Survival Rate and Death Rate Prediction Using Machine Learning

Introduction

This analysis employs machine learning to predict cancer survival and death rates by leveraging a comprehensive dataset containing demographic, socioeconomic, and geographic variables. The core goal is to enhance our understanding of how various factors contribute to cancer outcomes and to develop models that accurately predict these outcomes. Insights from this analysis can help guide public health policies and interventions, especially in the areas of education and poverty alleviation, where disparities in health outcomes are most pronounced.

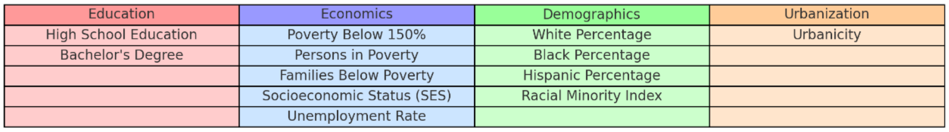
Methodology

Machine Learning Approach:

The machine learning models employed in this analysis use Random Forest Regression. Random Forests, an ensemble learning method, were chosen due to their robustness in handling complex, non-linear relationships between features. Random Forest models create multiple decision trees during training and combine their predictions, thus reducing overfitting, improving accuracy, and handling multicollinearity between variables.

Key Variables in the Analysis: The analysis considered a variety of predictors:

- Demographic factors: racial composition (White, Black, Hispanic percentages)
- Socioeconomic indicators: Education levels (Bachelor’s Degree rate, High School Education), unemployment, income, and poverty levels (Poverty Below 150%)
- Geographic information: Urbanicity(Urban vs. Rural)



Education	Economics	Demographics	Urbanization
High School Education	Poverty Below 150%	White Percentage	Urbanicity
Bachelor's Degree	Persons in Poverty	Black Percentage	
	Families Below Poverty	Hispanic Percentage	
	Socioeconomic Status (SES)	Racial Minority Index	
	Unemployment Rate		

Why Random Forest?

Random Forest models are highly effective for complex datasets that may have many interacting variables. This method also provides the benefit of ranking the importance of predictors, which is essential for understanding which factors are most influential in determining cancer survival and death rates. Additionally, Random Forest handles missing values and scaling issues well, making it suitable for this dataset's variety of variables.

Results

Random Forest Model for Death Rate:

- **Mean Squared Error (MSE):** 387.2382
- **R-squared Score:** 0.4762

The Random Forest model explained 47.62% of the variance in death rates. The moderately high R-squared value suggests that the predictors used in the model capture a substantial portion of the factors influencing death rates.

Random Forest Model for Survival Rate:

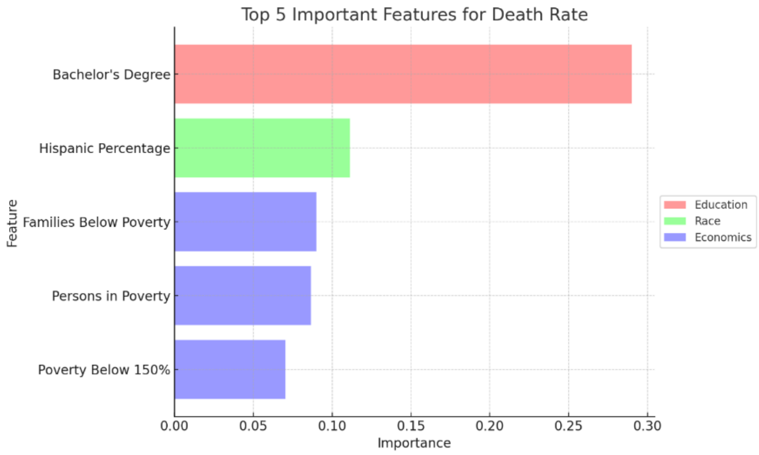
- **Mean Squared Error (MSE):** 25.0846
- **R-squared Score:** 0.2516

For survival rates, the model explained only 25.16% of the variance, indicating that survival rates might be influenced by factors not captured in the dataset or more complex non-linear interactions that the current model does not fully explain.

Feature Importance:

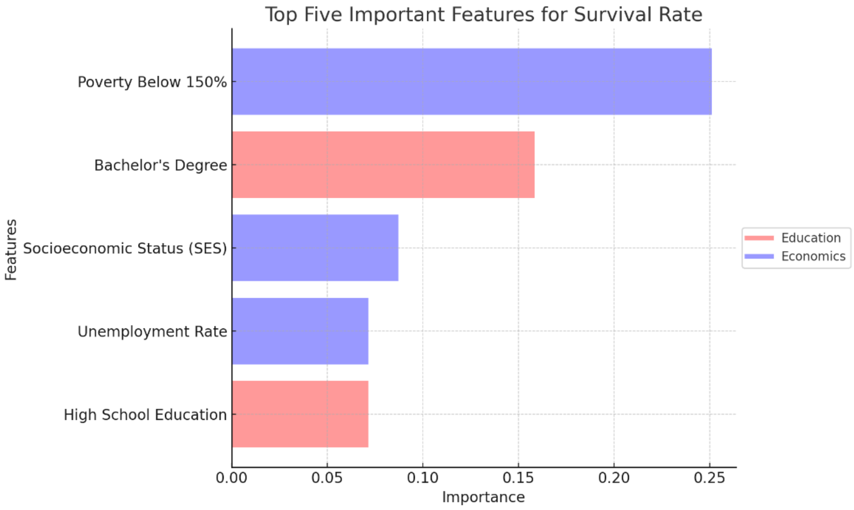
Death Rate Prediction: The most significant variable was the percentage of individuals with a Bachelor's Degree, followed by the Hispanic percentage. Education emerged as a consistent predictor across both death and survival models, signifying its central role in influencing health outcomes.

Figure 1: *Feature Importance for Death Rate Prediction*



Survival Rate Prediction: Similarly, education levels and poverty-related factors were crucial. Poverty (Poverty Below 150%) played a stronger role in predicting survival rates than death rates, suggesting economic factors significantly affect long-term cancer outcomes.

Figure 2: *Feature Importance for Survival Rate Prediction*



•

Discussion

1. Model Performance & Interpretations:

The Random Forest model for death rate performed notably better than the survival rate model. This could be due to the fact that death rates might be influenced by well-documented health disparities such as access to healthcare, lifestyle factors, or late-stage diagnosis, which are more readily quantifiable in the dataset. In contrast, survival rates might be influenced by complex factors not fully captured here, including treatment types, genetic predisposition, and other unrecorded individual-level data.

The lower R-squared for survival rates (25.16%) suggests that survival outcomes are more challenging to predict. Future research should explore adding clinical data (e.g., stage of diagnosis, treatment types) or patient-level data that can capture the complexities of cancer survivorship.

2. Significance of Education:

The finding that Bachelor's Degree percentage was the most influential predictor for both death and survival rates is a reflection of the strong correlation between education and health. Individuals with higher education levels typically have better access to healthcare, more awareness of preventative measures, and higher income—all of which positively impact health outcomes.

3. Ethnic Composition and Health Outcomes:

The Hispanic percentage emerged as a significant predictor for death rates but not survival rates. This could suggest that while mortality rates among Hispanic populations are affected by socioeconomic and healthcare access disparities, survival rates may not be as significantly impacted by ethnicity alone. This raises important questions about the specific barriers facing different ethnic groups and the role of healthcare access in mitigating these disparities.

4. Economic Disparities:

The prominence of poverty-related variables (such as Poverty Below 150%) in predicting survival rates emphasizes that poverty directly affects long-term health outcomes. Economic instability can limit access to timely treatments, quality care, and long-term follow-up, leading to worse survival rates.

5. Non-linear Relationships:

The improved performance of the Random Forest models over linear models indicates the presence of non-linear relationships between predictors and outcomes. Given this, exploring more advanced non-linear models, such as **Gradient Boosting Machines** (GBM) and **Neural Networks**, may further improve predictive accuracy. GBM, in particular, has the potential to provide higher accuracy by sequentially correcting model errors, while Neural Networks could capture more intricate patterns.

6. Model Refinements:

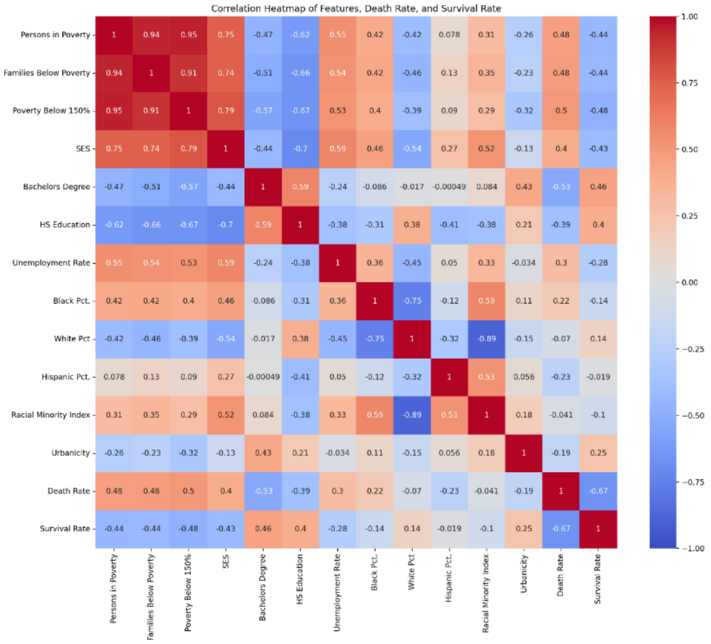
Future improvements could be achieved through:

- **Hyperparameter tuning**: Optimizing the Random Forest model parameters (number of trees, maximum depth, etc.).
- **Feature engineering**: Creating interaction terms (e.g., interaction between poverty levels and race) to better capture complex relationships.
- **Incorporating additional variables**: Integrating clinical data, individual health behaviors, and healthcare access metrics to improve model comprehensiveness.

7. Geospatial Analysis:

Incorporating geospatial data could offer new insights into how geographic disparities contribute to health outcomes. For example, regional variations in healthcare access, environmental factors, and social determinants of health could be explored using **spatial machine learning** techniques.

Figure 3: *Correlation Heatmap of Predictors*



Conclusion

This analysis successfully used machine learning to identify key predictors of cancer survival and death rates, with education and poverty emerging as the most critical factors. The Random Forest model performed well for death rate prediction, though further improvements could be made for survival rates. Policymakers should consider these findings in designing interventions that target education and poverty to improve health outcomes, especially in underprivileged and ethnically diverse populations.

Future Directions

1. **Exploring Other Algorithms**: Testing additional models like **Gradient Boosting Machines (GBM)**, **Support Vector Machines (SVM)**, and **Neural Networks** could uncover more complex relationships.
2. **Survival Analysis**: Applying survival analysis techniques such as **Cox proportional hazards** or **Kaplan-Meier curves** could help better understand survival times and the risk factors associated with survival.
3. **Treatment Effect Modeling**: Incorporating treatment and healthcare data to predict the effectiveness of specific treatments on survival rates.
4. **Longitudinal Analysis**: Collecting and analyzing longitudinal data to better capture the changes in patient health status over time and the impact of interventions.

Suggested Machine Learning Applications**:**

- **Time-to-Event Modeling**: Using survival analysis techniques (Cox models, Kaplan-Meier analysis) to predict not just survival rates but time-to-event (i.e., time until death or remission).
- **Healthcare Resource Optimization**: Machine learning models can predict healthcare resource needs (hospital beds, staff) based on predicted death or survival rates, improving resource allocation.
- **Risk Stratification Models**: Build models that identify high-risk individuals based on socioeconomic and demographic factors, guiding targeted interventions to improve cancer outcomes.

References

- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.