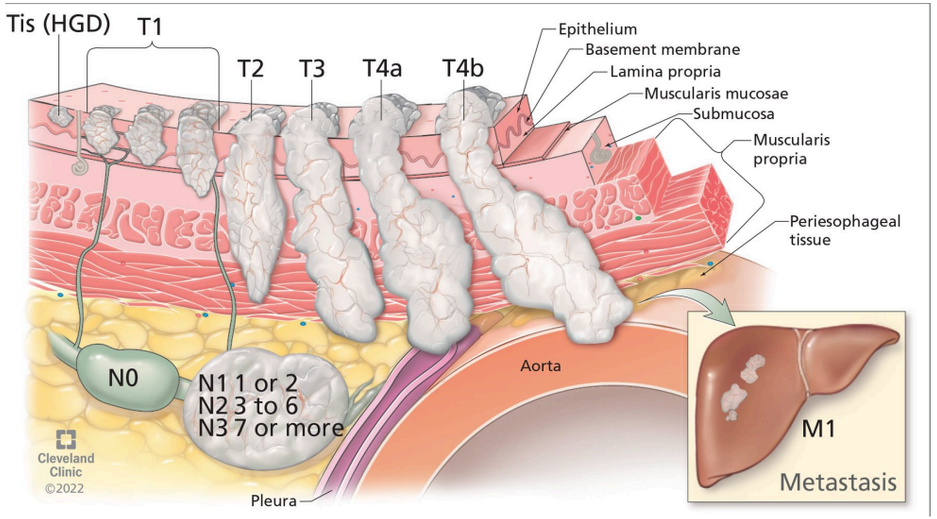
Esophageal cancer and machine learning

Ivor Rubio Tamplin, Jan 2025
Data Science, IA & Machine Learning Bootcamp

Introduction

"Esophageal cancer is a type of cancer that occurs when the cells lining the esophagus mutate and grow uncontrollably. In some cases, the cancer can grow through the esophageal lining and penetrate the esophageal wall."

MD Anderson Cancer Center



The tumor, node, metastasis (TNM) staging system for esophageal cancer. Source: Cleveland Clinic Journal of Medicine

Esophageal cancer is the **sixth most-common cause of cancer-related deaths**. Its prognosis is quite poor because of a lack of initial symptoms. This means the cancer is **usually detected when the disease has already progressed**, offering on average a 15% five-year survival rate.

Types of esophageal cancer

"Esophageal cancer can start anywhere in the esophagus. Your type of esophageal cancer depends on the type of cell that it starts in."

Cancer Research UK

Two types of esophageal cancer are mainly considered:

- Esophageal **squamous-cell carcinoma** (SCC) comprises close to 70% of all cases worldwide and is more common in the developing world, especially south-east Asia.
- Esophageal **adenocarcinoma** (ADC) affects the glandular cells lining the esophagus, is more common in the developed world and its incidence is rising very quickly.

Although other types of cancer can present in the esophagus, they do so very rarely.

Dataset

Mr. Abhinaba Biswas, Mr. Akash Nath and Ms. Shreya Dutta from the JIS College of Engineering in Kalyani (India), have made available a **comprehensive Esophageal Cancer Dataset for AI-Driven Early Detection & Research on Kaggle** to support the global fight against this disease**.**

The dataset comprises 85 columns with information regarding **patient demographics, medical and clinical history, cancer-specific data and clinical outcome data**.

There are, however, many null values in the dataset. Hence, it **requires extensive pre-processing**.

Pre-processing

· More than half the columns were deleted, either because the information they contain is irrelevant to modelling or because the **proportion of null values is high**.

· Null values in categorical columns (YES/NO) were **substituted by using the back-fill and forward-fill methods** alternately.

· Null values in weight and height were substituted with the **mean value according to gender**.

· Certain columns containing similar information, such as race/ethnicity, were **combined by creating a new class** in one column with information regarding the other.

· Null values regarding tumor staging were **calculated according to the TNM values provided** in each case, based on information found online in this regard.

· Numerical values (age, weight, height) were **normalized to (0,1)**.

Question

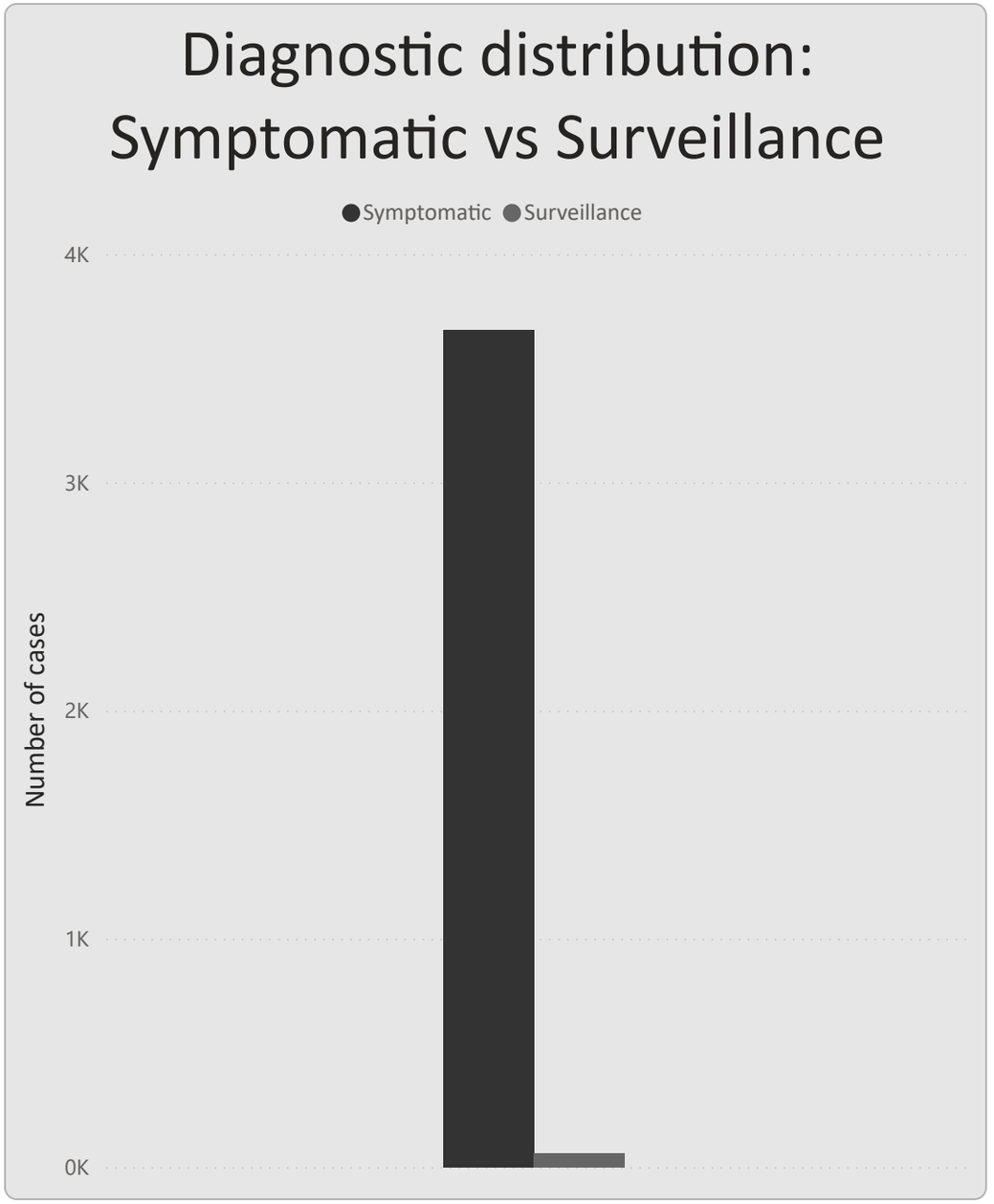
"The clinical, histologic, and oncologic differences between SCC and ADC justify a differentiated therapeutic concept for these two tumor entities and distinct consideration in clinical reports."

National Library of Medicine

Given the extensive nature of the dataset, it provides **many possible approaches** to studying esophageal cancer.

In order to help with the earliest possible setting of treatment, I would like to create a machine learning model that helps with **tumor classification based on demographic, medical and behavioural history information** prior to receiving the histological biopsy report, and comparing the accuracy of this model with one that also includes further information obtained through scanning (affected lymph node count).

But first, let us see what the online literature generally says about cancer type and other variables, and how the dataset reflects this.



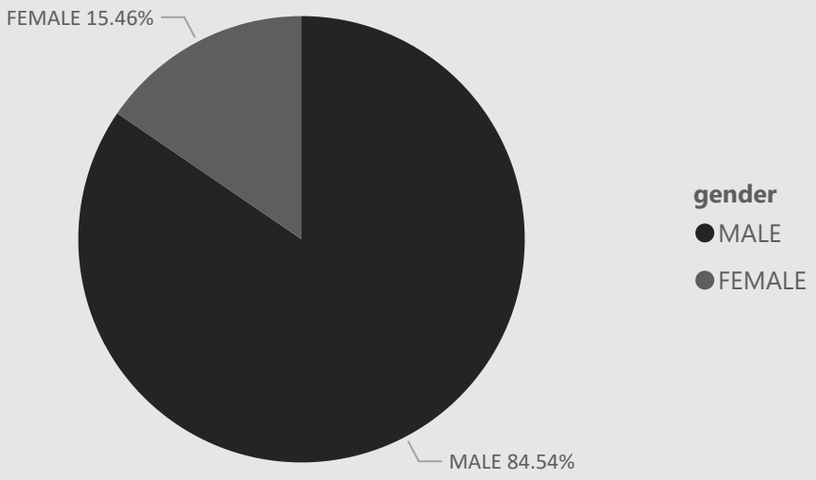
"Most people with esophageal cancer are diagnosed because they have symptoms. It's rare for people without symptoms to be diagnosed with this cancer. When it does happen, the cancer is usually found by accident because of tests done for other medical problems.

Unfortunately, most esophageal cancers do not cause symptoms until they have reached an advanced stage, when they are harder to treat."

Cancer.org

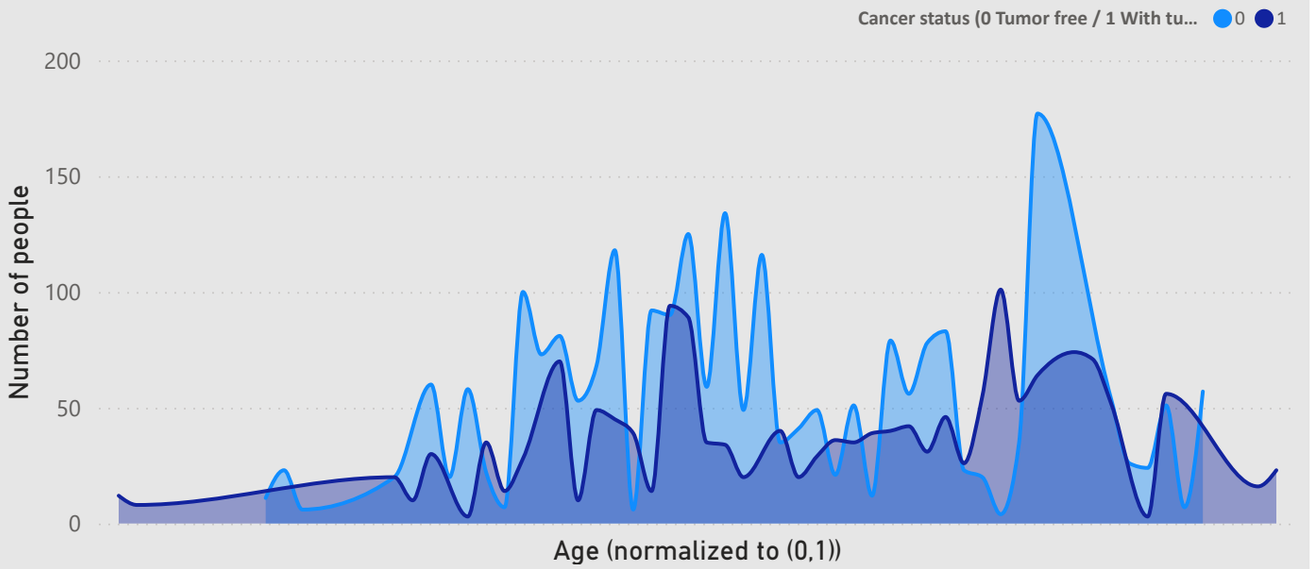
The dataset **clearly reflects the symptomatic nature** of diagnosis.

Gender distribution



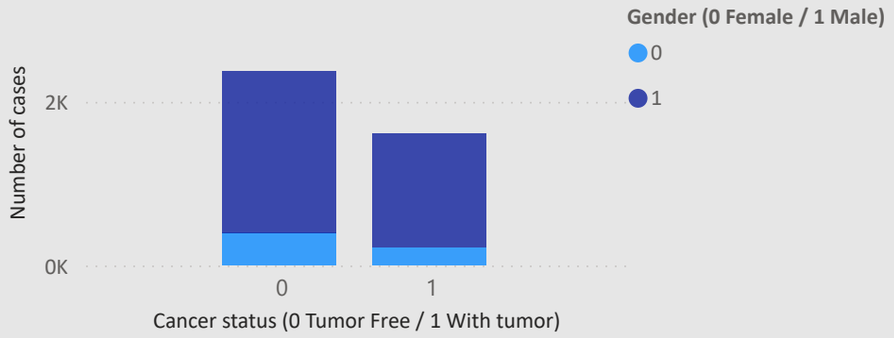
Age distribution

according to cancer status



Gender distribution

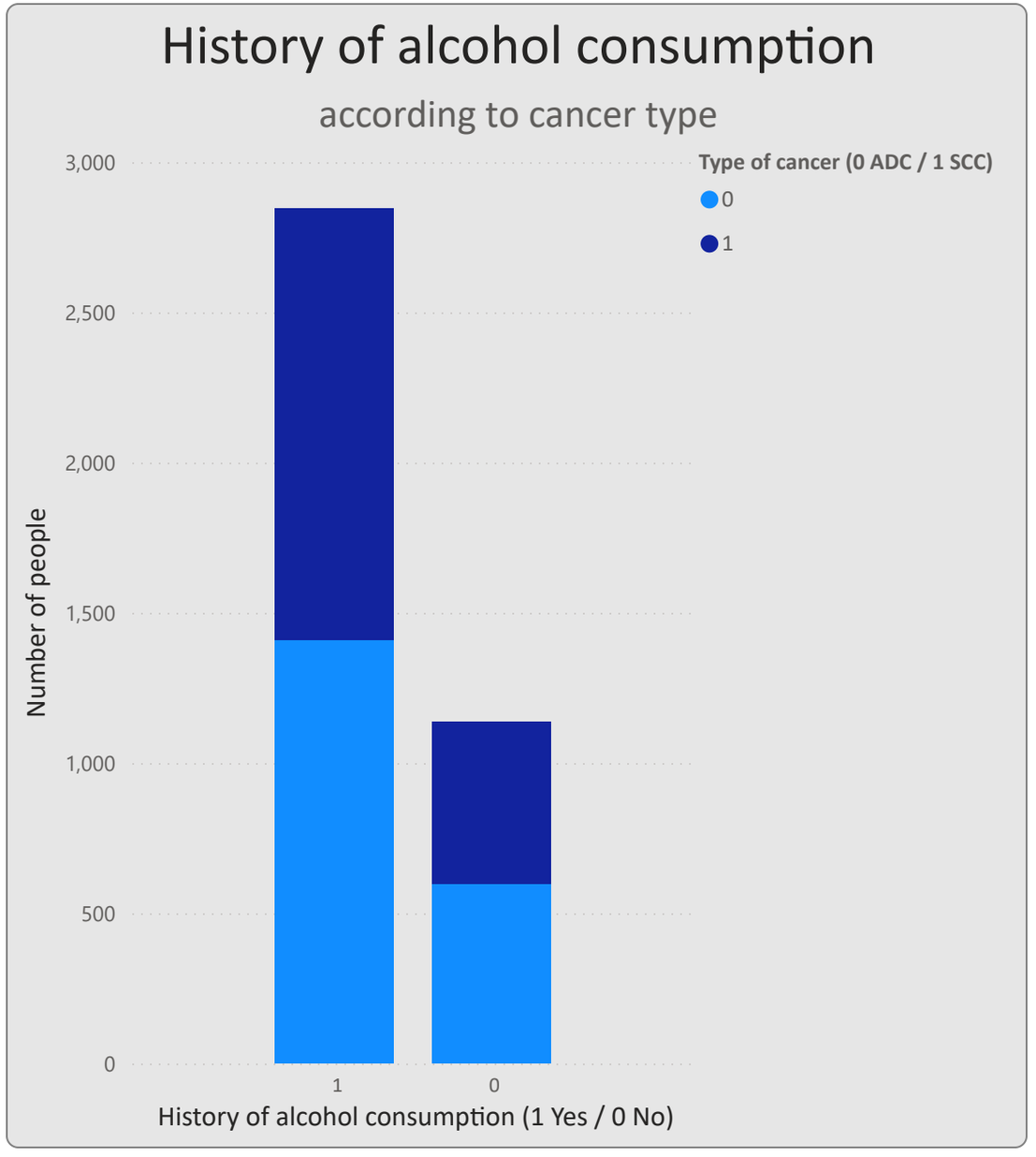
according to cancer status

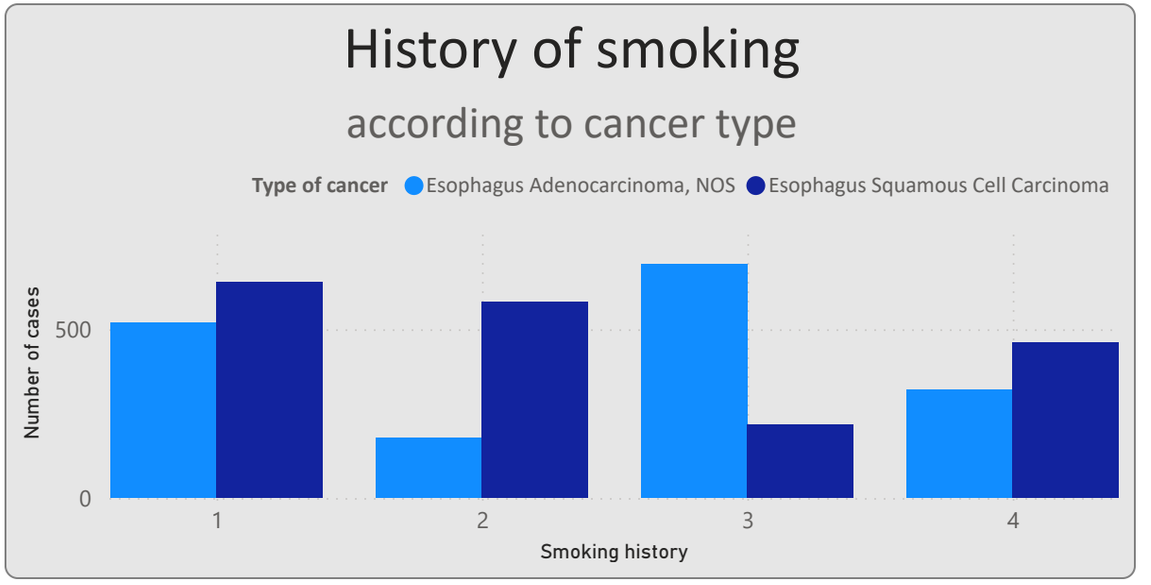


Most cases of esophageal cancer are found in people aged over 55 and men are three times more likely than women to develop esophageal cancer.

MD Anderson Cancer Center

The **dataset reflects the gender bias well** and is similarly balanced in tumor-free cases. However, though it reflects higher proportions of cancer in middle-to-old age (data has been normalized to (0,1)), the proportion of younger people with cancer is higher than would be expected.

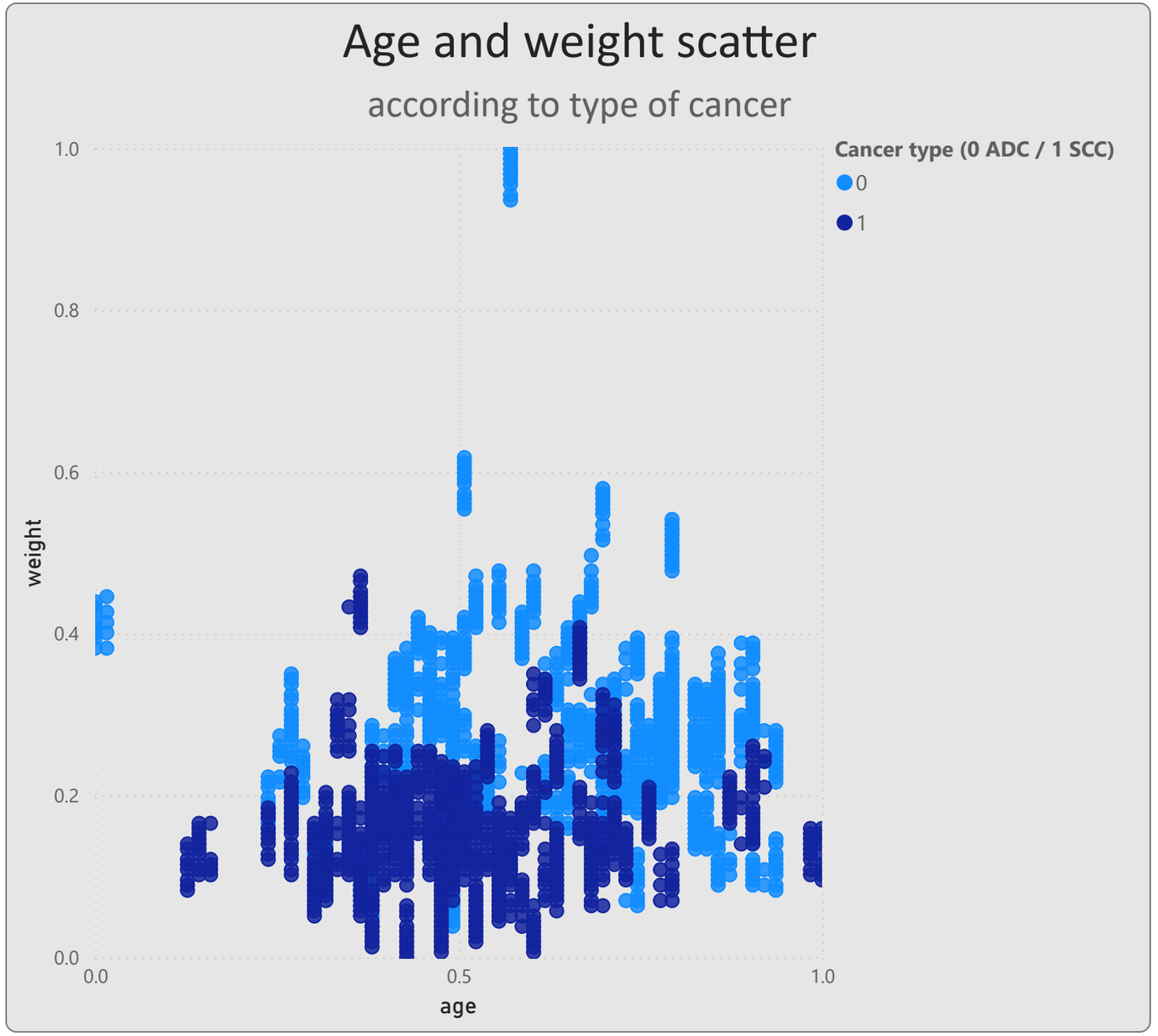




"Squamous cell carcinoma of the esophagus is strongly linked with smoking and drinking too much alcohol."

Memorial Sloan Kettering Cancer Center

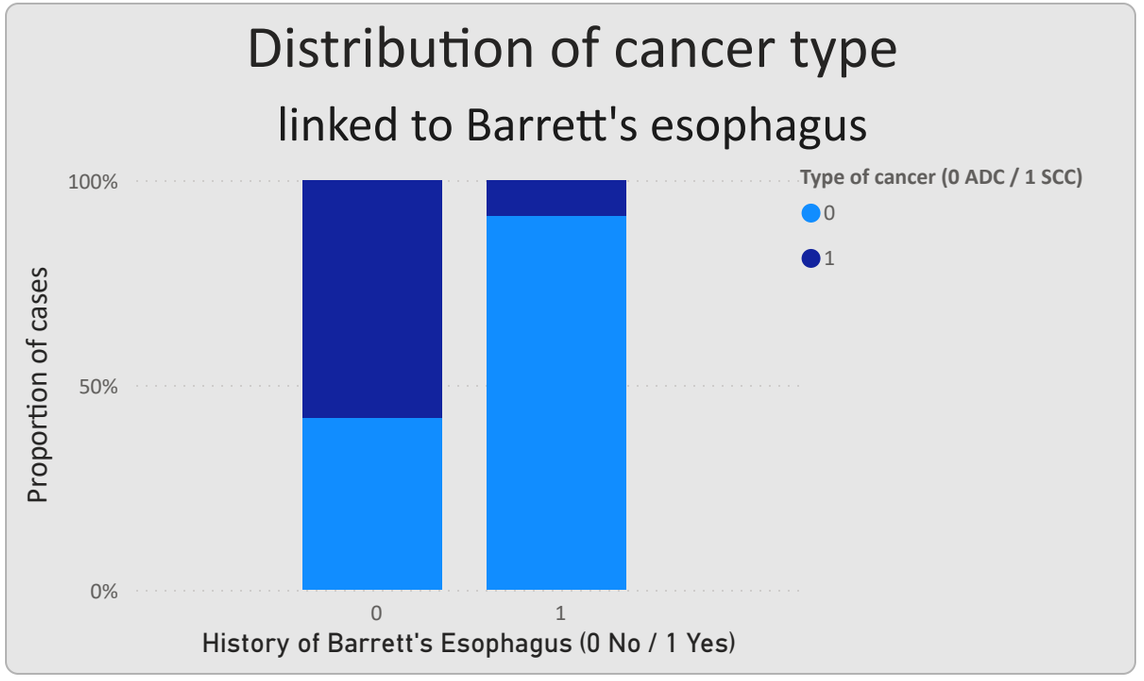
The dataset does not clearly reflect this in general, though perhaps **a deeper understanding of the level of alcohol consumption is required**. There are, however, too many null values in the dataset in this regard.

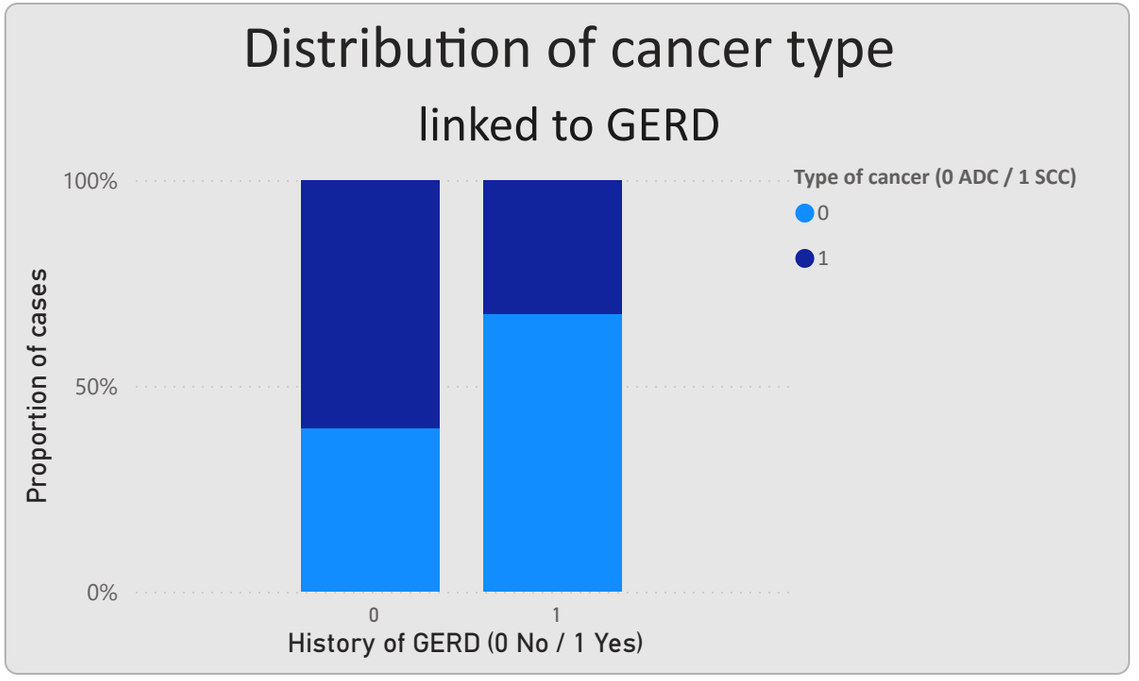


"Adenocarcinoma of the esophagus occurs most often in middle-aged, overweight, white men."

Memorial Sloan Kettering Cancer Center

The dataset appears to confirm this, showing also that **cancer type may be amenable to "clustering" by using logistic regression with these variables only**.

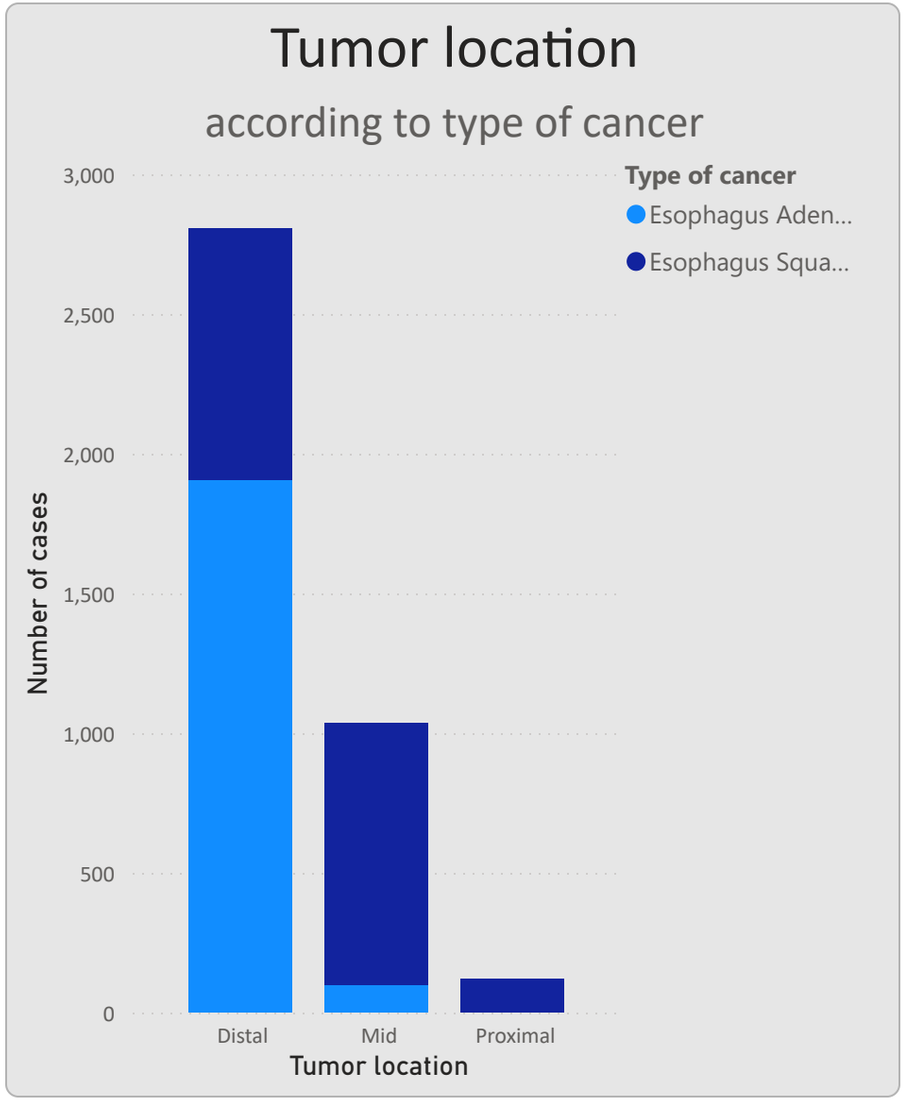




Gastroesophageal reflux disease (GERD) is a chronic condition that involves a backflow of stomach contents into the esophagus. [...] it can advance to Barrett's esophagus which is considered a precancerous lesion and increases the risk for esophageal cancer, more commonly adenocarcinoma.

Memorial Sloan Kettering Cancer Center

The dataset reflects the presence of **adenocarcinoma more commonly among patients who have previously suffered from Barrett's esophagus and GERD**.



Overall, the literature points out that, although this isn't necessarily the case always, **adenocarcinoma** generally tends to appear **in the lower section** of the esophagus while **squamous cell carcinoma** tends to appear **in the mid and upper sections**.

The dataset adequately reflects the described incidence of different cancer types according to where the tumor is located.

Machine Learning

20 columns will be fed to different ML models, containing data regarding:

- **Demographic** information:
  - sex
  - age
  - weight
  - race
- Relevant **medical history**:
  - a history of cancer
  - smoking habits
  - alcohol consumption
  - tumor location

3 models were trained and tested

- **KNN**
- **XGBoost** (with GridSearch cross-validation)
- **CatBoost** (with default hyperparameters and no cross-validation)

- Given the significant proportion of **categorical variables** in the clean dataset (with **dummy variables** created for modelling), the decision was made to try out CatBoost and compare the results of this model with the results provided by XGBoost.
- Accuracy figures for XGBoost show **some overfitting** in the model, with 0.995 in training vs. 0.987 in testing.
- The **results were excellent** for all models, with **accuracy levels ranging between 0.987** for XGBoost **and 1.0** for CatBoost and KNN (with similar best results for values of k=2 and k=4).

Conclusions

Insights from previous studies broadly cluster esophageal cancer types according to certain demographic characteristics and behaviours. This may be the reason behind the high accuracy levels displayed by all the models trained.

Does this render an AI to predict cancer type prior to histology useless?

To answer this, the **dataset could perhaps include the physician's preliminary diagnosis** based on the limited set of variables they first encounter (demographic and behavioural data, and endoscopy results), and whether they made the right diagnosis (confirmed or not by the histology report).

Also, whether or not such a model is impactful will ultimately depend on **whether such information can change the patient's outcome significantly by initiating treatment earlier**. This is currently beyond my scope of medical understanding.

Only by knowing all of this can we **understand whether the medical profession needs a solution of the sort**.

In regard to this study, the **dataset lacks information regarding any food and beverage intake besides alcohol**. For instance, some literature already hints to a connection between the disease and drinking hot beverages regularly.

Such information should be considered in future datasets as an **input that very likely affects the area linked to this disease and could, potentially, be linked to cancer type** (especially in regard to pesticides, hormones and other natural and synthetic chemicals involved in the agriculture and food industry).

Though the **initial question posited a comparison** with a model that used further information obtained from other radiological testing (lymph node count), the accuracy **results for the initial models seem to render this comparison useless**.

However, a comparison was carried out using XGBoost (which provided the worst results overall). As was to be expected, **accuracy results improved slightly with additional information in the model**, increasing from 0.987 to 0.993, still below the results for the limited KNN and CatBoost models.

Lastly, by not initially approaching the model with a question in mind, the pre-processing function created in the course of this study provides **a cleaner all-purpose data file that can be used by anyone** for a variety of further research and models.

References

- Abhinaba Biswas, Akash Nath, Shreya Dutta. *Esophageal Cancer Dataset*. Extracted from: https://www.kaggle.com/datasets/abhinaba1biswas/esophageal-cancer-dataset (accessed Dec 2024)
- Abel Joseph, MD, Siva Raja, MD, PhD, Suneel Kamath, MD, Sunguk Jang, MD, Daniela Allende, MD, Mike McNamara, MD, Gregory Videtic, MD, Sudish Murthy, MD and Amit Bhatt, MD). *Esophageal adenocarcinoma: A dire need for early detection and treatment* - Cleveland Clinic Journal of Medicine, May 2022 89 (5) 269-279; DOI: https://doi.org/10.3949/ccjm.89a.21053 (accessed Dec 2024)
- American Cancer Society. *Signs and Symptoms of Esophageal Cancer*. https://www.cancer.org/cancer/types/esophagus-cancer/detection-diagnosis-staging/signs-and-symptoms.html (accessed Dec 2024)
- Memorial Sloan Kettering Cancer Center. *Types of esophageal cancer*. https://www.mskcc.org/cancer-care/types/esophageal/types-esophageal (accessed Dec 2024)
- J. Encinas de la Iglesia,, M.A. Corral de la Calle, G.C. Fernández Pérez, R. Ruano Pérez, A. Álvarez Delgado. *Cáncer de esófago: particularidades anatómicas, estadificación y técnicas de imagen.* DOI: 10.1016/j.rx.2016.06.004. https://www.elsevier.es/es-revista-radiologia-119-articulo-cancer-esofago-particularidades-anatomicas-estadificacion-S0033833816300741 (accessed Jan 2025)
- C. Mariette, L. Finzi, G. Piessen, I. Van Seuningen, J.P. Triboulet. *Esophageal carcinoma: prognostic differences between squamous cell carcinoma and adenocarcinoma.* National Library of Medicine. DOI: 10.1007/s00268-004-7542-x. https://pubmed.ncbi.nlm.nih.gov/15599738/ (accessed Jan 2025)
- Cancer Research UK. *About the stages, types and grades of oesophageal cancer.* https://www.cancerresearchuk.org/about-cancer/oesophageal-cancer/stages-types-and-grades/about (accessed Dec 2024)

Thank you.