**Predicting Happiness: A Look at Global Well-Being Through Machine Learning**



This article explores a project that utilizes machine learning to delve into the World Happiness Report dataset. This report, published annually, ranks countries based on their citizens' perceived happiness. By analyzing this data, we can gain valuable insights into the factors that contribute to national well-being. However, the key challenge lies in predicting happiness scores without simply relying on the sum of contributing factors. This project aspires to move beyond a basic summation by uncovering the underlying relationships between various social, economic, and cultural indicators, and how they interact to influence a nation's happiness level. Ultimately, this knowledge can empower policymakers and researchers to design and implement strategies that cultivate greater well-being for all.

1. **Problem Definition**

The objective of this project is to develop a machine learning model that accurately predicts a country's happiness score based on various social and economic indicators. We aim to move beyond a simple summation of these factors by identifying the underlying relationships between them and happiness. This knowledge can prove valuable for policymakers and researchers seeking to understand and improve global well-being.

1. **Data Analysis**

The World Happiness Report dataset provides crucial information for our analysis. It includes data points for over 150 countries, with the following key columns:

* **Happiness Score:** Overall happiness level of a country (target variable).
* **GDP per Capita:** Economic well-being indicator.
* **Family:** Importance placed on family life.
* **Life Expectancy:** Average lifespan in the country.
* **Freedom:** Level of personal freedom experienced by citizens.
* **Generosity:** How charitable and giving the population is.
* **Trust (Government Corruption):** Perception of government corruption
* **Dystopia Residual:** First **dystopia** refers to an imaginary country that has the world’s least-happy people. The purpose in establishing Dystopia is to have a benchmark against which all countries can be favorably compared (no country performs more poorly than Dystopia) in terms of each of the six key variables, thus allowing each sub-bar to be of positive width. The lowest scores observed for the six key variables. The **residuals**, or unexplained components, differ for each country, reflecting the extent to which the six variables either over- or under-explain average life evaluations. These residuals have an average value of approximately zero over the whole set of countries .

The Dystopia Residual metric actually is the Dystopia Happiness Score (1.85) + the Residual value or the unexplained value for each country.

Our initial analysis focuses on understanding the relationships between these factors and the happiness score. This may involve techniques like correlation analysis and visualization to identify potential trends and patterns. In data analysis, we have paid attention to the graphical analysis of the dataset, with the main aim of finding the relationship between each variable i.e. features and how are they related to the target variable which is the happiness score.

In EDA, all the columns and their datatypes are verified whether they are correct or not. If there is any mismatch in the datatype the datatype is changed manually. Then, we check if there are any null values or not. If there are null values, if the column is categorical, the Nan values are replaced by the mode of the column and if they are numerical, the average of the column takes the place of the Nan values. In this dataset there are no null values.

Next comes the **visualization of dataset,** Data analysis is a powerful tool, but raw numbers and tables can be overwhelming and difficult to interpret. This is where data visualization comes in - it's the art of transforming data into visual representations like charts, graphs, and maps. These visuals make trends, patterns, and relationships within your data clear and easily understandable. This is important because it helps to unveil hidden relationship, simplify complex information, gives more clarity and helps in decision making.

Followed by visualization, we also perform **outlier detection test**, because as these outlier points fall significantly outside the pattern of the dataset, they can disturb the result of the overall model and can reduce the accuracy. So, in this dataset outliers were handled using z-score method and there were only 3 entries where z value was more than 3(threshold value), so decided not to remove them as they might not pose a big threat to the model. Then we performed **skewness test** also to be on the safer side because both positively and negatively skewed data can be a big threat and has the capacity to change the accuracy of the model. Usually, the acceptable range of skewness is -0.52 to +52, but here there were three features where they had skewness not falling between this. As a result, positive skewness was treated using boxcox method and negative using square method.

Finally, EDA is completed by finding **correlation** and heatmap of the correlation between the features and the target variables. In that we found that the column happiness rank is highly negatively correlated with the target variable so decided to drop that column from the dataset for further analysis.

1. **EDA Concluding Remarks:**

Exploratory Data Analysis (EDA) helps us understand the data's distribution, identify outliers, and uncover potential relationships between variables. Exploratory Data Analysis (EDA) provided valuable insights into the World Happiness Report data. We observed anticipated correlations, like positive associations between happiness and factors like GDP per capita and life expectancy. Conversely, corruption likely has a negative impact on happiness. The importance of family life and generosity might show more nuanced relationships depending on cultural contexts. These initial findings set the stage for further exploration. By delving deeper into potential interactions between variables, we can move beyond simple correlations and identify the underlying factors that truly contribute to national happiness. Here are some expected key findings:

* The dataset is mainly focused on sub-Saharan Africa followed by central -Eastern Europe and Western Europe.
* South-east Asia and Southern Asia contributes to the highest GDP.
* High happiness score relatively corresponds and contributes to more economic GDP.
* There is a gradual and visible increase in the health expectancy when there is high happiness score.
* Dystopia residual has a positive relationship with the target variable happiness score.
* Happiness rank is highly negatively correlated with the target variable so decided to drop that column from the dataset for further analysis.

1. **Pre-processing Pipeline:**

In the world of machine learning, data pre-processing acts as the foundation for best model development. This crucial step meticulously prepares the data, analogous to a scientist carefully calibrating instruments before an experiment. The pipeline addresses missing values, ensuring no crucial information is absent. It then standardizes features, akin to ensuring all measurements are on the same scale for accurate comparisons. Finally, it transforms non-numerical data into a format interpretable by the model, similar to converting qualitative observations into quantitative data points. This meticulous process guarantees the data's quality and consistency, empowering the model to learn effectively and deliver trustworthy predictions.

Pre-processing of data is a critical step in machine learning model development that involves transforming raw data into a format that is suitable for analysis and modeling. The pre-processing pipeline typically includes several key steps aimed at addressing issues such as feature scaling, categorical variable encoding.

The features of the dataset might have different scale, if we try to build a model with this, there might reduction in the accuracy. So that is why feature scaling is important in the dataset. Here the datasets are preprocessed using standard scalar method.

Machine learning algorithms typically require numerical input, necessitating the encoding of categorical variables into a numerical format. Common encoding techniques include:

•One-Hot Encoding: Converting categorical variables into binary vectors, with each category represented by a binary indicator variable.

•Label Encoding: Replacing categorical labels with numerical values, assigning a unique integer to each category.

•Target Encoding: Encoding categorical variables based on the target variable's mean or other statistical measures, capturing the relationship between the categorical variable and the target.

In this dataset the Label encoding method is implemented.

1. **Building Machine Learning Models:**

Finally with this clean dataset then comes the task of building a relevant and high accuracy machine learning model. The first step involved strategically splitting the data into two distinct sets: a training set and a testing set. The training set serves as the model's learning ground, where it identifies patterns and relationships between the various factors (e.g., GDP per capita, life expectancy) and the target variable - the happiness score. The testing set, untouched by the model during training, acts as a real-world simulation. By testing the model's performance on unseen data, we ensure it doesn't simply memorize the training examples but can effectively apply its knowledge to predict happiness scores for new countries.

Next, we unleashed a diverse arsenal of machine learning models onto the data. Here are the key players in this exploration:

**Linear Regression:** This fundamental model establishes a linear relationship between the happiness score and other factors. It depicts as a straight line drawn through the data, attempting to capture the overall trend. The results are as follows:

r2\_score is 0.998145333682481

r2\_score of training is 0.9953083709032756

mean\_absolute\_error is 0.04202606129370167

mean\_squared\_error is 0.002457228528182295

root\_mean\_squared\_error is 0.04957044006444057

**Random Forest:** This powerful technique combines the wisdom of multiple decision trees. Each tree makes its own prediction based on a subset of features, and the final score is a democratic vote, the average prediction of all the trees. This approach helps overcome the limitations of any single tree and leads to more robust predictions. The results are as follows:

r2\_score is 0.9311930671714193

r2\_score of training is 0.984334701839045

mean\_absolute\_error is 0.23439416666666688

mean\_squared\_error is 0.09116160502083326

root\_mean\_squared\_error is 0.3019298014784782

**K-Nearest Neighbors (KNN):** This intuitive method predicts a country's happiness score based on the happiness scores of its "nearest neighbors" in the training data. Imagine comparing a new country to its most similar counterparts - their happiness scores offer valuable clues about the new country's potential happiness level. The results are as follows:

r2\_score is 0.8920189449676235

r2\_score of training is 0.901800961483633

mean\_absolute\_error is 0.2977291666666666

mean\_squared\_error is 0.14306300083333334

root\_mean\_squared\_error is 0.3782366994797482

**LASSO and Ridge Regression:** These regularization techniques are like personal trainers for our models. They prevent the model from becoming overly complex and overly reliant on specific features in the training data. This helps to avoid overfitting, a scenario where the model performs well on the training data but struggles to generalize to unseen data. The results are as follows respectively :

r2\_score is -0.0015487464526167116

r2\_score of training is 0.0

mean\_absolute\_error is 0.9636886363636364

mean\_squared\_error is 1.3269417408953168

root\_mean\_squared\_error is 1.1519295728886019

r2\_score is 0.9980659673036202

r2\_score of training is 0.9952774423522255

mean\_absolute\_error is 0.042022444335900355

mean\_squared\_error is 0.00256238023578226

root\_mean\_squared\_error is 0.050619958867844414

**Extra Trees:** This ensemble method shares similarities with Random Forest but utilizes a slightly different approach for generating decision trees. Both techniques aim to leverage the collective wisdom of multiple trees for improved accuracy. The results are as follows:

r2\_score is 0.9341497382094125

r2\_score of training is 1.0

mean\_absolute\_error is 0.21485708333333342

mean\_squared\_error is 0.08724434165416674

root\_mean\_squared\_error is 0.2953715315567273

Scatterplot has been plotted for all the models comparing the y\_test and the predicted y variable for the model.

A 5-fold cross-validation technique is followed after building the model. Here's the idea: the training data is shuffled and divided into five folds. The model is trained on four folds and tested on the remaining fold. This process is repeated five times, ensuring each fold is used for testing once. This rigorous approach provides a more reliable estimate of the model's generalizability compared to a single train-test split. By looking at the scores of cross validation, we can saw that the linear regression has the least difference and that is our best fitting model.

Finally, the model is optimized using hyperparameter tuning. This identifies the key and the best parameters that would attribute to the best accuracy of the model.

1. **Conclusion:**

By employing machine learning, this project transcends simply summing factors to predict happiness scores. The resulting model can delve deeper, uncovering the relative importance of each factor and how they interact to influence national happiness. This knowledge transcends mere correlation, transforming into actionable insights. Armed with this deeper understanding, policymakers can craft national policies that prioritize the factors with the most significant impact on citizens' well-being, fostering a data-driven approach to building happier nations.