**Introduction**

**Data Description**

For the final, I will continue analyzing my midterm dataset about world happiness. Data for this final is sourced from Kaggle in CSV format, providing information on countries' happiness through a ladder score (happiness scale) and each country's scores in categories that can affect a country's happiness score.

After my pre-analysis of my dataset, I looked at the various graphs and charts of my data. The question that stood out to me was what determines a country's happiness level, which led me to look to model a country's happiness score. This is how I determined my model question below.

By the way, throughout this paper, I refer to the ladder score as a happiness score and use it interchangeably.

**Predictive Task**

Model Question: How can we predict a country's happiness score (ladder score) based on socioeconomic and health factors to identify the country with the highest score?

Breakdown of question

1. Target Variable:

* Ladder score

2. Predictors (Features):

* Log GDP per Capita
* Social support
* Healthy life expectancy
* Freedom to make life choices
* Generosity
* Perceptions of corruption
* Regional indicator

**Summary Findings**

After analyzing my dataset, I found that computing regression models are the most suitable choice. Regression models predict continuous values, and since the ladder score and all other feature scores are continuous values, the regression model will be appropriate. Although classification could divide countries into groups of high to low happiness scores, this approach doesn't help directly predict the highest ladder score country. This way, it would simplify the prediction, which loses precisions. Times series also wouldn't work because a time series model predicts trends over time. Still, the dataset I chose appears static and does not include time series data, meaning time series wouldn't be applicable.

**Modeling Data**

**Overview of Modules**

I will use various regression models to predict the ladder score. Multiple models will allow me to understand better the relationships between the ladder score (dependent variable) and the features/predictors (independent variables). For each model, I will use an 80-20 train test split, training 80% of my data and then testing the remaining 20%.

I will start with a baseline model so I have a baseline against which to compare my additional models later. My baseline model will be a mean squared error, focusing on minimizing significant errors. Interpreting MSE: The lower the MSE, the more accurate the model is. For example, an MSE of 0 means perfect predictions with no error. I will compare my various MSEs with my baseline MSE to determine the effectiveness of my models.

To specify, my data's original baseline MSE is 1.38. I will compare all my models to this baseline, interpreting each result and its accuracy.

**Implementation**

I will first start with a multiple regression model to use a baseline to assess the linear relationship between features and the target variable (ladder score). Next, I will use decision trees to visualize feature splits. This model works well for understanding nonlinear relationships between features and the target variable, helping me explore the nonlinear relationship of my data. I also used a random forest model, which combines multiple decision trees to reduce overfitting and improve predictive accuracy. I will also use a K–Nearest Neighbors model to discover the average ladder score of the nearest neighbors. Finally, I used the gradient boosting model for my additional model, which wasn’t discussed in class. This model sequentially builds trees to correct errors made by previous ones. I chose this for its ability to handle complex patterns and wanted to see if my dataset had any of those patterns.

Next, I will discuss and compare the results of each of my models.

**Results and Interpretation**

**Multiple Regression Model:**

**Intercept:** After constructing the model, the intercept is 2.05, representing the predicted outcome when all features have a zero value. This means that the model predicts a happiness score of 2.5 when all the features are zero. This intercept is not intuitive because it is unrealistic that every feature is zero, but this simply gives us a starting point.

**Categorical Features (Regional Indicators):** The coefficients represent the impact of each feature on the target variable. First, look at the regions. The reference region I am using is Central and Eastern Europe, meaning that the coefficients, whether high or low, depend on the average happiness score of Central and Eastern Europe. Starting with the Commonwealth of Independent States, it has a feature importance score of -0.403, showing that, on average, happiness scores are 0.403 lower than the reference region. East Asia has a score of -0.216, meaning happiness scores are 0.216 lower than those of the reference region. Latin America and the Caribbean have a score of 0.318, meaning their happiness scores are 0.318 compared to the reference region. The scores of the rest of the regions have the same meaning. Middle East and North Africa score is -0.370, North America and ANZ score is 0.129, South Asia score is -1.094, Southeast Asia is -0.478, Sub-Saharan Africa is -0.333, and Western Europe is 0.101.

**Numerical Features:** These features represent quantitative predictors, and their coefficients indicate the expected change in the target variable for a one-unit increase in the feature. Log GDP per capita is 0.616, meaning that for every unit increase in log GDP per capita, the happiness score will increase by 0.616. This is the same concept for the rest of the features. Social support score is 0.810, healthy life expectancy is 0.902, freedom to make life choices is 1.665, generosity is 1.177, and perceptions of corruption is 0.999.

These results show that the ladder score's most positive feature is the freedom to make life choices, at 1.655. The region with the most negative impact is South Asia, at -1.094. These results are not necessarily surprising, but some questions are raised. For example, looking at the feature that affects happiness the most, although freedom to make life choices is essential, other features like log GDP per capita and healthy life expectancy should play a more significant role.

**Mean Square Error:** The MSE train value is 0.1907, and the test data is 0.309. I calculated both the test and train MSE to ensure that the data did not overfit and that there was an accurate MSE. We want to look at MSE for test data because the ultimate goal is to see how well this model performs on unseen data. An MSE of 0.309 indicates that, on average, the squared diffferences between the predicted and actual ladder scores are approximately 0.235 units squared. Compared to the baseline MSE of 1.385, this is already lower, and considering the ladder score’s range is 0-10, we can say that the model is performing well.

Since we got a low MSE using a linear model, it indicates the relationship between the predictor's Log GDP per capita, Social support, etc., and the target variable ladder score is at least partially linear. This is expected because factors like GDP, social support, and healthy life expectancy are known to be strongly correlated with happiness.

While these results are solid, there may still be room for improvement if I explore nonlinear models like Random Forests, which I will do next.

**Random Forests Model**

**Categorical Features:** I have also found the region's feature coefficients, such as the multiple regression model above. Again, I will be using the reference region of Central and Eastern Europe to stay consistent and be able to compare the two. Commonwealth Independent states score is 0.000599, East Asia is 0.001509, Latin America and Caribbean is 0.002684, Middle East and North Africa is 0.003084, North America and ANZ is 0.000310, South Asia is 0.009287, Southeast Asia is 0.0000765, Sub Saharan Africa is 0.001650, and Western Europe is 0.000418.

**Continuous Features:** This is the same as how the feature coefficients affect the ladder score. Log GDP per capita is 0.172, social support is 0.563, health life expectancy is 0.102, freedom to make life choices is 0.894, generosity is 0.026, and perceptions of corruption is 0.025.

The top continuous features are social support (0.563), log GDP per capita (0.172), healthy life expectancy (0.103), and freedom to make life choices (0.089). The regions all show very low importance. These features and stats make more sense, as we still see the importance of social support, but log GDP and healthy life expectancy features are also more important than the multiple regression model. This data also seems more accurate for the region factors as these regions are massive and consist of various countries, so the ladder score shouldn’t differ as much. These findings are not surprising. The only thing that is somewhat surprising is the low importance of generosity and perceptions of corruption, as these were highly significant in the multiple regression model.

**MSE**: I will be analyzing the MSE of the testing data now. An MSE of 0.0433 backs my predictions and analyzes the coefficients I was analyzing above. They suggest that the random forest model is much better at capturing the patterns in the dataset and making more accurate predictions. It significantly outperforms the multiple regression model, telling us that the dataset also has nonlinear relationships.

The datasets above provide contradictory information on the region's importance, so I will use KNN to explore this further.

**K-Nearest Neighbors Model**

**Actual vs. Predicted Values:** A scatter plot shows the actual ladder scores compared to KNN’s predicted values. The blue dots represent the actual happiness scores, and the orange dots represent the happiness scores predicted by the KNN model. Analyzing the plot itself, we see an alignment between actual and predicted values, but these are not perfect. The predicted models are relatively close to the blue dots, indicating that it is still a good prediction.

**MSE**: The MSE of this model is 0.334, representing the average squared difference between actual and predicted happiness scores. An MSE of 0.334 means that the predictions deviate moderately from the exact amount of ladder scores. KNN MSE has a slightly higher MSE than multiple regression but a much higher MSE than random forests, but is still significantly better than the baseline model.

This tells us that the local similarity of data points will have similar ladder scores. For example, countries with similar log GDP per capita, social support, and healthy life expectancy will have similar target scores. However, its MSE being a lot higher tells us that countries with comparable data don’t always produce the same latter score.

Because of this, I wanted to revisit and examine more complex and nonlinear relationships, which led to my next model of decision trees.

**Decision Tree Model**

**Best Parameter:** I used Grid Search to find the optimal maximum depth of the tree, which was 3. This is considered a low depth, indicating that the data is not incredibly complex. A shallow tree prevents overfitting and helps the model generalize better to unseen data.

**Categorical Features**: Again, it is the same as we’ve done before. The regional areas all have a score of 0 except for Latin America and the Caribbean, with a score of 0.002207, and South Asia, with a score of 0.019263. This tells us that regions don’t affect the score except for those regions mentioned above. They barely affect the ladder score.

**Continuous Features:** Log GDP per capita is 0.139, social support is 0.658, healthy life expectancy is 0.0556, freedom to make life choices is 0.1023, generosity is 0.0219, and perceptions of corruption are 0.0015.

The key features for continuous variables are social support and log GDP per capita, but social support is the most essential feature. And it also tells us that the regions don’t matter at all. Although not fully surprising, I feel these results might be stretched. Although region shouldn’t affect happiness significantly with the number of countries in each region, I feel there is still an effect, not necessarily 0, like the model predicted. For the continuous features, although social support is a crucial factor, I’m not sure that the significant difference is representative of the dataset itself.

**Decision Tree:** The decision tree diagram shows the structures of the splits and the rules it uses to predict the ladder score. Although the visualization is hard to interpret due to its small sizing, after downloading the image, it becomes clearer. The root node (top of the tree) contains the first split, which is the regional indicator, and in this case, it starts with North America and ANZ. This means that the decision tree first decides whether the data point belongs to this region and uses it to separate the dataset. Next, each internal node represents a decision based on a feature like GDP per capita, social support, etc. Then, the final nodes represent the final predictions, consisting of each predicted happiness score. From this, it tells us that the tree relies heavily on regions and that the numerical features refine predictions.

**MSE:** The MSE of 0.494 suggests that this model is just a moderate performance compared to the rest. This score makes sense as we conducted a random forest model above, a more robust version of the decision tree. This also tells us that a more complex model is needed to explain how the MSE of random forests was so good. Still, the MSE of the decision tree significantly outperformed the baseline but underperformed compared to the KNN and multiple regression.

**Gradient Boosting Model:**

Gradient Boosting model is a machine learning method that incrementally builds a model by combining several weak learners' predictions, typically decision trees. Unlike random forests, which build trees independently, gradient boosting builds them sequentially, each one correcting the errors of the previous ones.

**Categorical Features**: The region's scores are the same as the scores of the decision tree, with all the features being 0 except Latin America, the Caribbean region, and South Asia.

**Continuous Features:** The continuous features are also identical to the decision trees.

This is surprising initially, but understanding how gradient-boosting regression works makes this clear. Decision trees are used as the model's core structure in both cases.

**Residual Plot:** The plot shows the difference between actual and predicted values. The residuals appear scattered around the horizontal line, meaning a residual of 0, which is the ideal result. However, some residuals are significantly above or below 0, indicating that the model does struggle with certain data points. The spread of the residuals tells us that although the model performs moderately well, there is still room for improvement.

**MSE:** The MSE is 0.350, indicating that the model performs moderately well and aligns with other models like KNN and multiple regression models. However, this model still performs significantly worse than the random forest model, likely meaning that the dataset is more suitable for average than the random forest model.

**Conclusion:**

**Key Findings**  
In analyzing world happiness data, I aimed to predict a country’s happiness score based on various factors using multiple regression models. Each model offered valuable insight into the relationship between the target and predictor variables. It became evident that certain features like social support, log GDP per capita, and freedom to make life choices consistently remained significant contributors to happiness. This aligns with pre-existing knowledge and research that these factors are important determinants in shaping happiness.

The random forest model performed the best with an MSE of 0.0433, far surpassing other models. This suggests the dataset contains nonlinear relationships and complex patterns that random forests can capture. Multiple regression was performed with an MSE of 0.309, indicating that the predictors have at least a partially linear relationship with the target variable. The KNN model could have been more accurate, showing that the importance of the local pattern may have been insignificant with an MSE of 0.334. The decision tree and gradient boosting models performed with MSEs of 0.494 and 0.350 highlight that the model could experience overfitting with gradient boosting, looking to build trees off each other.

Relating this back to the question: How can we predict a country's happiness score (ladder score) based on socioeconomic and health factors to identify the country with the highest score? The best way to predict this is through the random forest model.

**Implications**

The results reinforce the critical role of economic and social factors in determining happiness. So policymakers looking to improve national happiness levels should prioritize enhancing social support systems, increasing economic opportunities, and promoting individual freedoms.

**Future Analysis**

To improve the analysis further, I can experiment with deeper trees to capture more complex patterns and avoid overfitting. I can test various depths to see if this will improve my decision tree models. I can also change my reference variable region to see how the regions vary around a different region, providing more analysis of my research. I can also manipulate my features by combining more important features like GDP per capita and healthy life expectancy to verify their importance and to see if new variables provide more predictive power. Finally, I can also pull the world happiness score report from previous years and analyze those as well to test the generalization of the models and ensure accurate modeling in diverse datasets.