



DSA210- Introduction to Data Science Final Report

Project Title

Data Driven Gene Expression Analysis Related to Tobacco and Alcohol Risk

Course Instructor

Özgür Asar

Selim Balcısoy

Assigned Teaching Assistant

Kerem Aydın

Submitted by

Arda Taşkoparan – 32512

Sabancı University

Istanbul, 2025

INTRODUCTION

Lung cancer remains one of the most lethal cancer types worldwide, with gene expression alterations playing a pivotal role in its diagnosis and treatment. In particular, genes like KRAS, EGFR, and ALK have been widely studied in relation to environmental and lifestyle factors such as tobacco and alcohol use (Tam et al., 2006; Soo et al., 2017). This project investigates the relationship between these genes and behavioral risk factors through exploratory data analysis (EDA), hypothesis testing, and machine learning (ML) techniques.

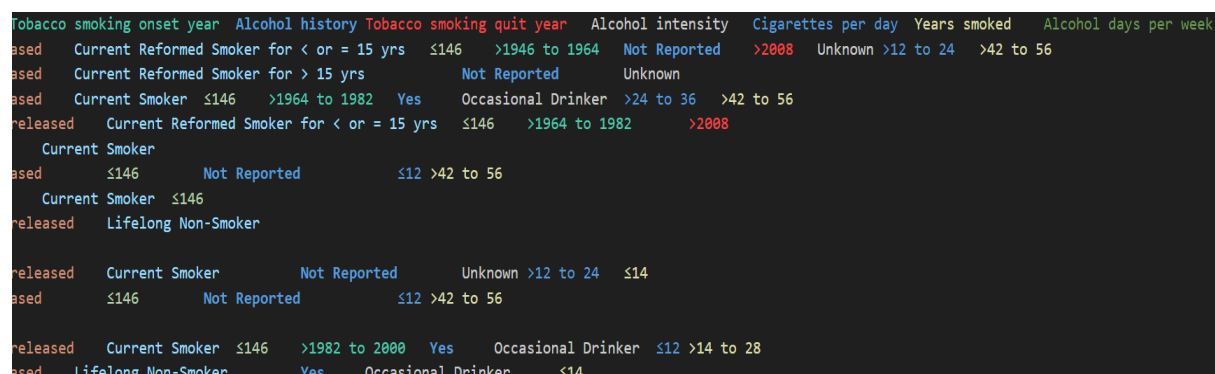
PART 1: DATA COLLECTION and PREPROCESSING

Data Source

The dataset was manually retrieved from the National Cancer Institute GDC Data Portal with selecting “lung” as the cohort. After filtering and sorting patient data through the site interface, the clinical and genomic data were exported in “.tsv” format

Cleaning and Standardization

To ensure consistency in all the data, all placeholders for missing or unknown values (e.g., “Not Reported”, “Unknown”) were replaced with NaN. Multiple non-informative or fully missing columns (like “Alcohol days per week”) were dropped to improve analysis quality. **Figure 1** contains some of dropped columns.



```
Tobacco smoking onset year Alcohol history Tobacco smoking quit year Alcohol intensity Cigarettes per day Years smoked Alcohol days per week
ased Current Reformed Smoker for < or = 15 yrs ≤146 >1946 to 1964 Not Reported >2008 Unknown >12 to 24 >42 to 56
ased Current Reformed Smoker for > 15 yrs Not Reported Unknown
ased Current Smoker ≤146 >1964 to 1982 Yes Occasional Drinker >24 to 36 >42 to 56
released Current Reformed Smoker for < or = 15 yrs ≤146 >1964 to 1982 >2008
Current Smoker
ased ≤146 Not Reported ≤12 >42 to 56
Current Smoker ≤146
released Lifelong Non-Smoker
released Current Smoker Not Reported Unknown >12 to 24 ≤14
ased ≤146 Not Reported ≤12 >42 to 56
released Current Smoker ≤146 >1982 to 2000 Yes Occasional Drinker ≤12 >14 to 28
ased Lifelong Non-Smoker Yes Occasional Drinker ≤14
```

Figure 1) Some of the categories that were dropped.

Mapping, Risk Scoring and Data Transformation

For EDA and ML, I have transformed the data into these two key risk scores:

- **Tobacco Risk Score:** Based on smoking status, cigarettes per day and years smoked.
- **Alcohol Risk Score:** Based on alcohol history and intensity.

These were computed by mapping categorical values into numeric scores and summing the dimensions, excluding ambiguous cases like “Current Reformed Smoker, Duration Not Specified”.

For Tobacco Risk Score, four categories were ranked based on risk exposure: lifelong non-smokers were assigned a score of 0, reformed smokers who quit over 15 years ago were given 1, those who quit more recently were assigned 2, and current smokers were rated highest at 3. Additional smoking-related variables were also mapped numerically: cigarette consumption per day ranged from 0 (≤ 12) to 4 (> 48), and years smoked followed a similar scale from 0 (≤ 14 years) to 4 (> 56 years). For Alcohol Risk Score, a binary mapping was applied to drinking history (0 for “No”, 1 for “Yes”), and alcohol intensity was categorized from 0 to 2, with lifelong non-drinkers at the lowest risk and occasional drinkers at the highest. These mappings were then used to generate composite risk scores for both tobacco and alcohol exposure.

PART 2: EXPLORATORY DATA ANALYSIS (EDA)

Boxplot Analysis

I have conducted separate boxplot analyses to compare EGFR, KRAS, and ALK expression levels across tobacco and alcohol risk scores.

EGFR vs. Tobacco Risk

EGFR showed modest variance across risk levels. Outliers and slightly higher expression in mid-risk groups suggest a non-linear trend seen in **Figure 2**.

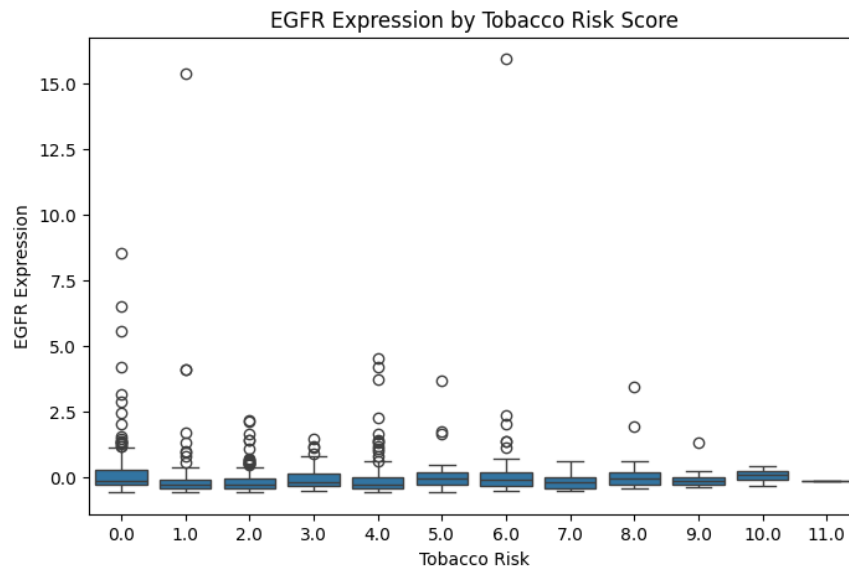


Figure 2) Boxplot – EGFR vs. Tobacco Risk.

KRAS vs. Tobacco Risk

KRAS displayed a broader spread across all tobacco risk levels, indicating a stronger potential correlation seen in **Figure 3**.

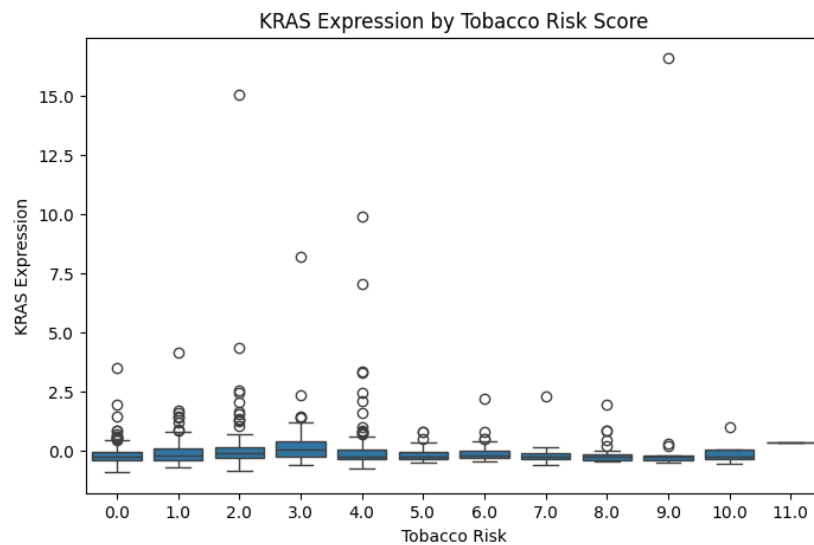


Figure 3) Boxplot – KRAS vs. Tobacco Risk.

ALK vs. Tobacco Risk

ALK was highly variable in low-risk groups, with many outliers. This instability might affect predictive modelling seen in **Figure 4**.

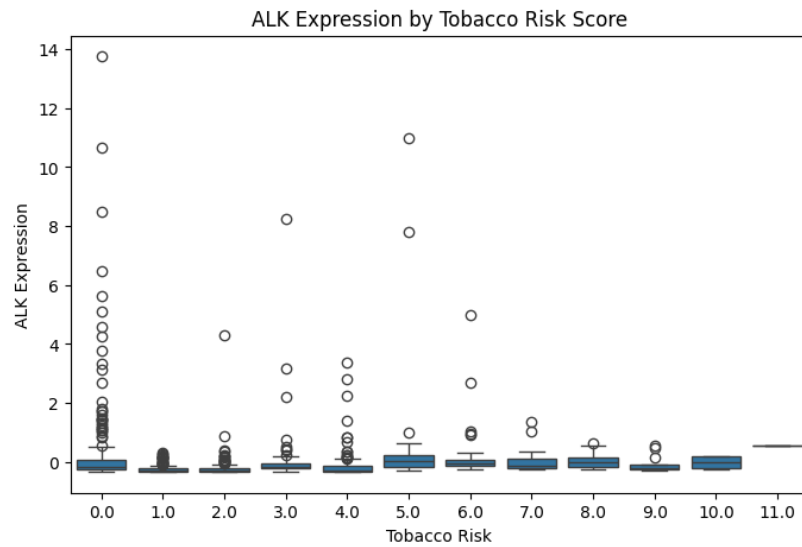


Figure 4) Boxplot – ALK vs. Tobacco Risk.

Gene Expressions vs. Alcohol Risk

Patterns here were much less conclusive. EGFR and ALK showed moderate variation, but the trends were not as strong as tobacco risk as can be seen in grouped boxplot as in **Figure 5**.

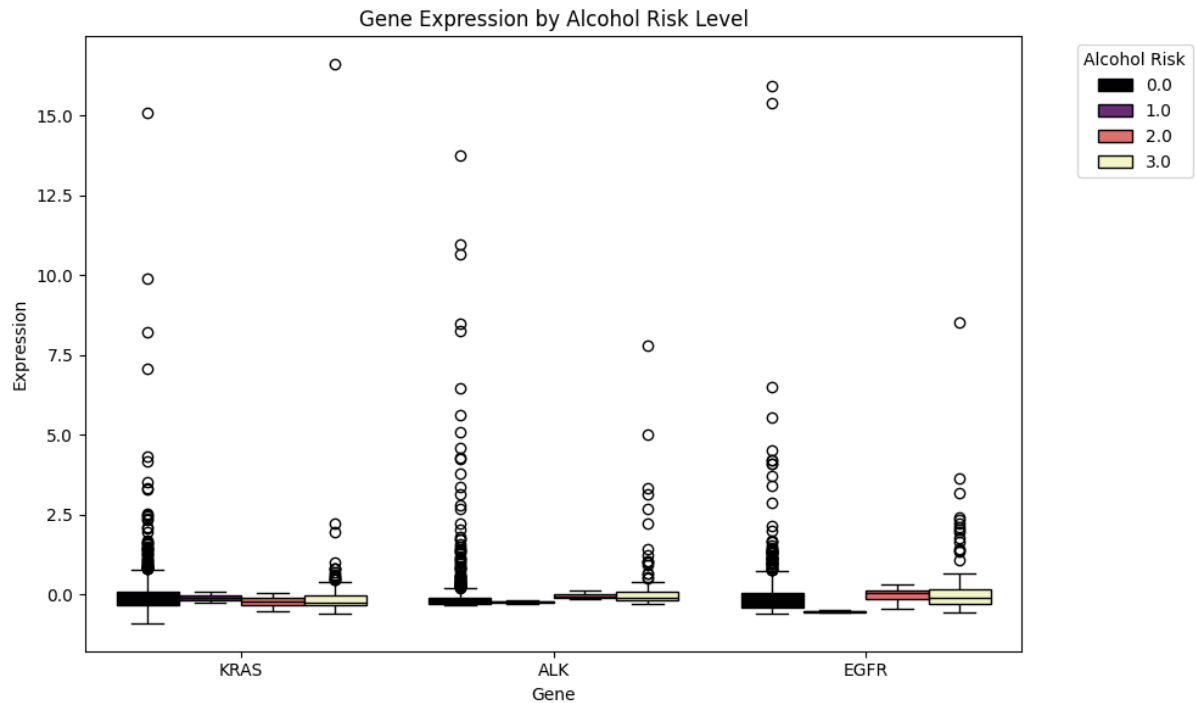


Figure 5) Grouped Boxplot – EGFR/KRAS/ALK vs. Alcohol Risk.

Analyzing EDA Results

The exploratory analysis revealed that gene expression values for EGFR, KRAS, and ALK were generally clustered around zero across most tobacco and alcohol risk groups, suggesting subtle differences or a high signal-to-noise ratio. Boxplots showed that EGFR had slightly wider dispersion in moderate tobacco risk levels, supporting previous findings that link environmental tobacco exposure to EGFR expression changes (Soo et al., 2017). KRAS displayed broader variation across tobacco risk categories, particularly in mid to high-risk levels, aligning with research indicating its stronger association with heavy smoking (Tam et al., 2006). In contrast, ALK showed notable outliers in low-risk groups, reflecting findings that ALK alterations may occur independently of tobacco exposure.

Grouped boxplot for alcohol risk suggested mild variability in EGFR and ALK, though overall patterns were weaker compared to tobacco. Additional visualizations, including grouped boxplots, pairplots and barplots were created to improve clarity and interpretation. These are available in the Jupyter notebook.

PART 3: HYPOTHESIS TESTING

Constructing Hypothesis

Null Hypothesis (H_0): Tobacco and alcohol usage has no significant effect on the expression levels of lung cancer-related genes KRAS, EGFR, and ALK.

Alternative Hypothesis (H_1): Tobacco and alcohol usage significantly influences the expression of cancer-related genes KRAS, EGFR, and ALK.

I have performed ANOVA and Spearman tests to reject or fail to reject null hypothesis.

Table for ANOVA Results

Comparison	F-Statistic	p-Value
EGFR vs Tobacco Risk	2.303	0.0113
EGFR vs Alcohol Risk	1.189	0.3129
KRAS vs Tobacco Risk	3.229	0.0004
KRAS vs Alcohol Risk	0.119	0.9490

ALK vs Tobacco Risk	5.342	0.00
ALK vs Alcohol Risk	1.253	0.2893

Table for Spearman Results

Gene	Tobacco (r, p)	Alcohol (r, p)
EGFR	-0.043, 0.1792	0.125, 0.0001
KRAS	0.085, 0.0069	-0.033, 0.2926
ALK	0.086, 0.0062	0.328, 0.0000

Analyzing Hypothesis Testing Results

The ANOVA results indicate that tobacco risk is significantly associated with expression levels of EGFR, KRAS, and ALK, while alcohol risk does not show a significant group-level effect. These findings are supported by the Spearman correlation analysis, which shows weak but statistically significant positive correlations between tobacco risk and KRAS ($r = 0.085$, $p = 0.0069$) and ALK ($r = 0.086$, $p = 0.0062$). Although EGFR had a significant ANOVA result, its correlation with tobacco risk was not significant ($r = -0.043$, $p = 0.1792$), suggesting non-linear or group-based differences.

On the other hand, ALK showed a notable correlation with alcohol risk ($r = 0.328$, $p < 0.0001$), even though the ANOVA did not detect significant differences which may be hinting at a possible continuous relationship rather than distinct group effects.

In summary, tobacco risk demonstrated more consistent statistical associations with gene expression across both tests, justifying its focus in downstream modeling.

PART 4: MACHINE LEARNING (ML) ANALYSIS

I have conducted All machine learning experiments were conducted within the same Jupyter Notebook used during the EDA phase. The objective was to evaluate whether the engineered tobacco and alcohol risk scores could be used to predict gene expression levels, or vice versa, through classification and regression models.

Binary Classification of Tobacco Risk

In the binary classification task, gene expression values were used to predict whether a patient belonged to a low or high tobacco risk group. Among all models, the Random Forest classifier performed best on ALK, achieving an accuracy of 74%, while Logistic Regression reached 66%. For EGFR, Random Forest yielded 65% accuracy, demonstrating relatively balanced performance across classes. In contrast, KRAS performed poorly in both models, with accuracy scores around 49–43%, reinforcing the weaker associations observed during EDA.

The feature importance plots from the Random Forest classifiers which is presented in **Figure 6 to 8**, highlight that tobacco risk was generally more influential than alcohol risk, especially for ALK and KRAS.

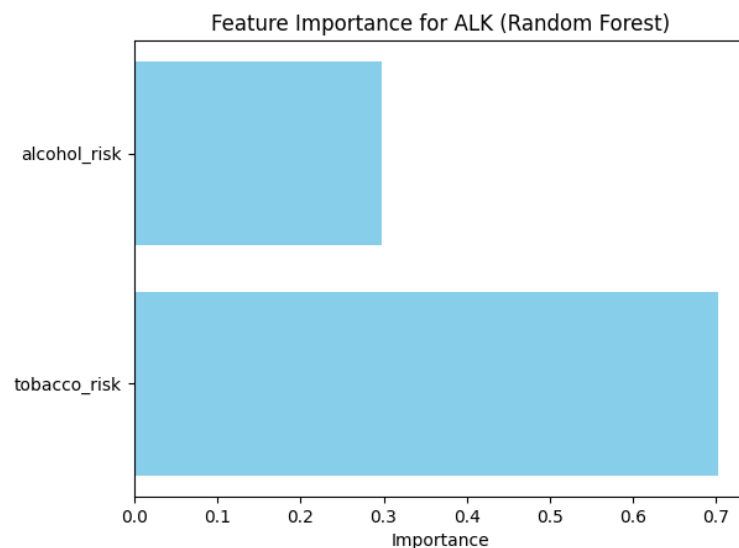


Figure 6) Feature importance for ALK.

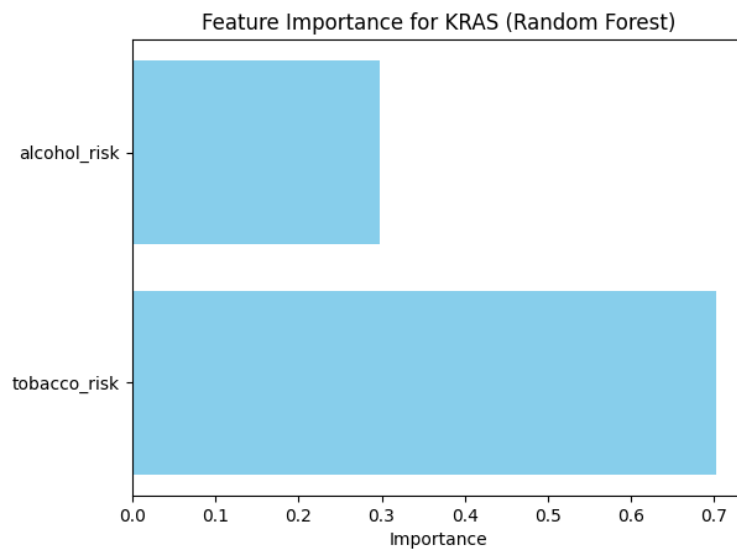


Figure 7) Feature importance for KRAS

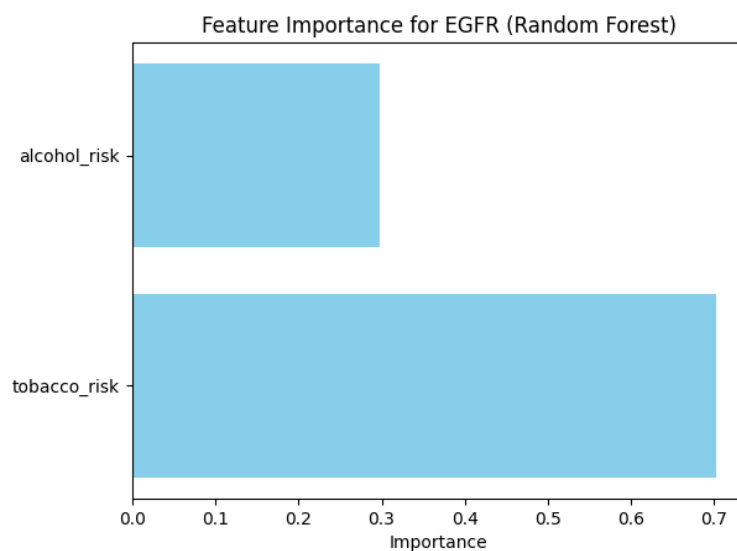


Figure 8) Feature importance for EGFR.

Multiclass Classification of Tobacco Risk

In this task, a Random Forest classifier was trained to predict one of three tobacco risk levels—Low, Medium, or High—using gene expression data. The model achieved an overall accuracy of 62% (with `random_state=45`). It showed high performance for the Low-risk group (precision = 0.67, recall = 0.87), but failed to correctly classify any instances in the High-risk group, with both precision and recall at 0.00. These imbalances are likely due to

class distribution skew. The confusion matrix and gene importance plot for this model are presented in **Figure 9 and 10**.

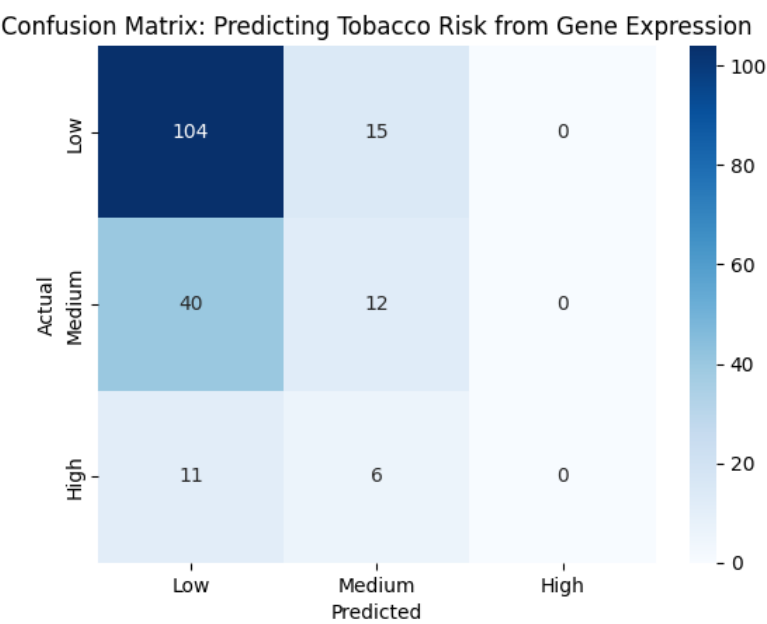


Figure 9) Confusion Matrix for Predicting Tobacco Risk from Gene Expression.

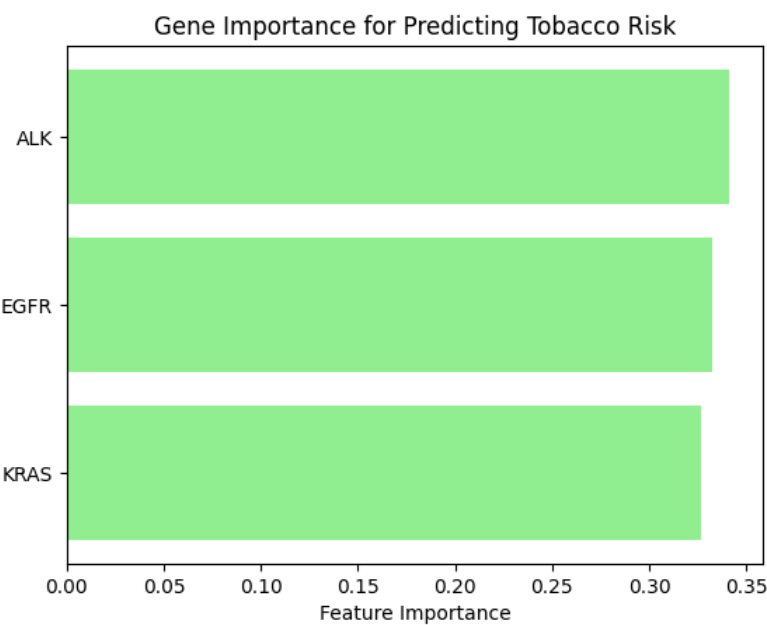


Figure 10) Gene Importance Plot for Predicting Tobacco Risk.

Regression on Gene Expression

Regression models were used to predict continuous gene expression levels based on the tobacco_risk and alcohol_risk features. Both Linear Regression and Random Forest Regressor were applied to each gene.

For ALK, the Random Forest model achieved an R^2 score of 0.0992, slightly outperforming Linear Regression ($R^2 = 0.0009$). EGFR models also performed poorly, with R^2 values between 0.01 and 0.05. The most underwhelming results came from KRAS, where both models returned negative R^2 scores, indicating performance worse than predicting the mean. Mean Absolute Errors (MAE) for all genes ranged between 0.21 and 0.35.

Figures 11 to 13 show the scatter plots of actual vs. predicted values for ALK, KRAS, and EGFR. As shown, predictions tend to cluster around the mean, indicating that the models failed to capture meaningful variability.

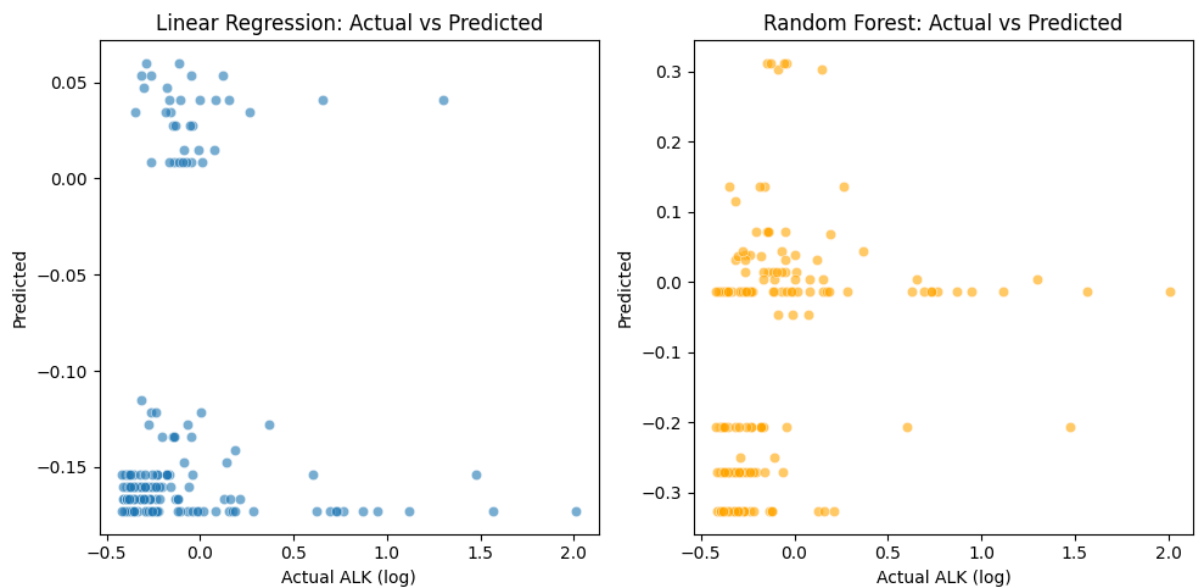


Figure 11) Scatter Plots of Actual vs. Predicted Values for ALK.

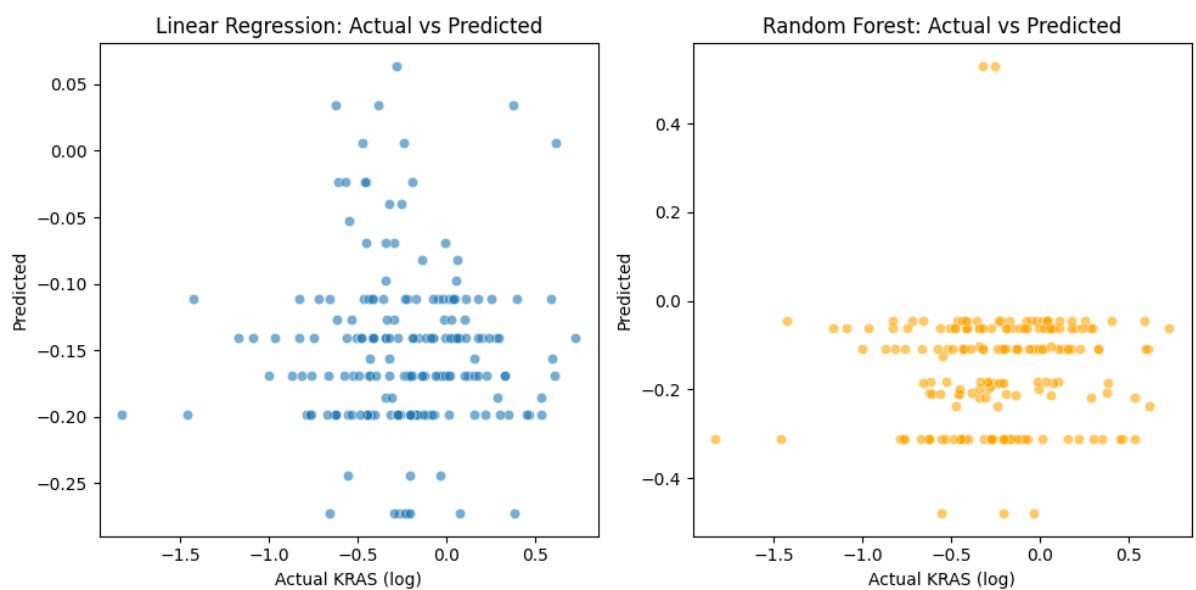


Figure 12) Scatter Plots of Actual vs. Predicted Values for KRAS.

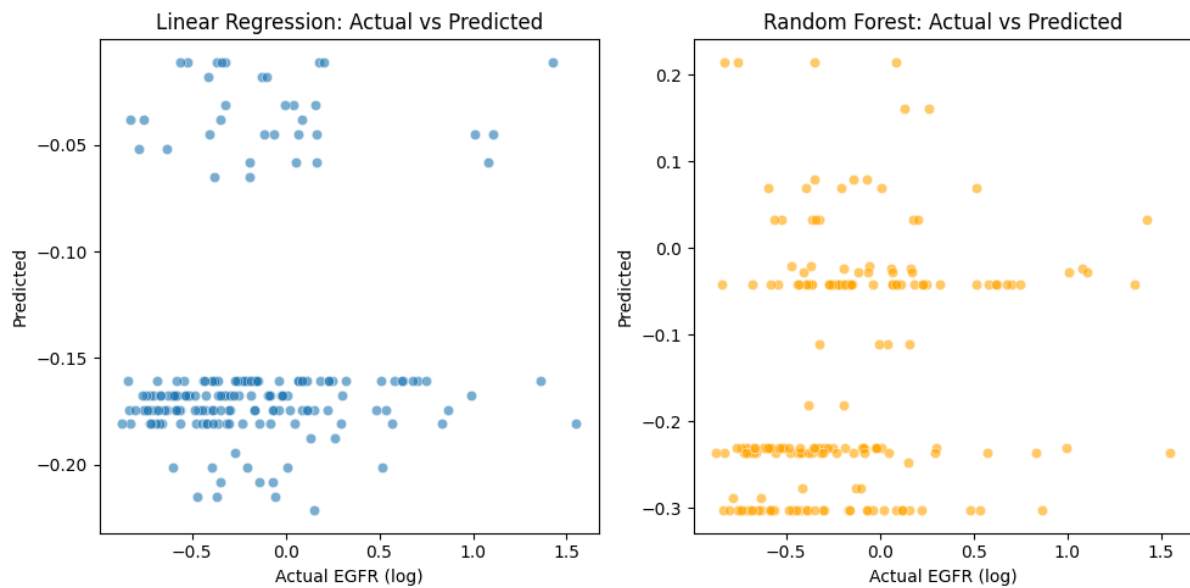


Figure 13) Scatter Plots of Actual vs. Predicted Values for EGFR.

Analysis of Machine Learning Results

Machine learning results yielded mixed outcomes. The Random Forest classifier performed best when classifying tobacco risk from gene expression, particularly for ALK and EGFR while KRAS consistently underperformed across all models. This was somewhat unexpected, as KRAS had shown a stronger correlation with tobacco risk during EDA and hypothesis testing. The discrepancy may be due to factors such as noise in expression values, weak signal strength relative to the limited number of features, or model sensitivity to class boundaries.

Feature importance plots confirmed that tobacco risk was generally more informative than alcohol risk, reinforcing earlier conclusions in EDA. In multiclass classification, model performance was heavily affected by class imbalance, which likely explains its poor handling of medium and high-risk groups. The regression results were weak across all genes, with near-zero or negative R^2 values which may indicate that the chosen features did not capture enough continuous variance to model gene expression with confidence even though they were somewhat meaningful for classification.

CONCLUSION

In this project, I aimed to explore the relationship between tobacco and alcohol usage and the expression levels of key lung cancer-related genes (EGFR, KRAS, and ALK) using data from the National Cancer Institute. Through EDA and hypothesis testing, I have found statistically significant links between tobacco use and gene expression, particularly for ALK and KRAS, whereas alcohol use showed weaker or inconsistent associations.

However, the results from machine learning models highlight key limitations. Predictive performance was moderate at best, and regression models in particular failed to meaningfully capture gene expression levels. These outcomes suggest that gene expression is influenced by a broader set of biological and clinical factors, and that lifestyle scores alone may be too narrow to explain complex expression patterns. To improve analysis, additional clinical or demographic features can be added such as age and gender. Also, the mutation status or genomic biomarkers could be added alongside the gene expression values.

REFERENCES

- National Cancer Institute. (n.d.). *Genomic Data Commons Data Portal*.
<https://portal.gdc.cancer.gov/>
- Tam, I. Y. S., Chung, L. P., Suen, W. S., Wang, E., Wong, M. C., Ho, K. K., ... & Lam, W. K. (2006). Distinct epidermal growth factor receptor and KRAS mutation patterns in non-small cell lung cancer patients with different tobacco exposure and clinicopathologic features. *Clinical Cancer Research*, 12(5), 1647–1653.
<https://doi.org/10.1158/1078-0432.CCR-05-1981>
- Soo, R. A., Lim, S. M., Syn, N. L., Teng, R., Soong, R., Mok, T. S. K., & Cho, B. C. (2017). Association Between Environmental Tobacco Smoke Exposure and the Occurrence of EGFR Mutations and ALK Rearrangements in Never-Smokers With Lung Cancer. *Clinical Lung Cancer*, 18(5), 535–542.
<https://doi.org/10.1016/j.clcc.2017.01.005>
- Wang, Y., Wang, R., Pan, Y., Wang, L., Shen, L., Chen, H., ... & Chen, H. (2014). The prevalence of ALK rearrangements and association with clinicopathologic characteristics in Chinese patients with lung adenocarcinoma. *Lung Cancer*, 83(3), 312–317. <https://doi.org/10.1016/j.lungcan.2013.12.016>
- Uchida, S., & Mizuguchi, Y. (2010). *Gene expression analysis in lung cancer*. In Lung Cancer—Molecular Targets and Therapeutic Uses. IntechOpen.
<https://doi.org/10.5772/13628>