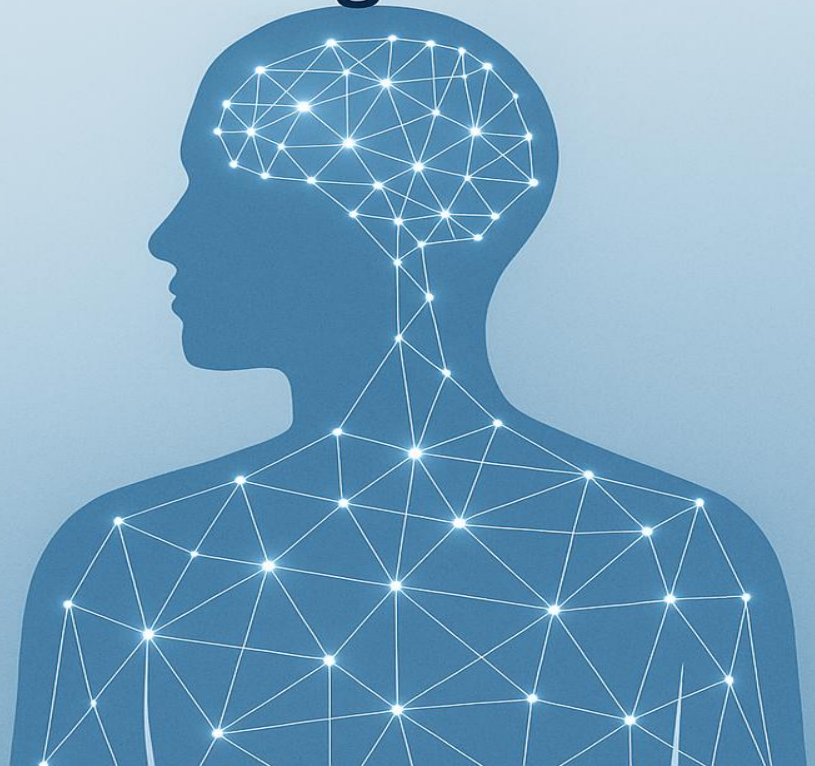
Predictive Modeling and Risk Analysis of Cancer Severity and Survival Outcomes Using Machine Learning



🔍 Business Problem

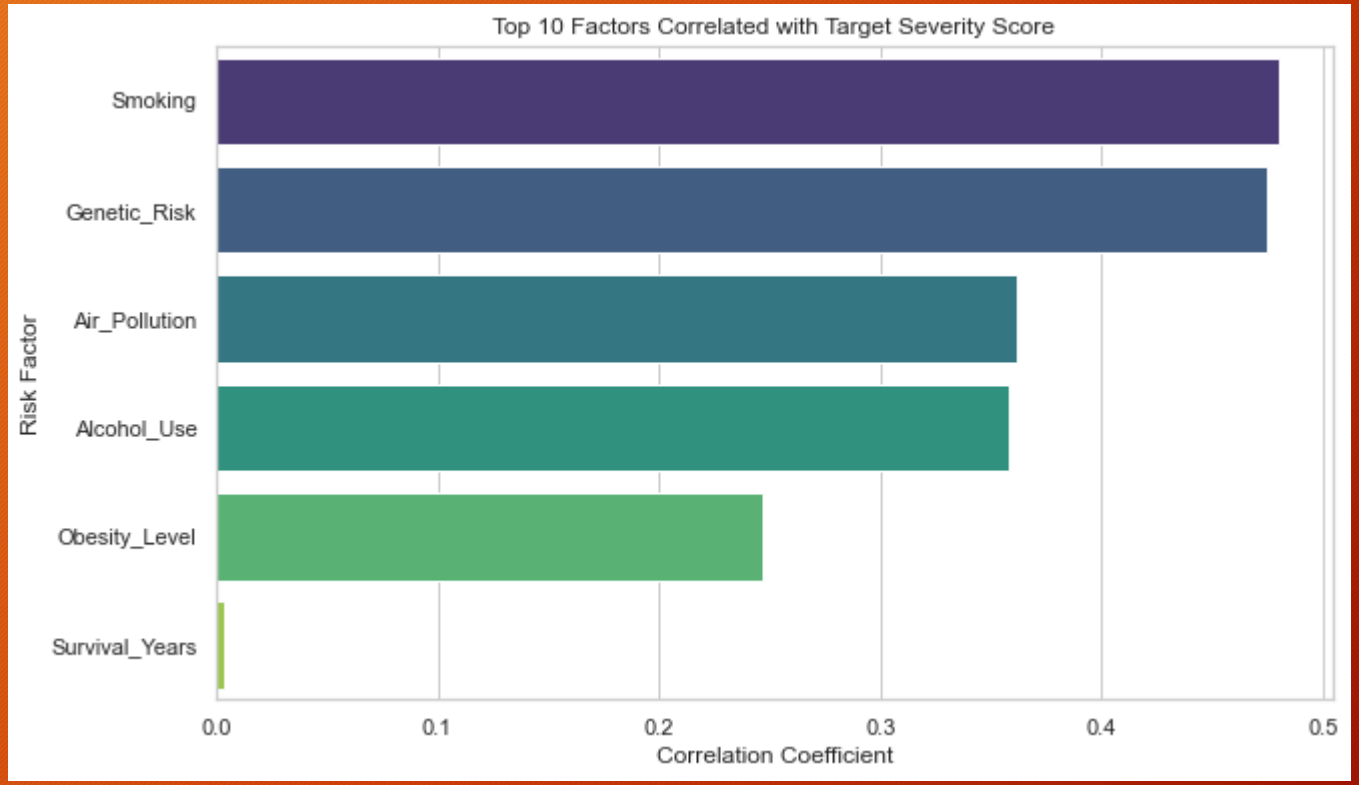
Cancer remains one of the world's deadliest diseases, presenting challenges not only in treatment but also in early prediction and resource allocation. With healthcare systems facing increasing demands, being able to predict cancer severity and survival likelihood allows providers to better allocate resources, prioritize care, and personalize treatment strategies. Our goal is to use data-driven insights to support smarter healthcare decisions.

🧠 Project Objectives

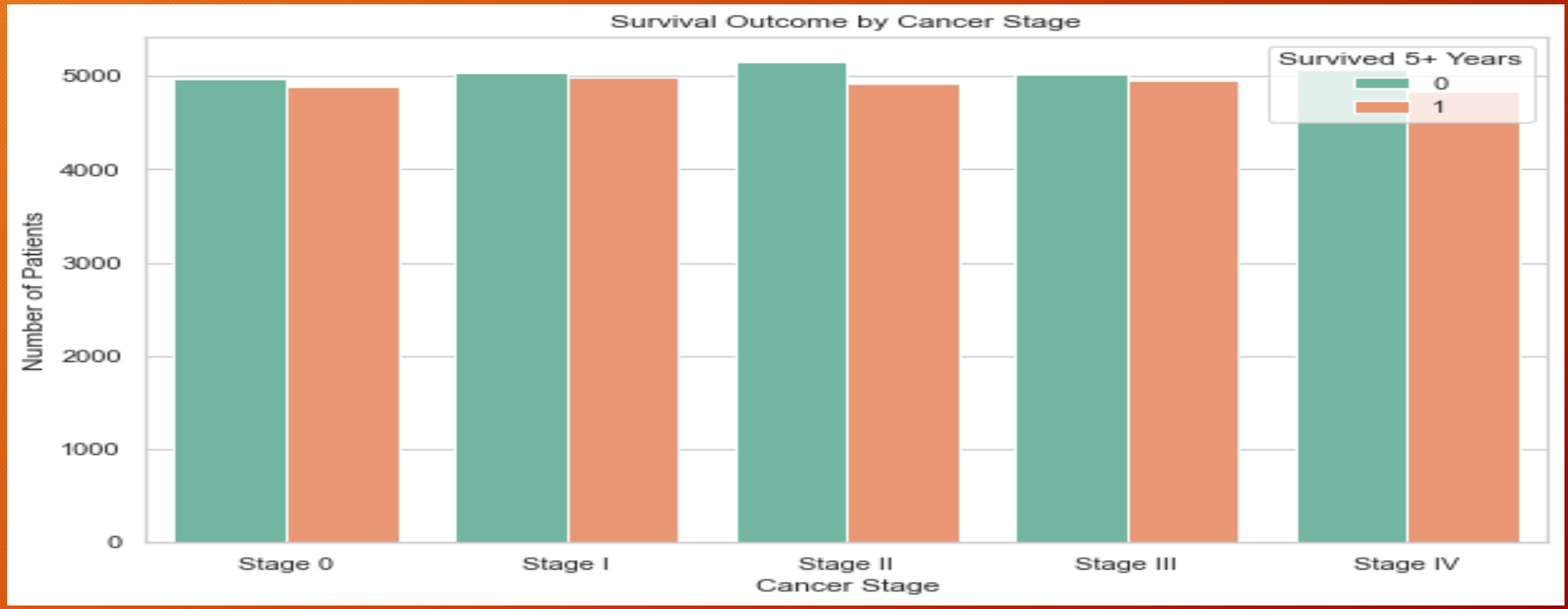
This project focuses on answering four essential questions:

1. What are the strongest continuous risk factors linked to cancer severity?
2. Do survival outcomes significantly vary by cancer stage?
3. Can we predict whether a patient's condition is severe (High vs Low severity)?
4. Can we predict whether a patient is likely to survive more than 5 years after diagnosis?

Top 10 Risk Factors for Cancer Severity

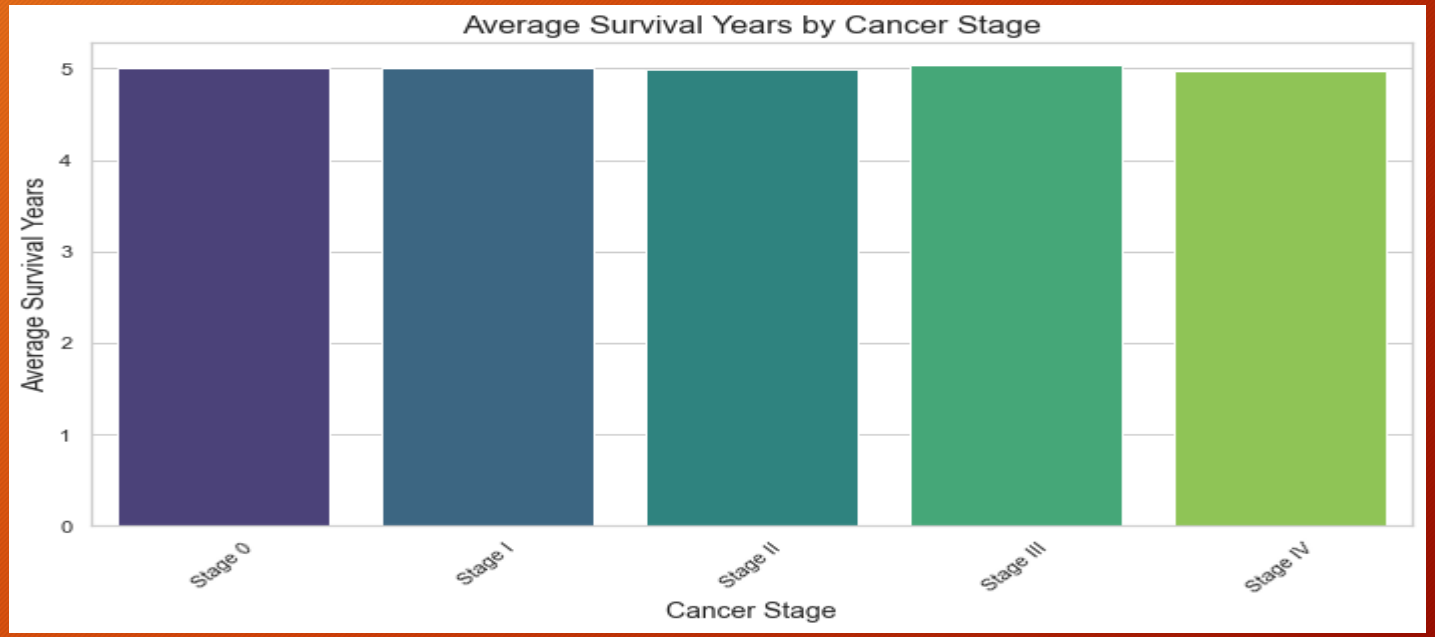


Survival outcome by cancer stage

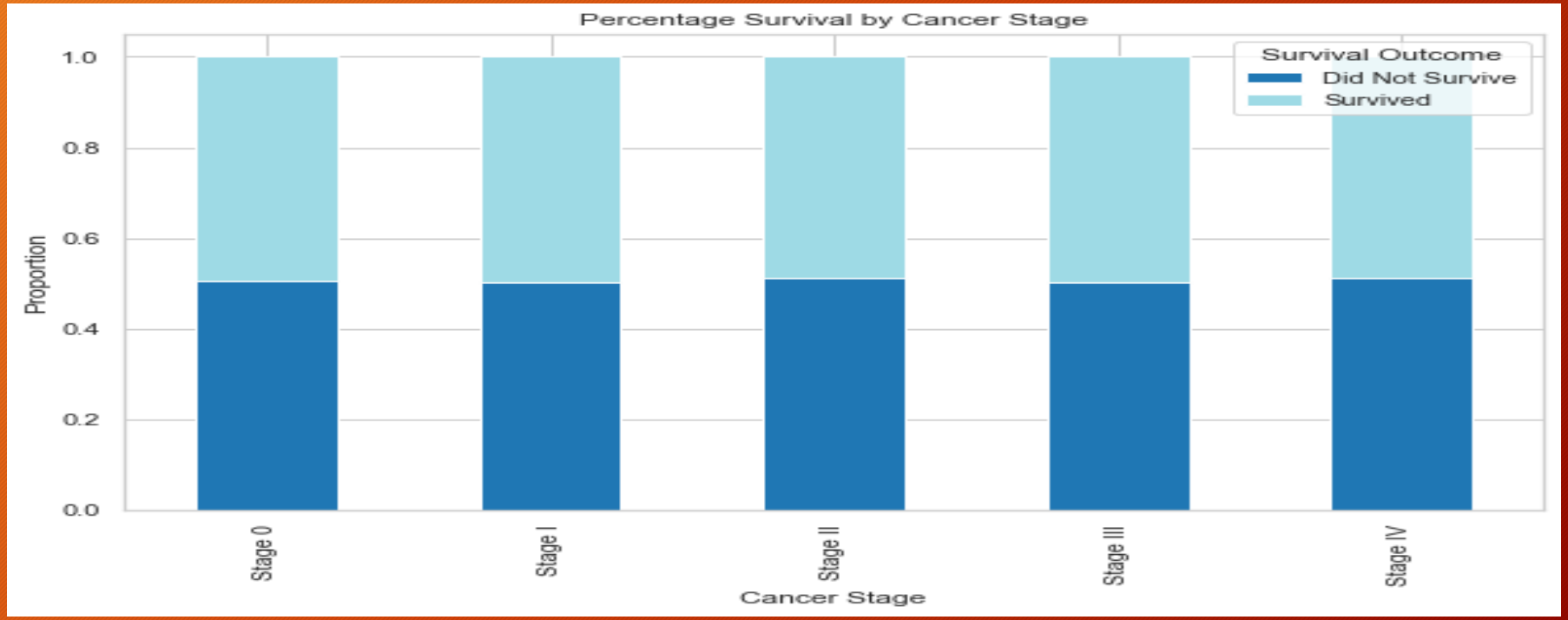


Average Survival Years by Cancer Stage

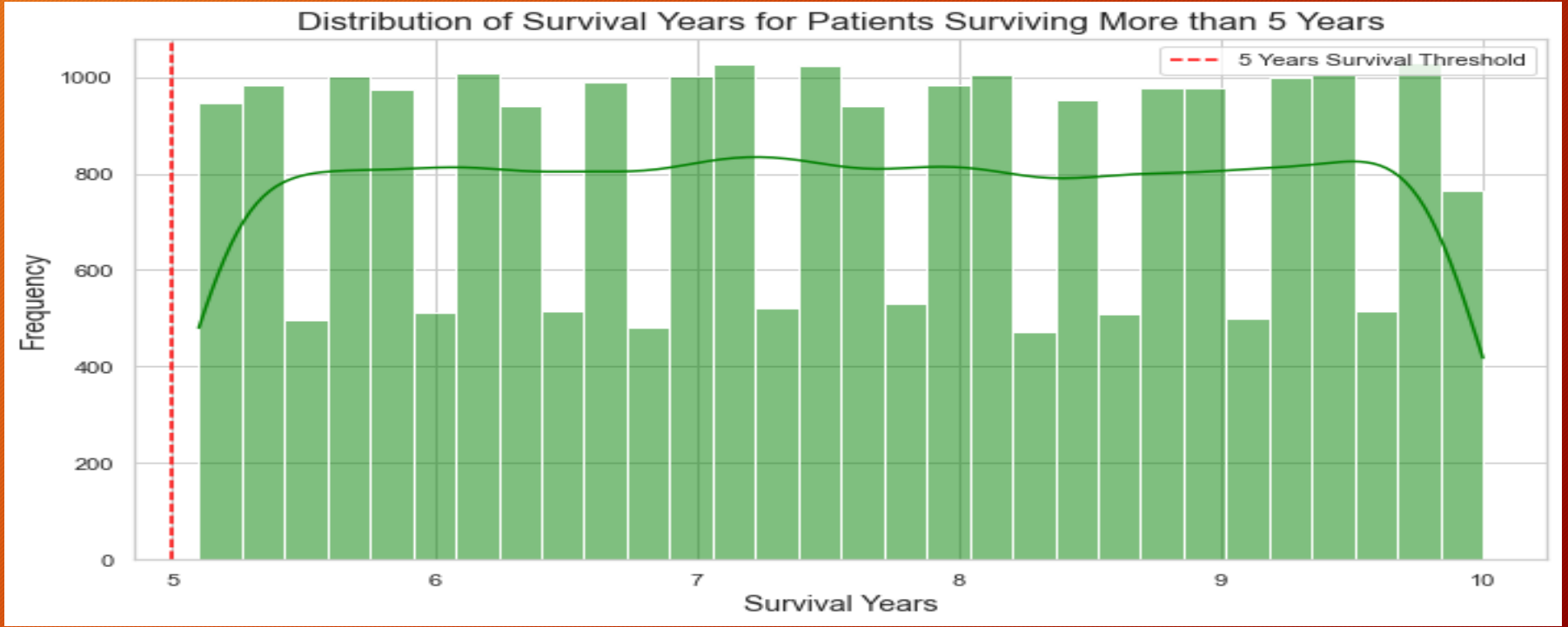
- The plot below shows that the average years across different stages does not vary significantly across different stages.



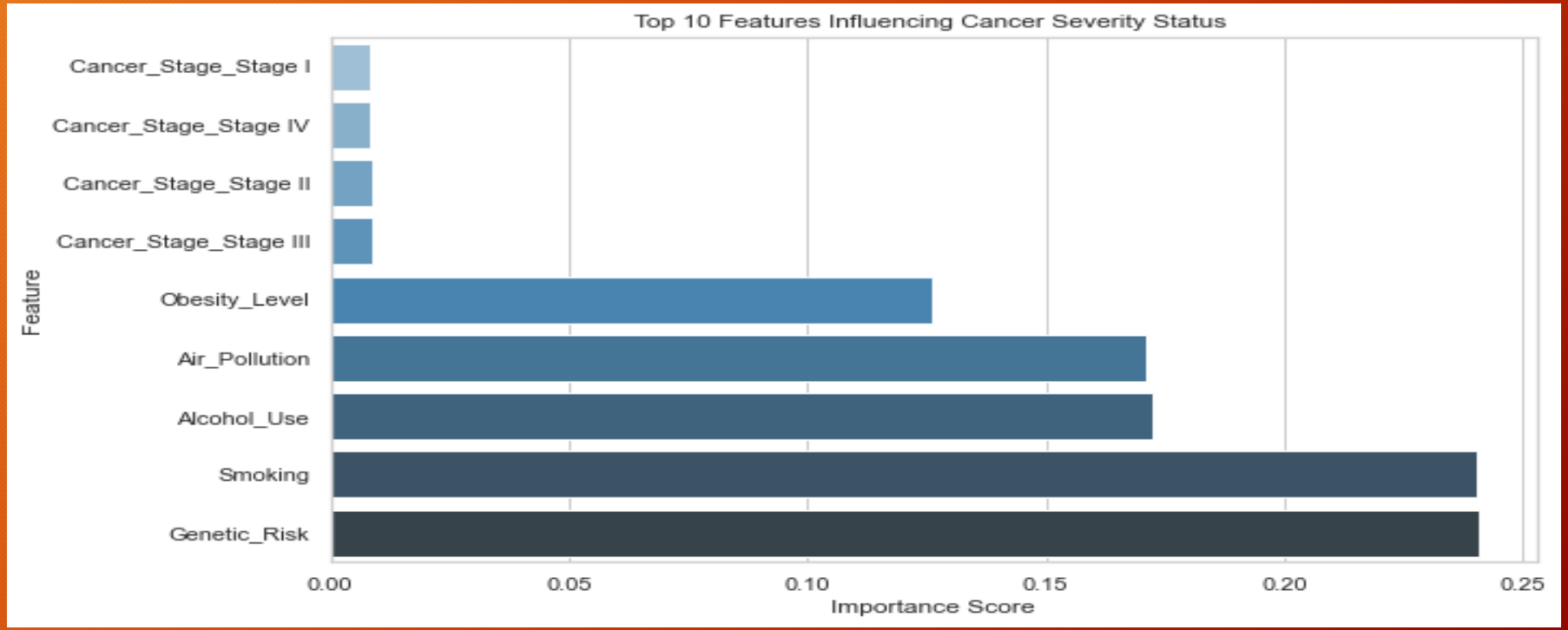
Percentage Survival by Cancer Stage



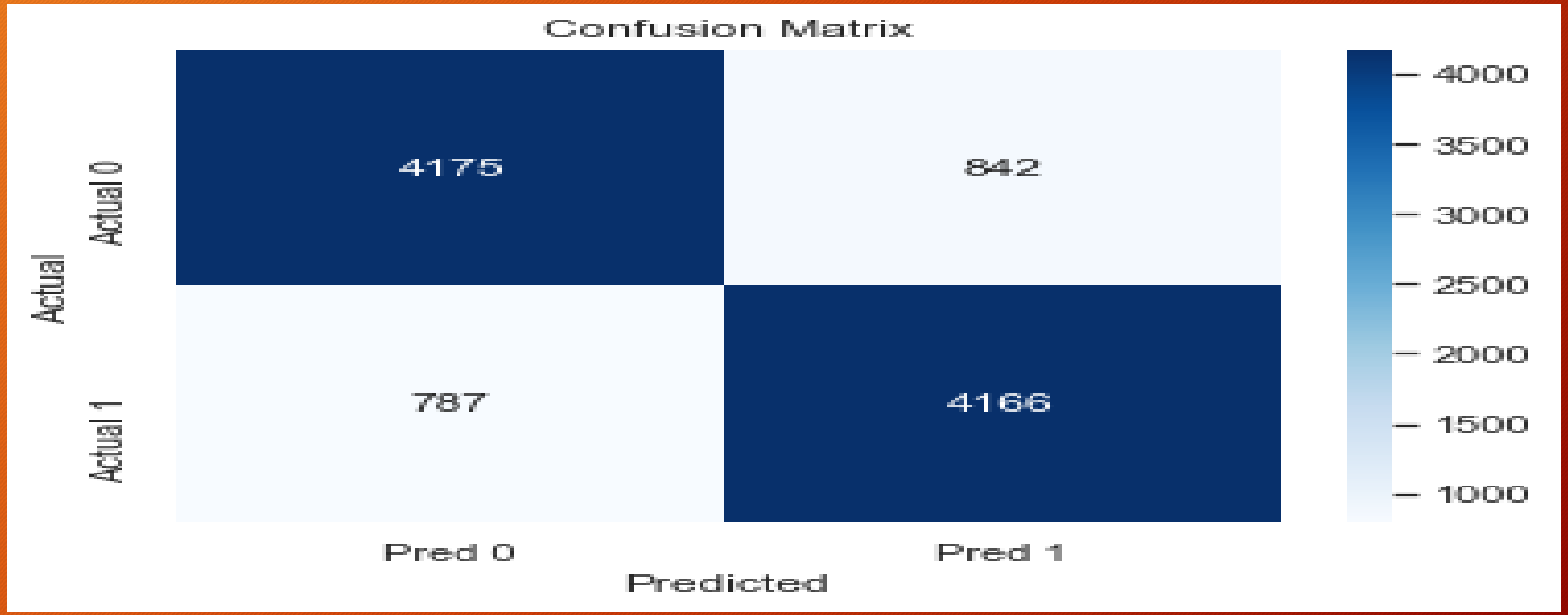
Survival Years Distribution for patients surviving 5+ years



Top 10 Features Influencing Cancer Severity Status



Confusion Matrix for the best performing model



Interpretation of the confusion matrix

True Negatives (Actual 0, Predicted 0): 4175

False Positives (Actual 0, Predicted 1): 842

False Negatives (Actual 1, Predicted 0): 787

True Positives (Actual 1, Predicted 1): 4166

Interpretation:

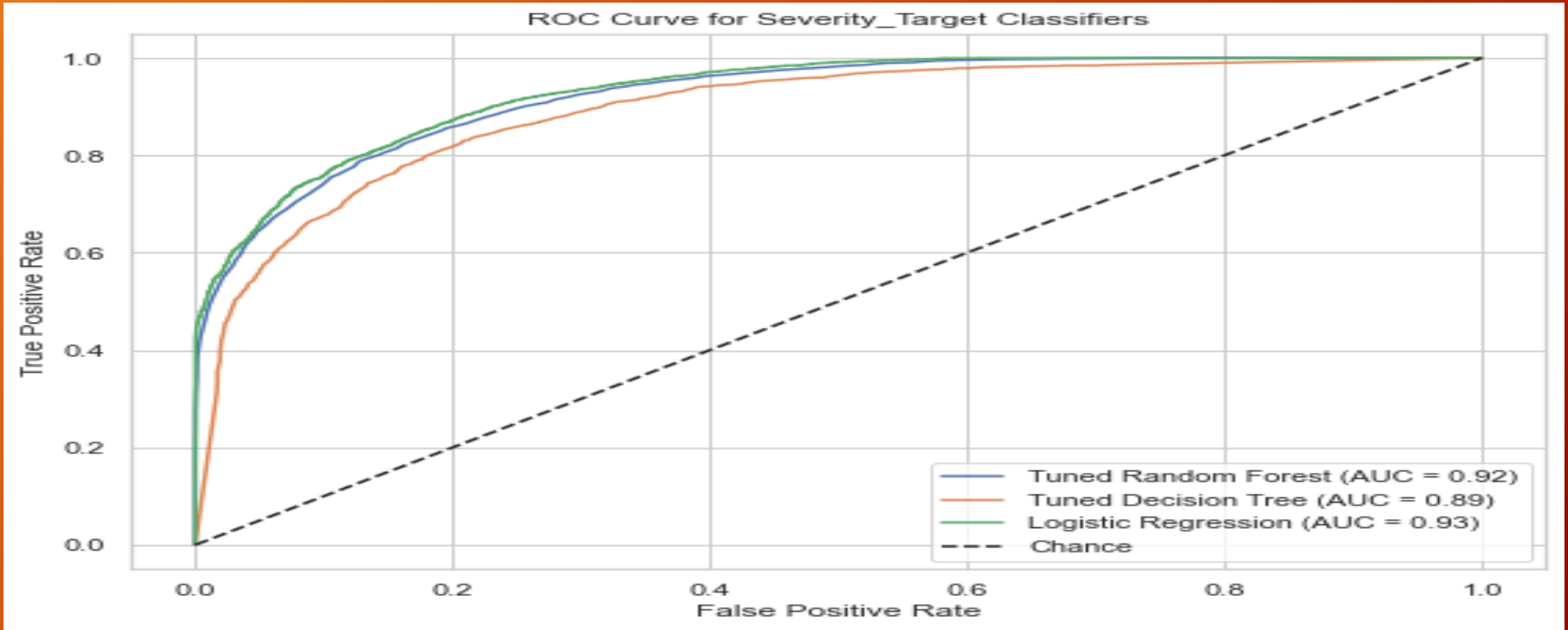
- The model correctly predicted 4175 samples as class 0 (True Negatives).

- The model incorrectly predicted 842 samples as class 1 when they are actually class 0 (False Positives).

- The model incorrectly predicted 787 samples as class 0 when they are actually class 1 (False Negatives).

- The model correctly predicted 4166 samples as class 1 (True Positives)

Model Performance



⚠️ Limitations

One of the main challenges was the poor performance of models in predicting 5-year survival outcomes, even after hyperparameter tuning and data cleaning. While severity classification performed well, survival prediction models like logistic regression, decision trees, and random forests showed limited accuracy. This suggests that important features may be missing and that the synthetic dataset lacks the complexity needed for modeling real-world survival outcomes.

🔑 Summary of Key Insights

- Severity prediction models performed well, with logistic regression leading in accuracy.
- Predicting 5-year survival was challenging, likely due to missing clinical features and the limits of synthetic data.
- Visuals revealed strong links between severity and factors like genetic risk, pollution, smoking, alcohol, and obesity.
- No significant survival differences were found across cancer stages.
- Tuned models consistently outperformed untuned ones, highlighting the value of optimization.

✅ Recommendations

- Explore advanced models (e.g., neural networks, ensemble methods) to improve accuracy
- Use survival analysis techniques tailored to time-based outcomes.
- Collaborate with healthcare experts to refine feature relevance.
- Apply explainability tools to make model predictions more transparent.
- Expand and regularly update the dataset to ensure accuracy and generalizability.
- Focus future research on lifestyle and genetic factors impacting cancer severity and survival.