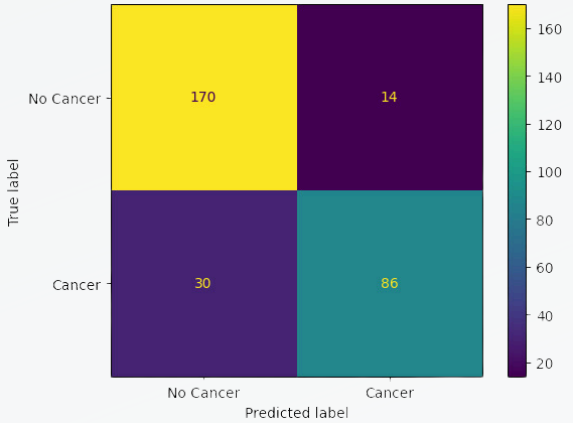


	Precision	Recall	F1	Support
0	0.85	0.93	0.89	184
1	0.87	0.75	0.81	116

RESULTS -QUAD. DISCR. ANALYSIS

TRAINING ERROR RATE: 0.153
TEST ERROR RATE: 0.147
ROC AUC: 0.85
ACCURACY: 0.85

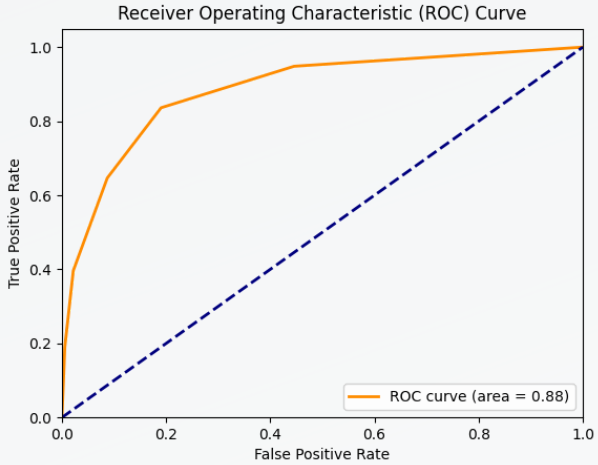


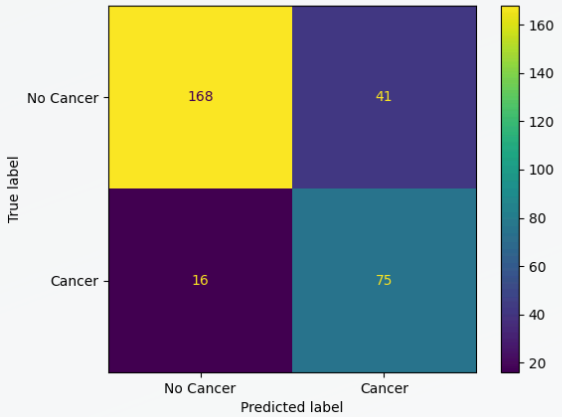


	Precision	Recall	F1	Support
0	0.85	0.92	0.89	184
1	0.86	0.74	0.80	116

RESULTS - KNN

TRAINING ERROR RATE: 0.113
TEST ERROR RATE: 0.190
ROC AUC: 0.89
ACCURACY: 0.81





	Precision	Recall	F1	Support
0	0.80	0.91	0.85	184
1	0.82	0.65	0.72	116

LIMITATIONS & FUTURE WORK

- Logistic regression & LDA produced same results - Why?
- KNN produced a **decreased recall value** in comparison with the other models
- Costs are health & life so error is very crucial

REFERENCES

- [1] API Reference. Scikit-Learn, scikit-learn.org/stable/api/index.html.
- [2] Google Developers. Classification: Precision and Recall | Machine Learning Crash Course | Google Developers. Google Developers, 5 Mar. 2019, developers.google.com/machine-learning/crash-course/classification/precision-and-recall.
- [3] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning. Springer Nature.
- [4] Rabie El Kharoua. “Cancer Prediction Dataset.” Kaggle.com, 2024, www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-datasetresource=download. Accessed 14 July 2024.
- [5] Yeil Kwon. Chap 4. Classification [PDF document].