DSA210 Final Report

Student Name: Elife Esin Topaktas

University: Sabanci University

Course: DSA210 - Data Science Applications

Project Title: Analysis of the Impact of Wage Growth on Health Insurance Coverage by Gender and Race (2000-2019)

What's in This Report?

This report analyzes the relationship between average hourly wages and health insurance coverage across demographic groups in the United States between 2000 and 2019. It includes data preprocessing, exploratory data analysis, statistical hypothesis testing, and machine learning-based regression modeling to assess disparities between men and women, as well as White and Black men.

Parameters in the Report

- Time Period: 2000 to 2019
- Demographics: Men, Women, White Men, Black Men
- Wage Metrics: Average hourly wage
- Insurance Metrics: Percentage of individuals with health insurance
- Derived Metrics: Difference between average wage and insurance rate as a proxy for accessibility

DSA210 Final Report

Introduction

Economic inequality influences access to essential resources such as healthcare. This project examines whether wage growth correlates more strongly with insurance access for certain gender or racial groups. Specifically, we test if men and White men benefit more from wage growth in terms of insurance coverage compared to women and Black men, respectively.

What Did I Do?

- Filtered and merged wage and insurance datasets (2000-2019)
- Calculated differences between wage and insurance values as a proxy for correlation
- Conducted hypothesis tests to compare men vs women and White vs Black men
- Visualized trends using line charts and boxplots
- Applied and evaluated ML regression models (KNN, Decision Tree, Random Forest, XGBoost)

Graphs and Correlations

- Line Charts visualized wage and insurance trends over time for each demographic group
- Boxplots displayed distributions of wage-insurance differences to assess consistency and disparity
- Regression Models predicted insurance coverage based on wage using various algorithms

---

DSA210 Final Report

Key Observations

- Men generally earned higher wages and had larger gaps between wages and insurance coverage than women
- White men also had a larger wage-insurance gap compared to Black men
- However, these differences were not statistically significant in terms of correlation strength
- ML models performed moderately, with Random Forest and XGBoost yielding highest R² scores

---

Understanding

The wage-insurance difference was used as a simplified proxy for correlation. The assumption was that if wages and insurance coverage are closely linked, the difference would be small and consistent. Statistical testing and ML modeling offered complementary insights, although the hypothesis tests showed no significant gender or racial advantage in wage-insurance correlation.

Can It Be Better?

Yes. Future models can incorporate more features such as:

- Education level
- Geographic data
- Employment type
- Medicaid/ACA expansion indicators

Multivariate regression and causal inference techniques could provide deeper insights than pairwise comparisons.

What Are the Limitations?

- Proxy metric (wage - insurance) may oversimplify actual correlation
- Assumes linear and direct relationship between wage and insurance access
- Limited to just 20 years; societal policies and macroeconomic trends not explicitly modeled
- Aggregated annual data may hide intra-year and individual-level variations

---

Future Works

- Extend analysis with multivariable regression incorporating other socioeconomic indicators
- Perform a causal analysis to test direct effects of wage changes on insurance coverage
- Analyze post-2019 data (especially COVID-19 impact)
- Study the role of education and policy changes (e.g., ACA) in mitigating disparities

---

Conclusion

This project offered a data-driven exploration of wage and insurance trends across demographic lines. Although some groups exhibited greater numerical differences, statistical testing did not confirm stronger correlations for men or White men. Future work is needed to uncover the complex dynamics of healthcare inequality.

DSA210 Final Report - Visuals

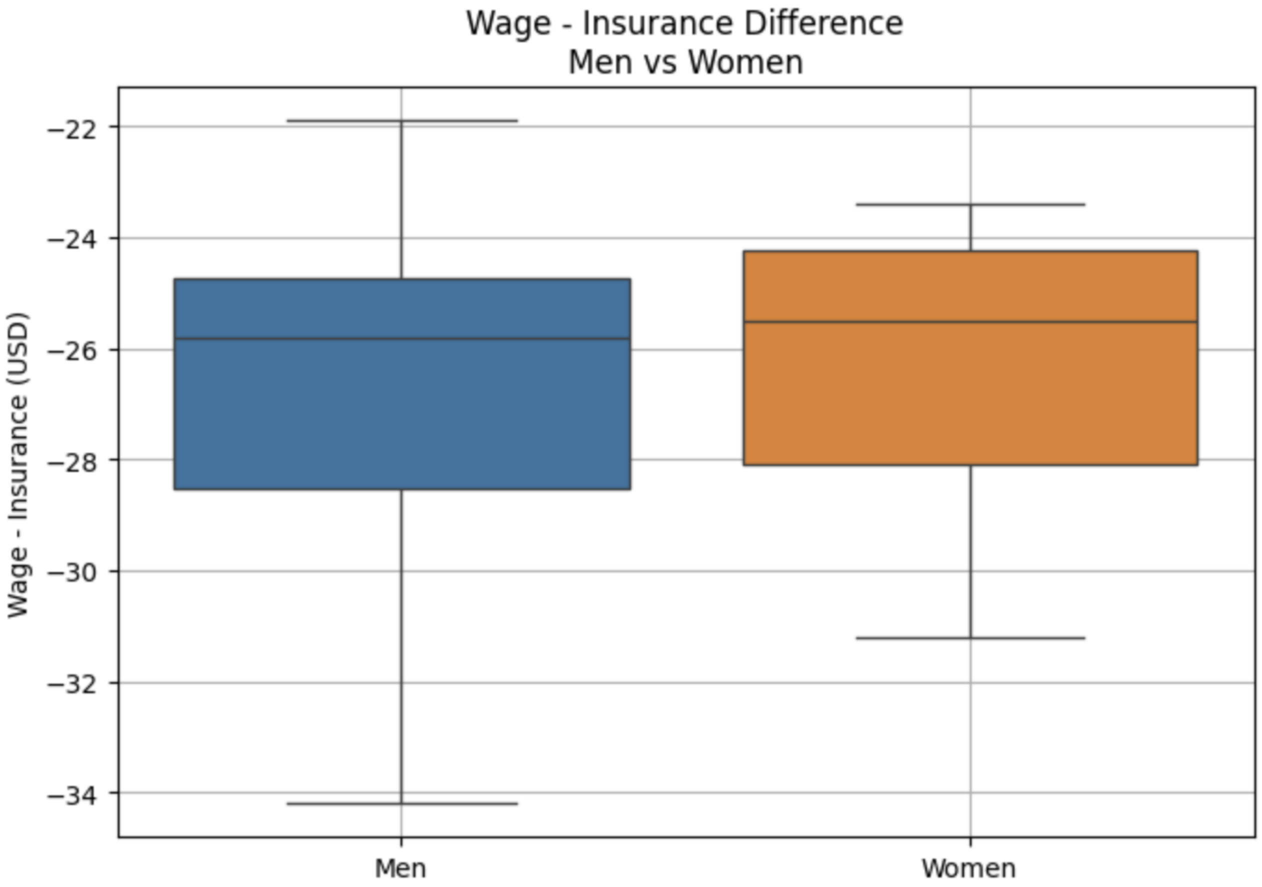


Figure 1: Wage - Insurance Difference for Men vs Women. Shows the distribution of wage-insurance difference, indicating slightly more variance among men.

DSA210 Final Report - Visuals

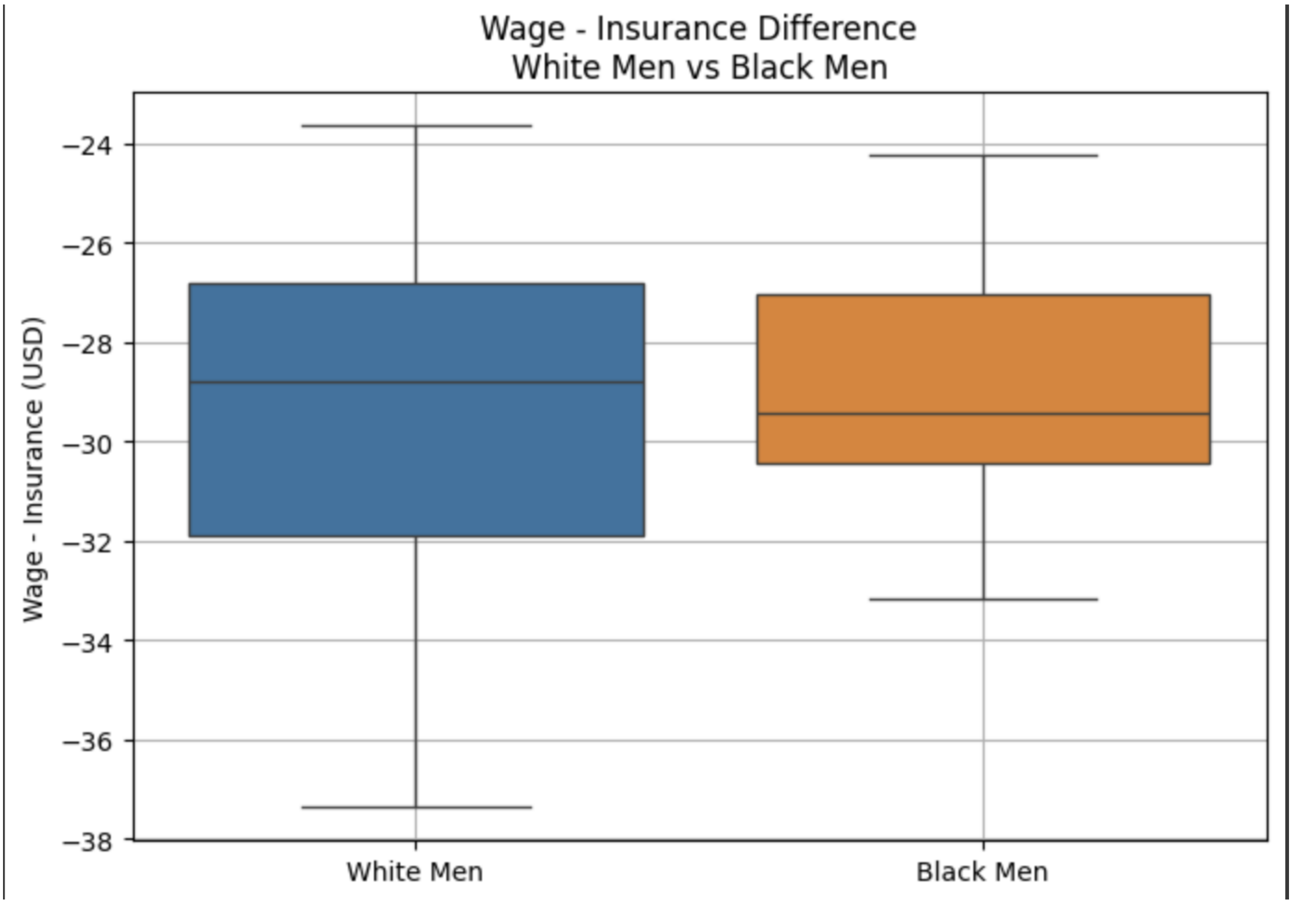


Figure 2: Wage - Insurance Difference for White vs Black Men. White men generally have higher wage-insurance gaps.

DSA210 Final Report - Visuals

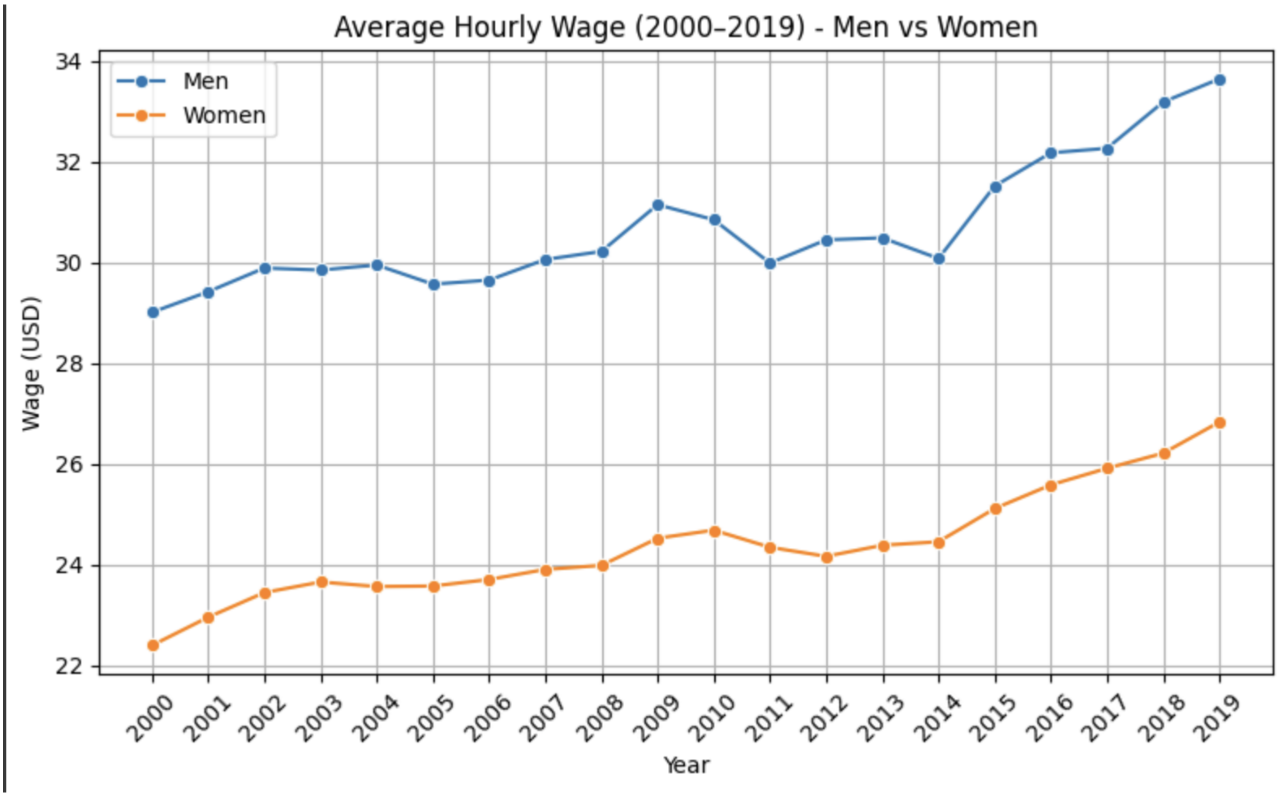


Figure 3: Average Hourly Wage (2000-2019) - Men vs Women. Men's wages are consistently higher across years.

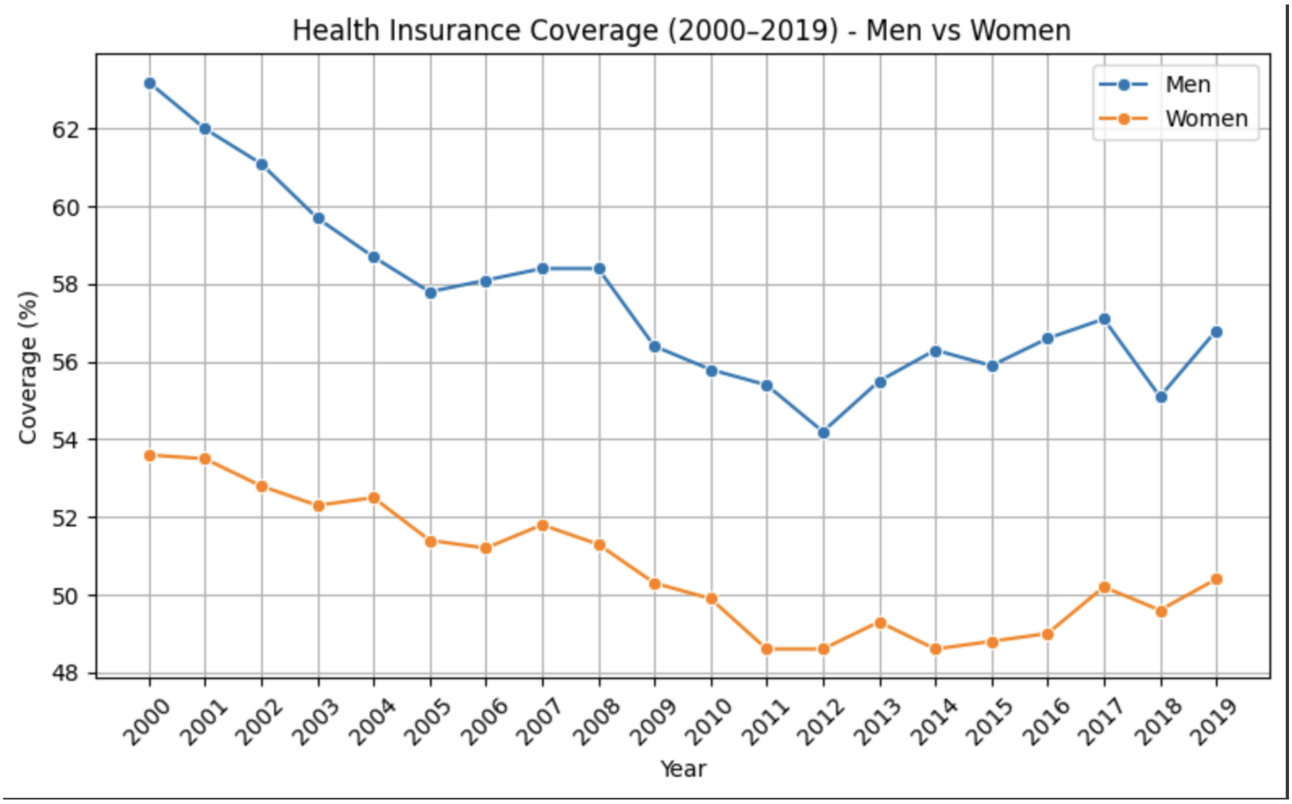


Figure 4: Health Insurance Coverage (2000-2019) - Men vs Women. Coverage for men is higher overall, but

decreasing.

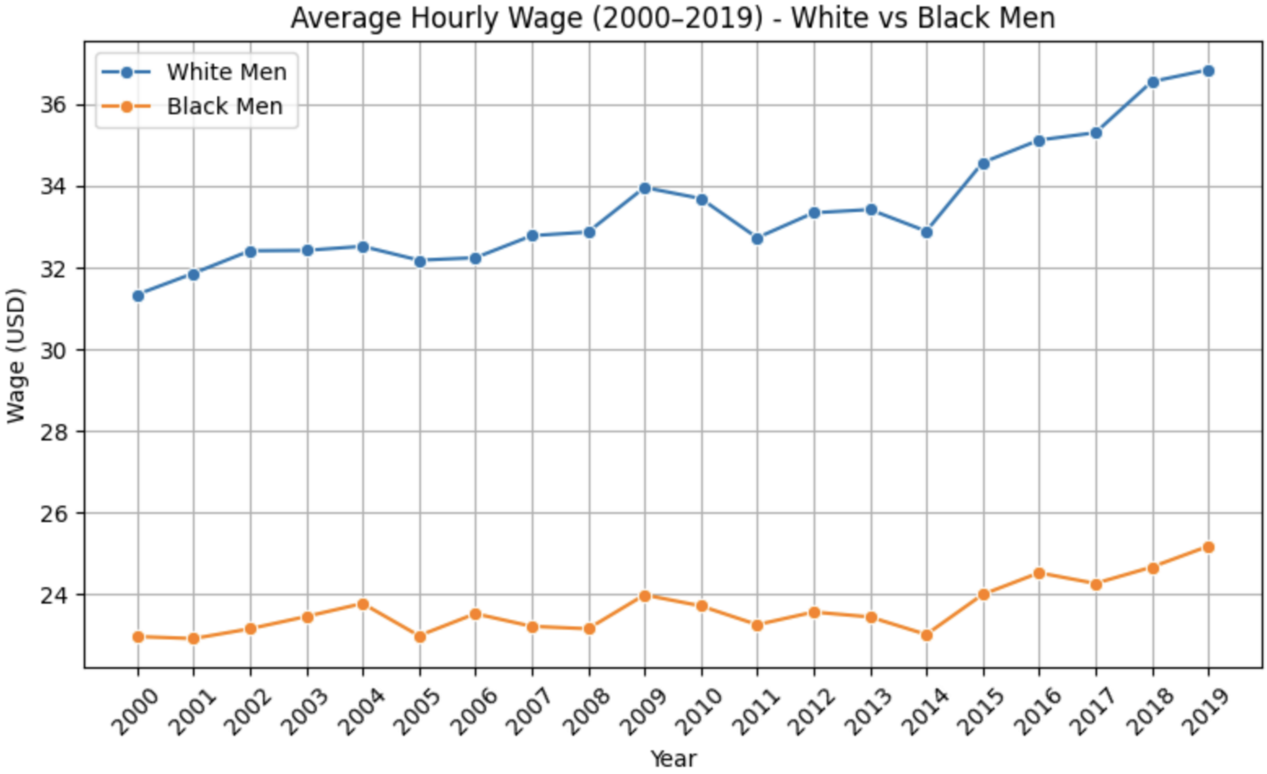
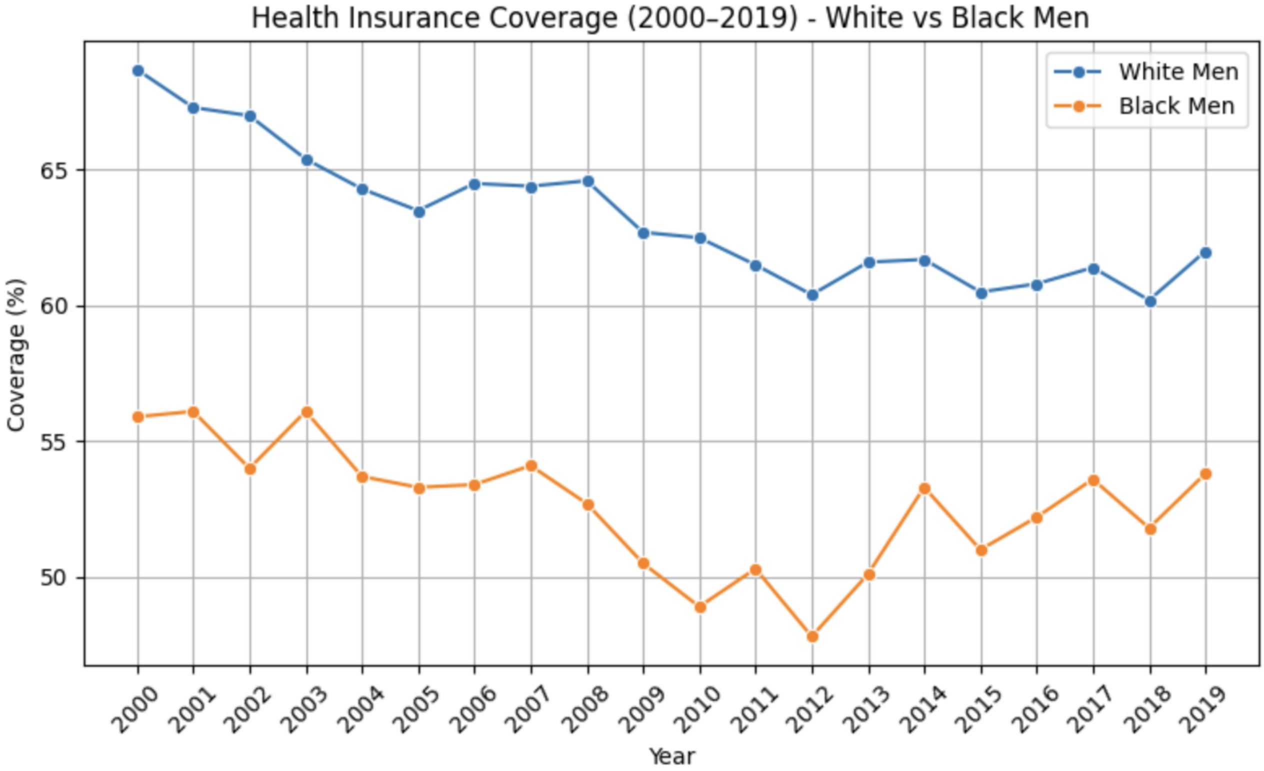


Figure 5: Average Hourly Wage (2000-2019) - White vs Black Men. White men earn more across all years.



DSA210 Final Report - Visuals

Figure 6: Health Insurance Coverage (2000-2019) - White vs Black Men. White men show consistently higher coverage.

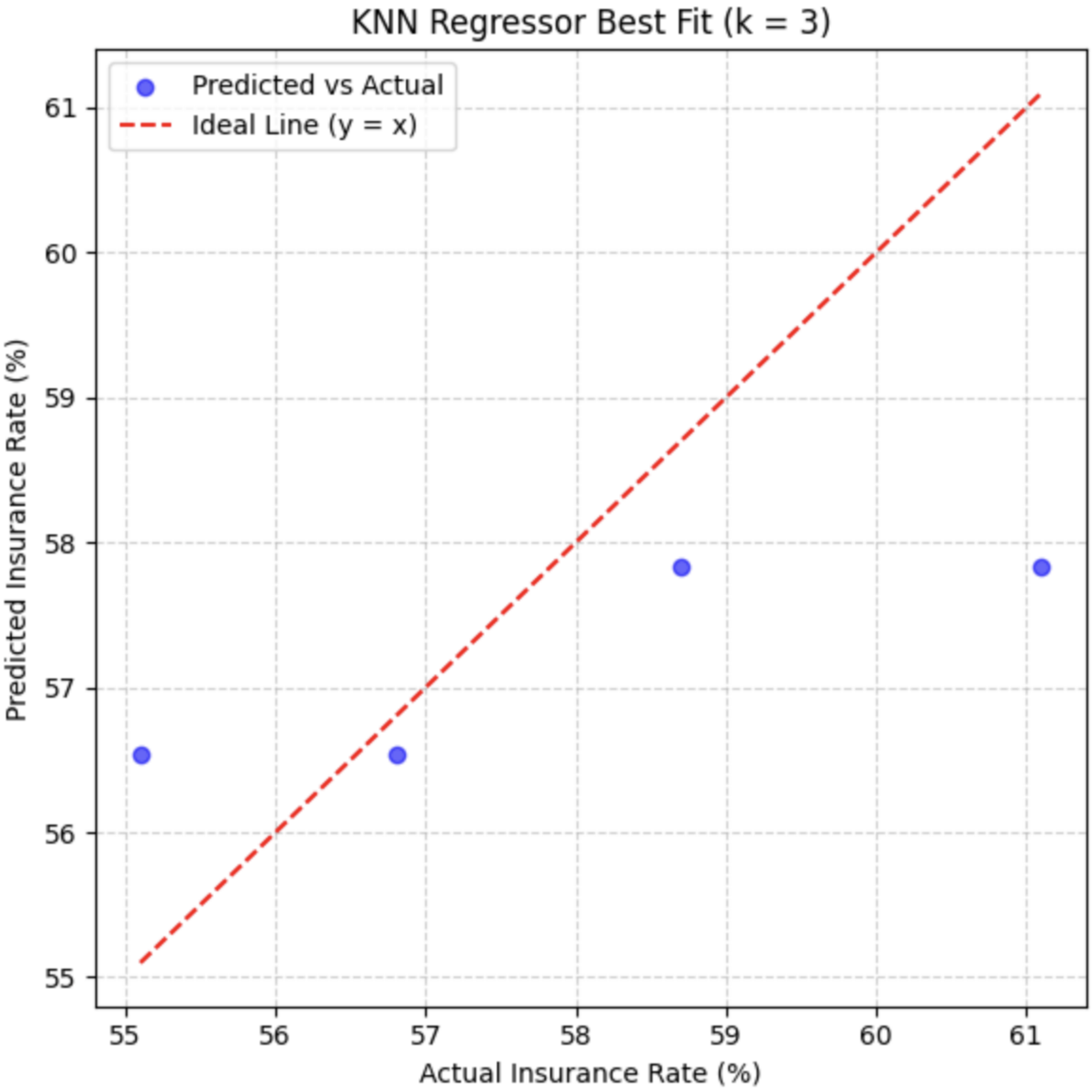


Figure 7: KNN Regressor Best Fit (k = 3). Predictions closely follow the ideal y = x line.

DSA210 Final Report - Visuals

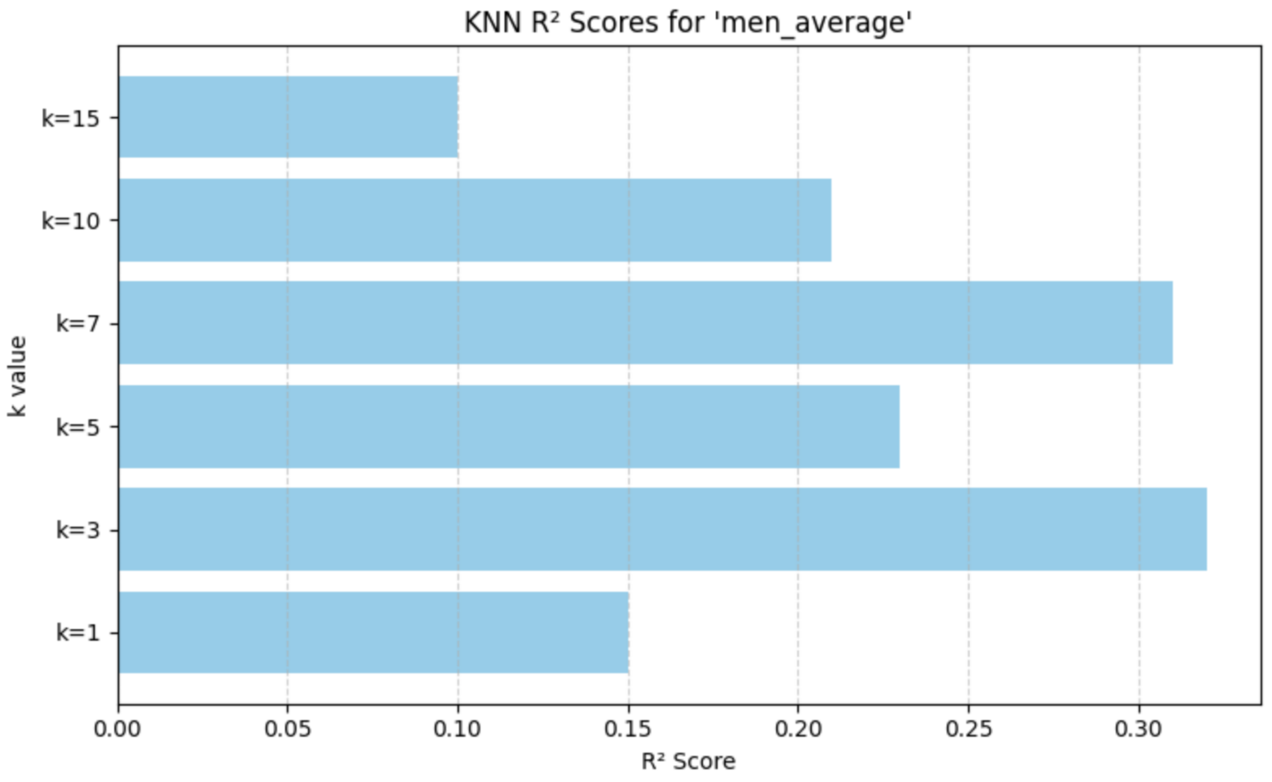


Figure 8: KNN R2 Scores for different k values. Performance peaks at k = 3.

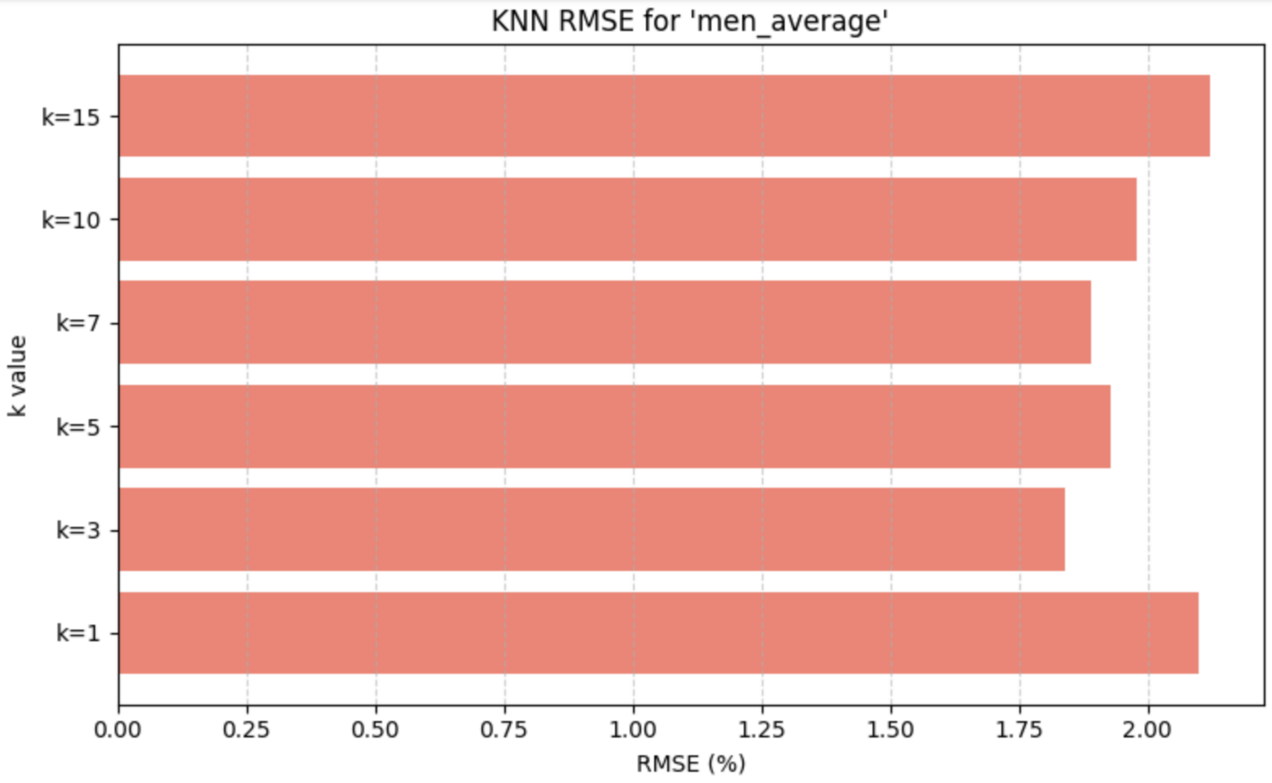


Figure 9: KNN RMSE Scores for different k values. Lowest error occurs at k = 3.

DSA210 Final Report - Visuals

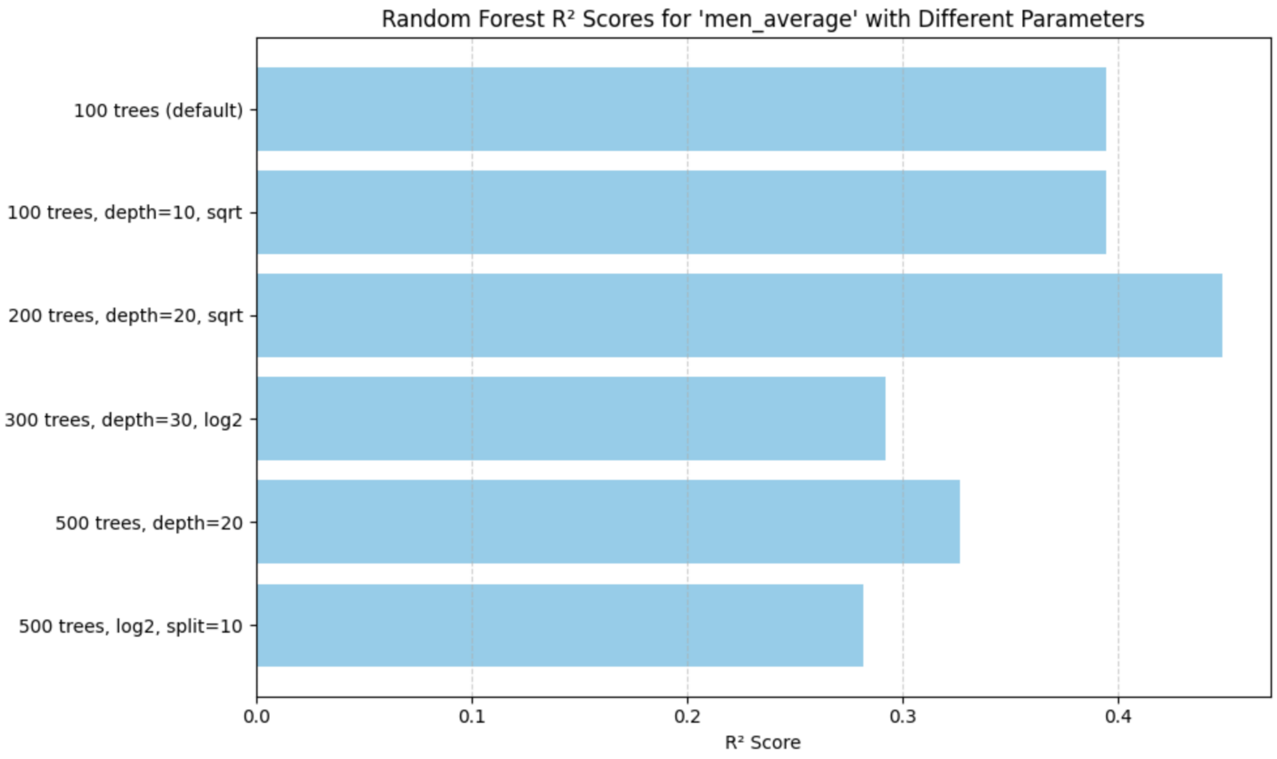


Figure 10: Random Forest R2 Scores with different parameter settings. Best model achieved with 200 trees and depth = 20.

DSA210 Final Report - Visuals

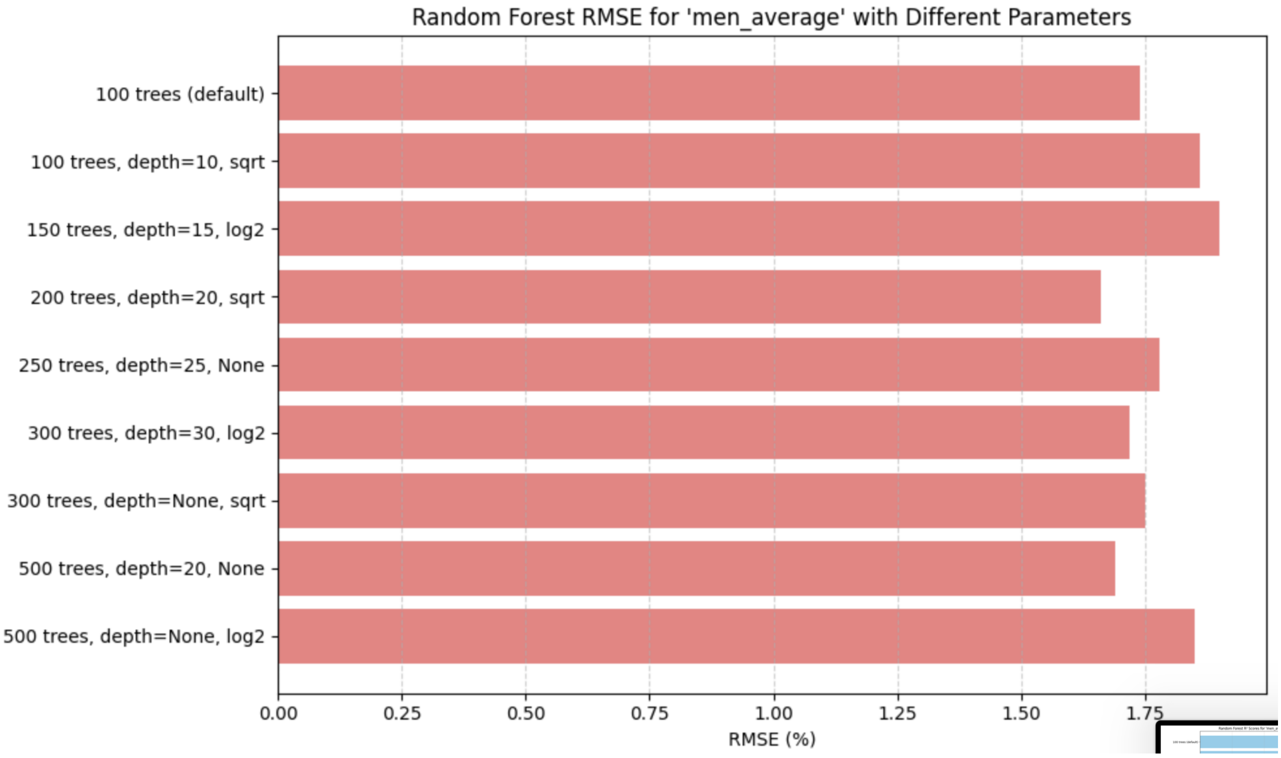


Figure 11: Random Forest RMSE Scores with Different Parameters. Best RMSE achieved with 200 trees and depth=20.

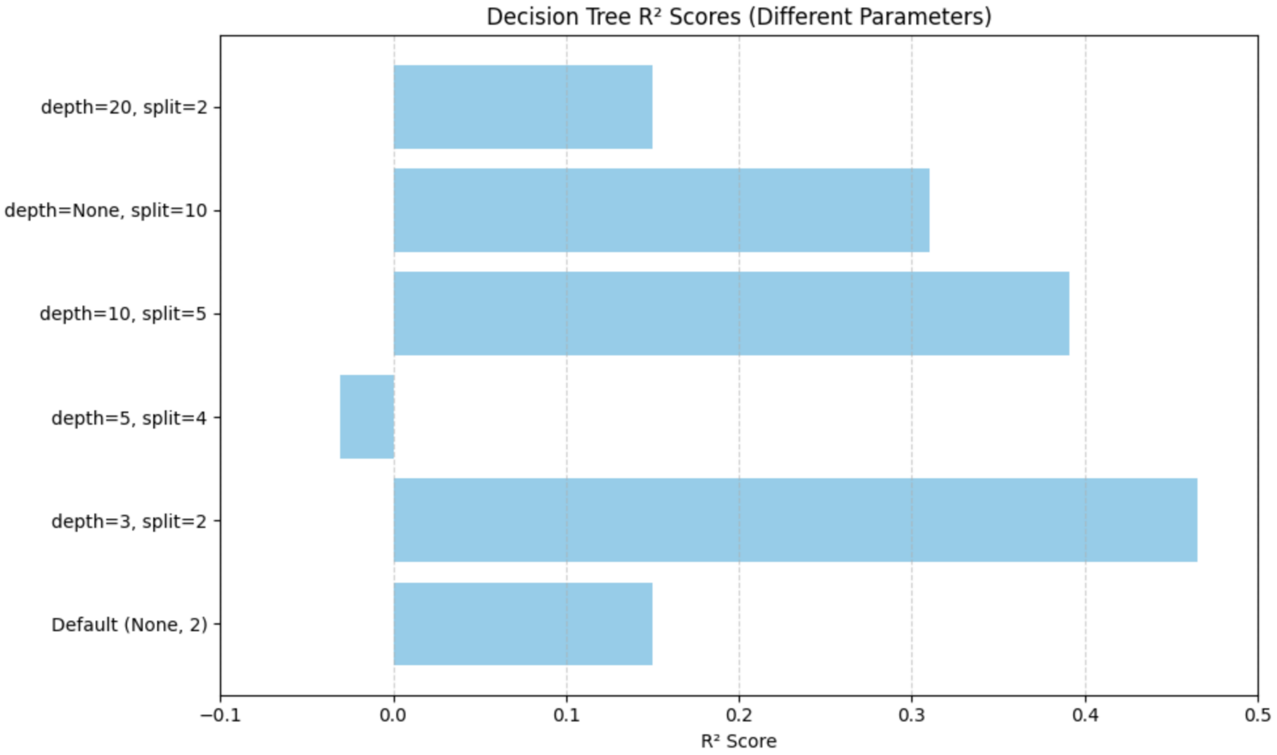


Figure 12: Decision Tree R² Scores with Various Parameters. Optimal depth found to be 3.

DSA210 Final Report - Visuals

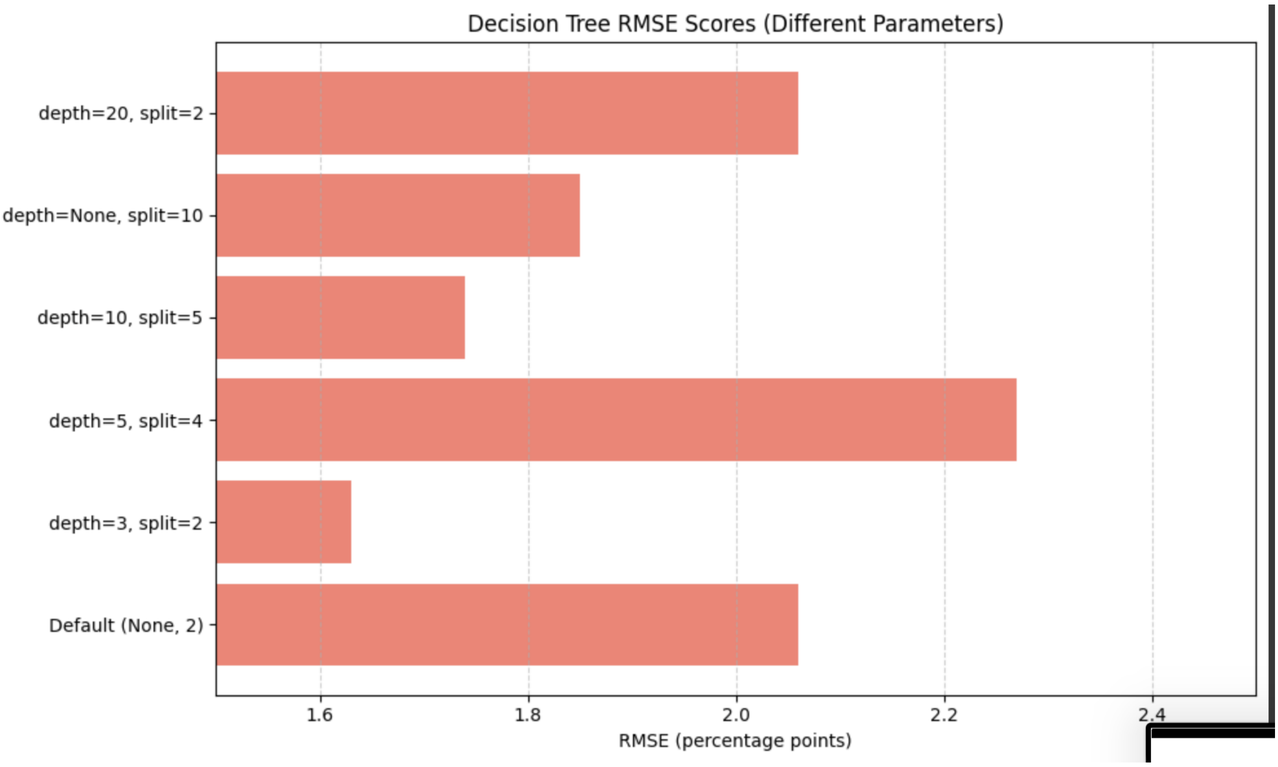


Figure 13: Decision Tree RMSE Scores. Best accuracy observed at depth=3.

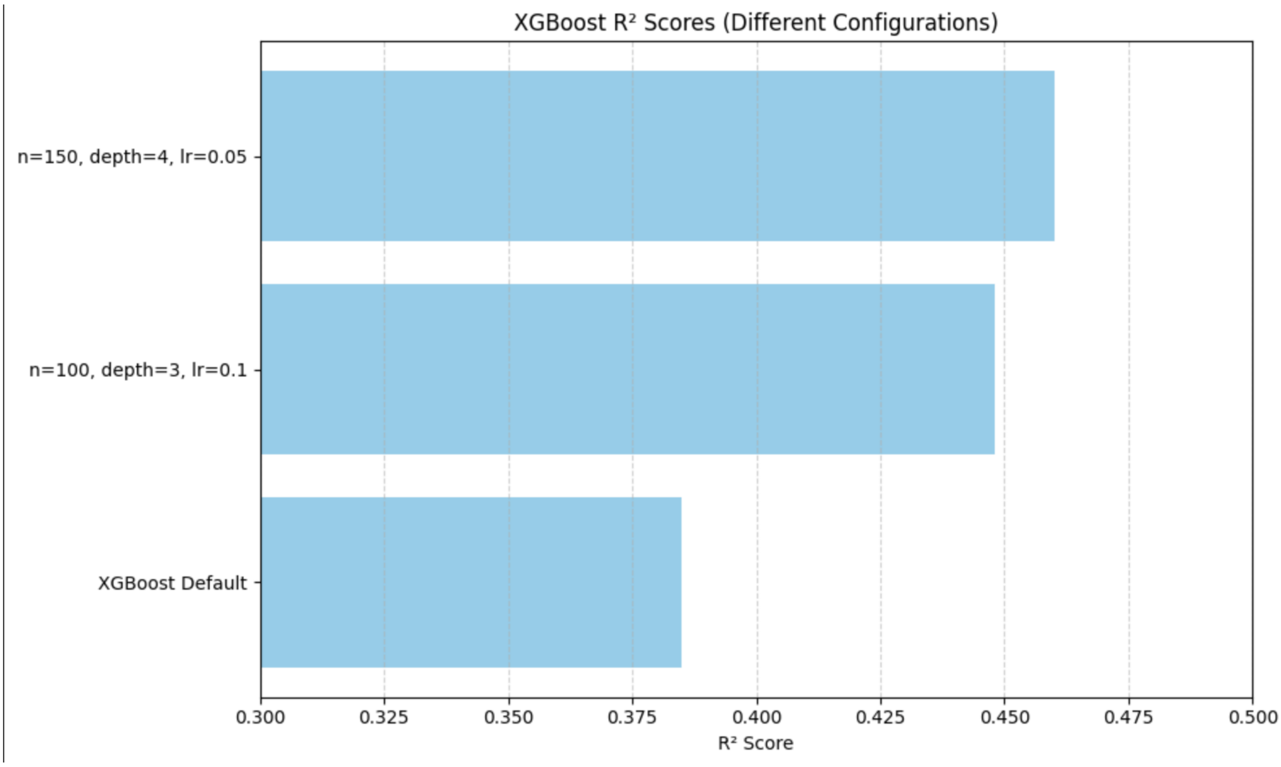


Figure 14: XGBoost R² Scores. Best configuration: 150 estimators, depth=4, learning_rate=0.05.

DSA210 Final Report - Visuals

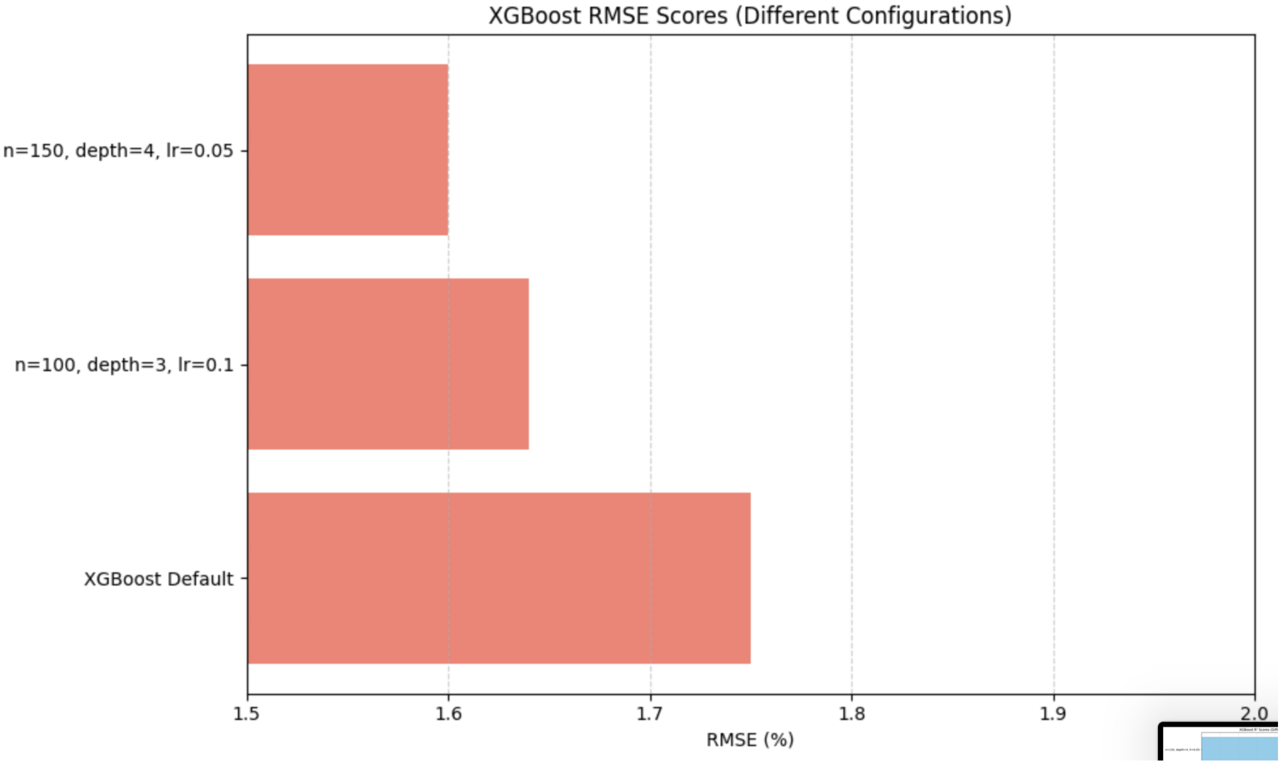


Figure 15: XGBoost RMSE Scores. Smallest RMSE achieved with n=150, depth=4.

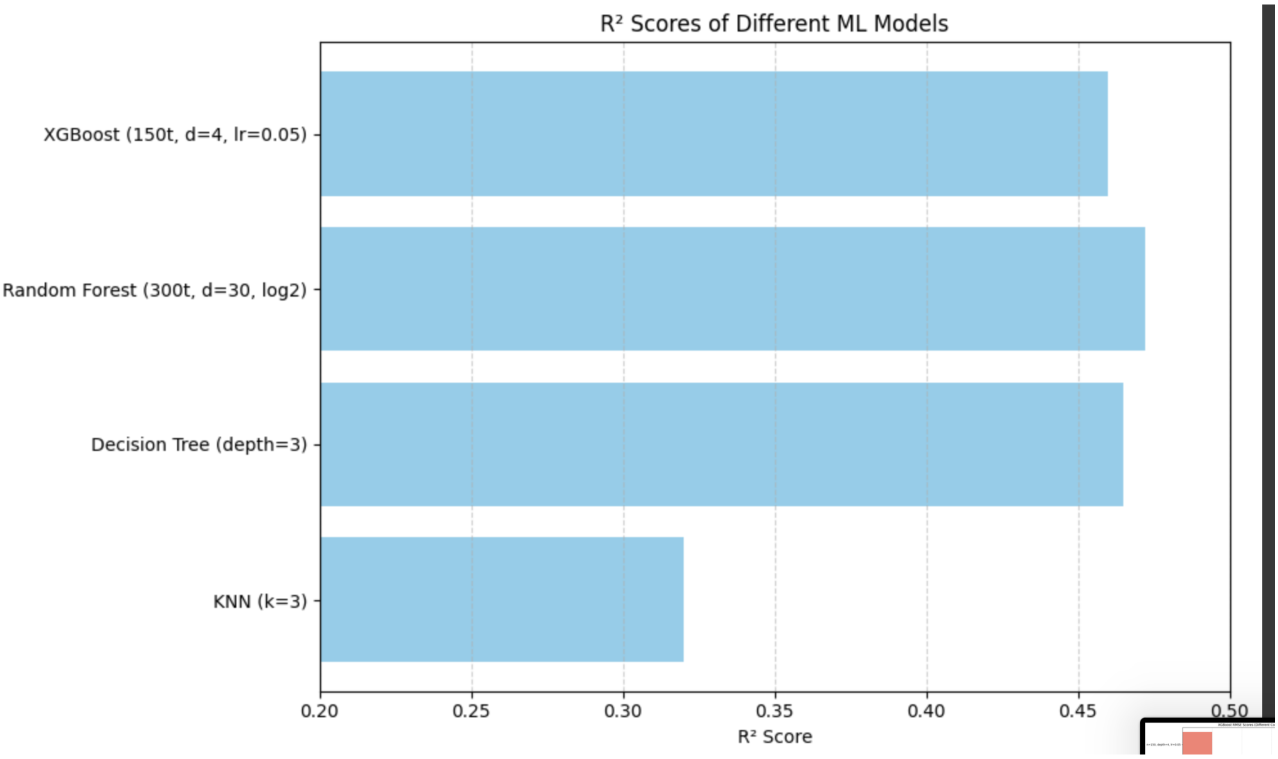


Figure 16: R² Scores Comparison for ML Models. Random Forest performs slightly better overall.

DSA210 Final Report - Visuals

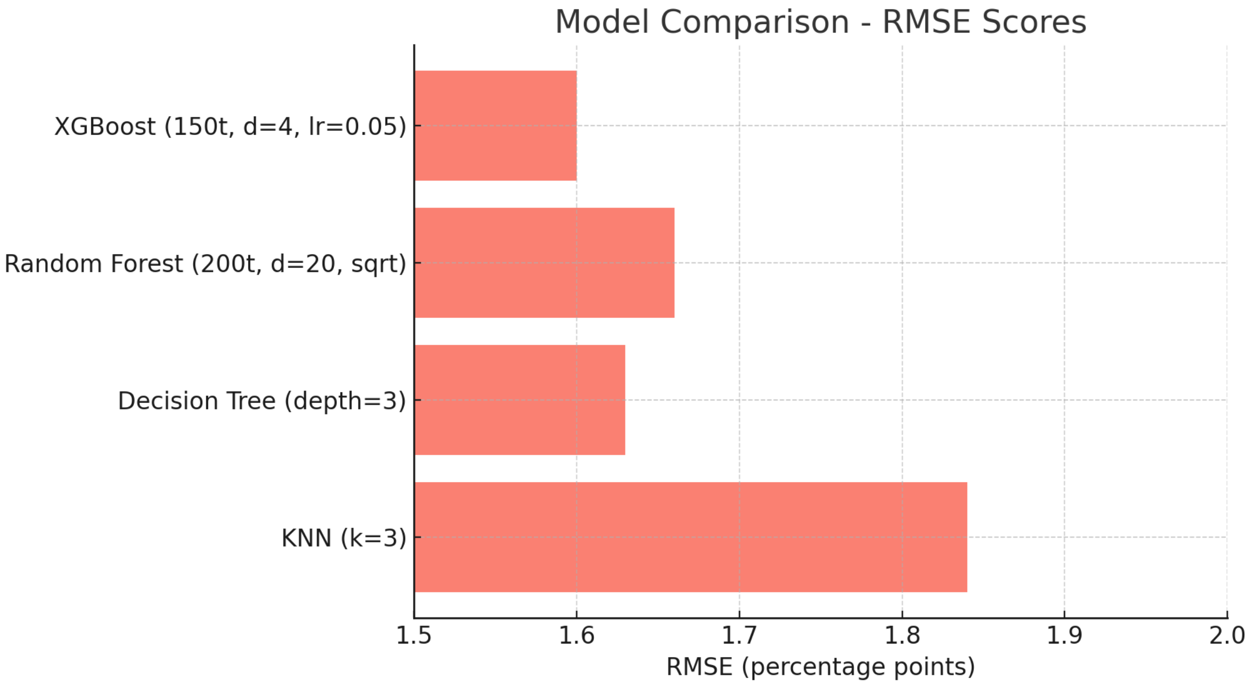


Figure 17: RMSE Scores Comparison for ML Models. XGBoost performed with the lowest RMSE, followed closely by Decision Tree and Random Forest.