SOCI354 Final Project Written Report

Dila Gürer
M. Fuat Kına

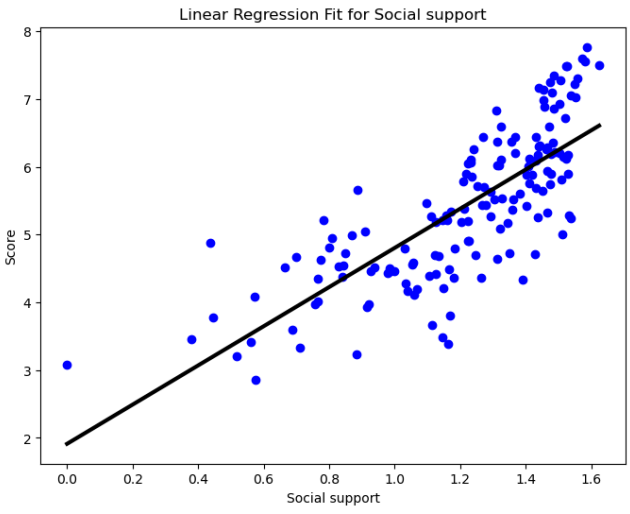
8th of June, 2024

Research Purpose: Aim of this project is to analyze the relationship between happiness of people in a given country and different factors in the country including the level of social support, being in the Europe continent. However, I also aim to delve into this relationship by taking into account the other factors that might have an impact on the happiness levels such as GDP per capita, freedom to make life choices, healthy life expectancy, perception of corruption and generosity. For those purposes, data from Kaggle is used, link is provided below:
https://www.kaggle.com/datasets/priyanka841/2019-world-happiness-report-csv-file
In the mapping of countries and happiness scores on the world map, https://www.naturalearthdata.com/downloads/110m-cultural-vectors/ site is used to get the shapefile.

ANALYSIS 1: LINEAR REGRESSION: Happiness level in relation to social support level in the country

Hypotheses:
H0(1): There is no significant relationship between happiness scores and level of social support in a country.
H1(1): There is a significant relationship between happiness scores and level of social support in a country: Increase in social support is associated with an increase in happiness scores in a country.

Model Evaluation:
1. Mean Squared Error: 0.98: A lower MSE indicates a better fit of the model to the data. In this case, an MSE of 0.98 suggests that on average, the squared difference between the predicted and actual happiness scores is around 0.98 units.
2. Coefficients: [0.77537163 1.12419158 1.07814273 1.45483237 0.97228022 0.48978335]: ***For every one-unit increase in social support, the predicted happiness score increases by approximately 1.1477.H1 is accepted.***



3. R^2 Score: 0.7368467575719668: A good correlation between the predictors and the happiness score is suggested by an R2 value of roughly 0.737, which indicates that the model accounts for 73.7% of the variation in the happiness score.
4. Cross-Validation Scores: [-5.57285067 -5.9477523 -6.23988074 -8.84930385 -2.39521998]: The model's poor performance on these validation sets is shown by the negative values, which may be the

result of data variability or overfitting on the training set.

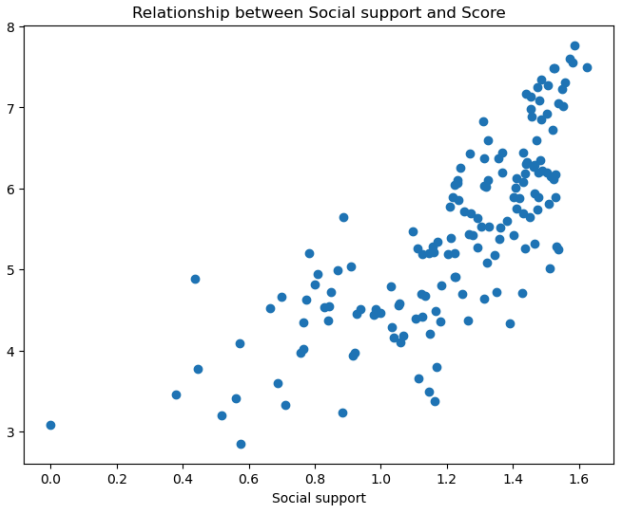
5. Mean Cross-Validation Score: -5.801001509838042: The model's average predictions are found to be substantially different from the actual values in the cross-validation folds, as shown by a mean score of about -5.801.

6. Root Mean Squared Error: 0.6386463842144325: An RMSE of approximately 0.639 indicates that, on average, the model's predictions differ from the actual happiness scores by around 0.639 units.

7. Lasso Coefficients: [1.35429473 0.40532089 0. 0. 0. 0.]:This indicates that only the first two predictors—possibly the most significant ones—have non-zero coefficients, indicating that the Lasso model only takes these two predictors into account when attempting to explain the happiness score.

8. best alpha for Lasso: 0.005815372239742716: The best alpha value of approximately 0.0058 suggests that the model performs best with minimal regularization

9. R^2 Score (Lasso): 0.6664963334361766 : about 66.0% of the variance in the happiness score is explained by the predictors.



10. Best alpha for Ridge: 0.6280291441834259:The degree of regularization in Ridge regression is determined by the alpha parameter. The model performs better with a moderate level of regularization, as indicated by the best alpha value of roughly 0.628.

11. R^2 Score (Ridge): 0.6673941235948362: The factors account for roughly 66.7% of the variance in the happiness score. This score indicates a moderate level of regularization that balances bias and variance; it is greater than the Lasso model but lower than the linear regression model.

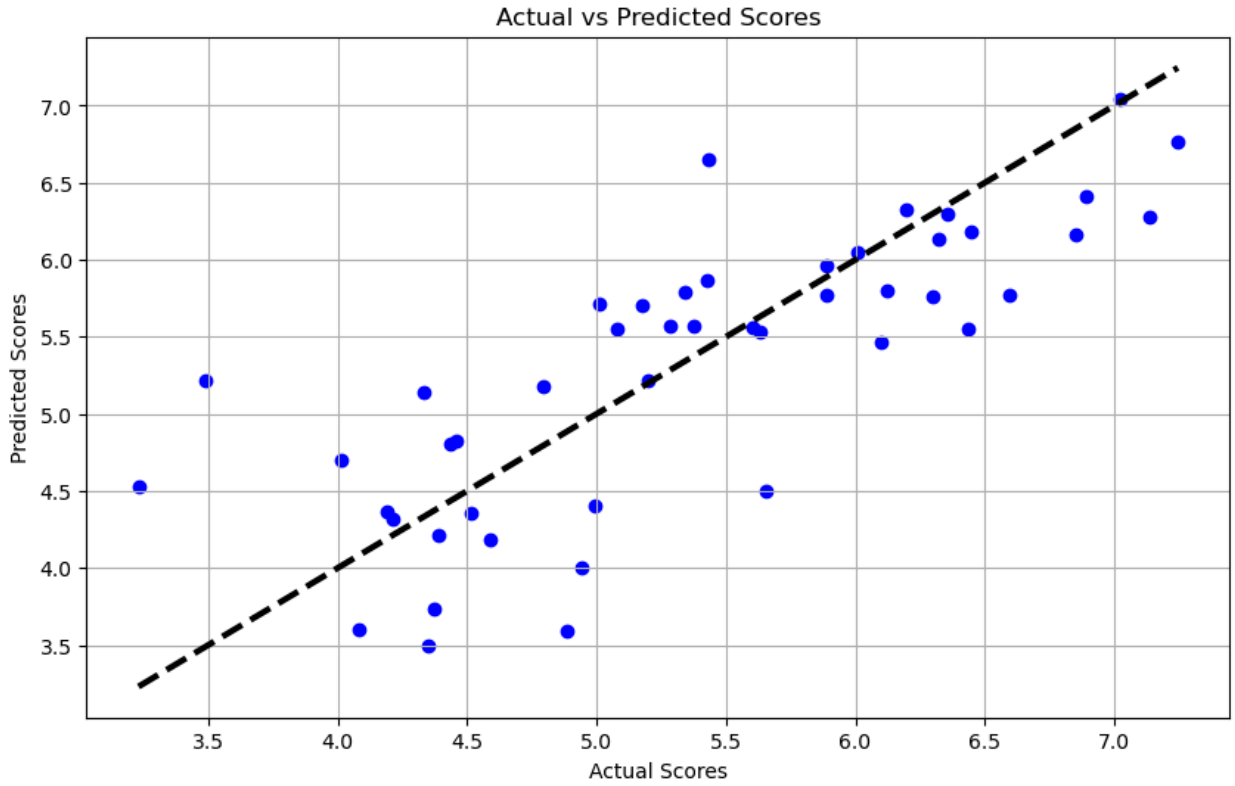
12.

A considerable amount of the variation (73.7%) is explained by linear regression, which has a surprisingly low RMSE (0.639). However, cross-validation scores point to possible overfitting or variability problems.

Lasso Regression: This method selects variables by reducing some coefficients to zero, which lowers the R2 value to 52.0% and highlights the most significant predictors. Minimal regularization is suggested by the best alpha value.

Ridge Regression: This model achieves a moderate R^2 score (66.7%) and an ideal alpha value, which indicates moderate regularization, while balancing the trade-off between variance and bias. While still not as good as the unregularized linear regression, this model outperforms Lasso.

Every model has advantages and disadvantages. Lasso uses regularization to highlight significant predictors, while Ridge strikes a balance between the two methods. Linear regression accounts for the majority of variation but has the potential to overfit.



OLS Regression Results

An OLS regression using the statsmodel library on the full dataset yielded an R2 value of 0.791. This number indicates the degree to which the dataset is fully fitted by the model, without any division into training and testing sets.

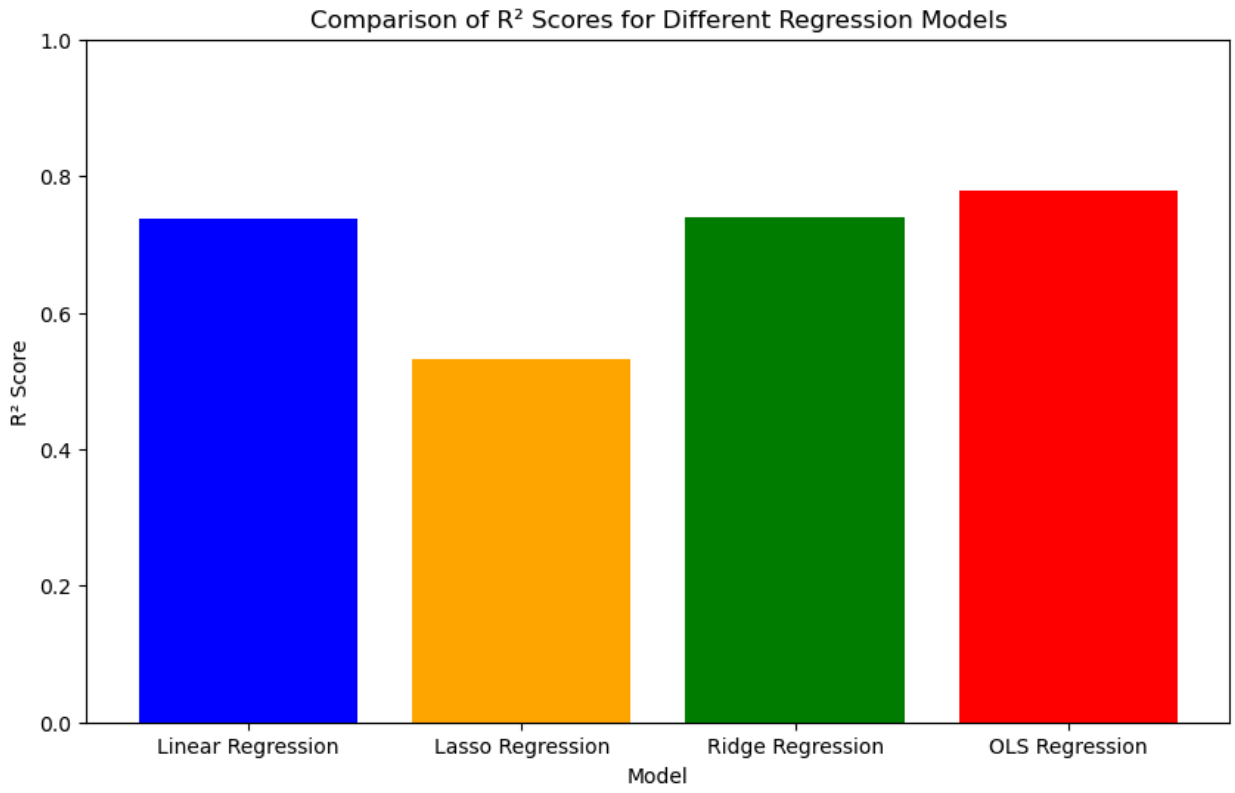
Social support:

- **Coef:** 1.1477
- **P-value:** 0.000

- Interpretation: A one-unit increase in social support is associated with an increase of approximately 1.1477 in the happiness score. This predictor is highly significant.

Analysis of the Plot

- The fit of the linear regression model to the test data is indicated by linear regression.
- Lasso Regression: Shows how well the Lasso model performs when regularized with L1, usually yielding a lower R2 because of its feature selection property.
- Ridge Regression: Shows how well the Ridge model performs when regularized with L2, striking a balance between regularization and fit.
- OLS Regression: Uses the training data in the evaluation process, so it reflects the model fit throughout the full dataset, which is typically greater.



<u>ANALYSIS 2:</u> LOGISTIC REGRESSION: Happiness level in relation to social support level
Hypotheses:
H0(1): There is no significant relationship between being happy & unhappy and the level of social support in a country.

H1(1): There is a significant relationship between being happy & unhappy and the level of social support in a country: Increase in social support is associated with an increase in the likelihood of being happy compared to being unhappy.

Best Hyperparameters: {'alpha': 1}: This indicates that the optimal alpha value for the model was found to be 1. This hyperparameter controls the strength of the regularization in the model.
Mean Squared Error on Test Set: 0.34730821508971366: This calculates the average squared difference between the actual outcomes that were observed and the outcomes that the model anticipated. A better match is indicated by a lower MSE.

Accuracy: 0.851063829787234
Confusion Matrix:
[[31 1]
[6 9]]

Classification Report:

	precision	**recall**	**f1-score**	**support**
0	0.84	0.97	0.90	32
1	0.90	0.60	0.72	15
accuracy			0.85	47
macro avg	0.87	0.78	0.81	47
weighted avg	0.86	0.85	0.84	47

Cross-Validation Scores: [0.6875 0.96774194 0.87096774 0.87096774 0.74193548]
Mean Cross-Validation Score: 0.8278225806451613

The range of cross-validation scores, which show performance variations across several data folds, is 0.6875 to 0.96774194.

The average cross-validation score, or 0.8278225806451613, shows how well the model performs on average over all folds. This number represents the model's overall success in capturing the correlation between the predictors and the dependent variable.

Overall, the cross-validation scores indicate that the logarithmic regression model performs fairly well, with a mean score of roughly 0.83, indicating that it can generalize well to fresh data.

Mean Squared Error (MSE): 0.10638297872340426
Root Mean Squared Error (RMSE): 0.3261640365267211

In this instance, the squared difference between the expected and actual numbers is roughly 0.106 on average, according to the MSE value of 0.10638297872340426.

The average magnitude of the mistakes in the predicted values is represented by the Root Mean Squared Error (RMSE), which is the square root of the Mean Squared Error (MSE). The average difference between the predicted and actual values is about 0.326, according to the RMSE value of 0.3261640365267211.

In conclusion, the MSE and RMSE both shed light on how accurate the model's predictions are. Better performance is indicated by lower MSE and RMSE values, which imply that the model's predictions are more accurate.

Score for Pseudo R2: 0.5115876602606143 A statistical model's goodness of fit is gauged by the pseudo R2 score, which is very useful in regression analysis. It shows the degree to which the variance in the dependent variable can be explained by the independent factors. With a Pseudo R2 score of 0.5115876602606143 in your case, the independent variables in the model account for around 51.16% of the variance in the dependent variable..

Lasso Logistic Regression Accuracy: 0.8936170212765957
Lasso Logistic Regression Confusion Matrix:
[[32 0]
[5 10]]

Lasso Logistic Regression Classification Report:

	precision	**recall**	**f1-score**	**support**
0	0.86	1.00	0.93	32
1	1.00	0.67	0.80	15
accuracy			0.89	47
macro avg	0.93	0.83	0.86	47
weighted avg	0.91	0.89	0.89	47

Best alpha for Lasso: 0.0011719128787983544: Indicates that the optimal alpha for Ridge regression was found to be 0.1.

R^2 Score (Lasso): 0.736525633365397: The R^2 score for the Lasso regression indicates that approximately 73.65% of the variance in the dependent variable (target) is predictable from the independent variables (features) using the Lasso regression.

Ridge Logistic Regression Accuracy: 0.8723404255319149

Ridge Logistic Regression Confusion Matrix:

```
[[32  0]
 [ 6  9]]
```

Ridge Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.84	1.00	0.91	32
1	1.00	0.60	0.75	15
accuracy			0.87	47
macro avg	0.92	0.80	0.83	47
weighted avg	0.89	0.87	0.86	47

Best alpha for Ridge: {'alpha': 0.1}

R^2 Score (Ridge): 0.7386395153144539: Similar to Lasso, the R^2 score for Ridge regression indicates that approximately 73.86% of the variance in the dependent variable is predictable from the independent variables using the Ridge regression.

Ridge Classification Report using Logistic Regression:

Precision can be defined as the ratio of accurately predicted positive observations to the total number of positive predictions. Class 0 precision is 0.84, while Class 1 precision is 1.00.

Recall: The proportion of all observations in the actual class that were accurately predicted to be positive. Class 0 recall is 1.00, and Class 1 recall is 0.60.

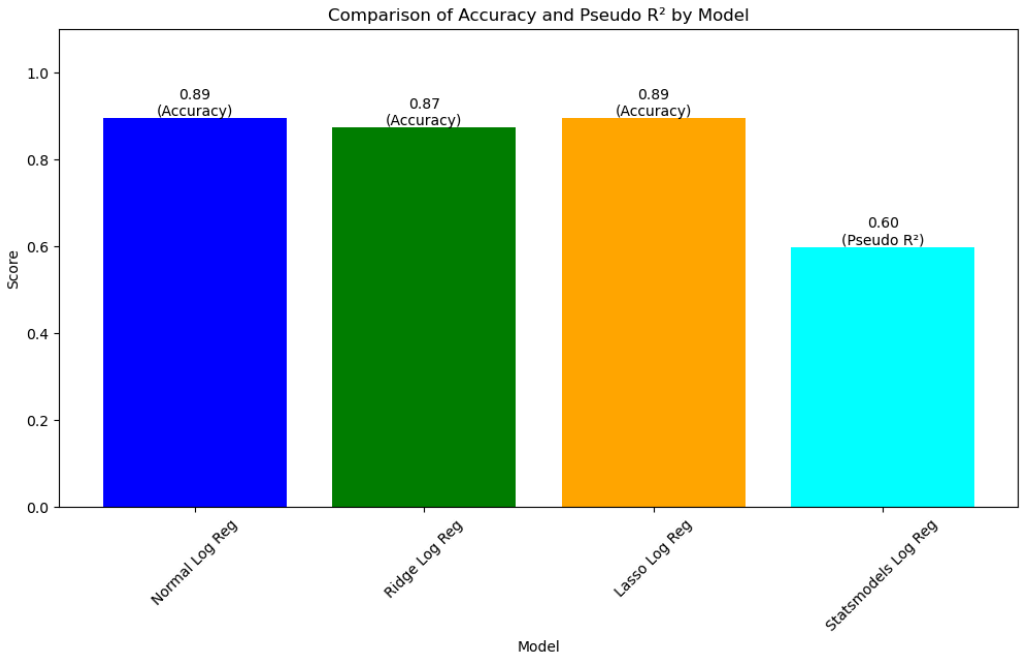
The precision and recall weighted average is known as the F1-score. Class 0's F1-score is 0.91, whereas Class 1's is 0.75.

The accuracy of Lasso Logistic Regression, measured as the percentage of correctly classified occurrences (true positives and true negatives) relative to all instances, is 0.8936170212765957.

Confusion Chart:

Precision can be defined as the ratio of accurately predicted positive observations to the total number of positive predictions. Class 0 precision is 0.86, while Class 1 precision is 1.00.

Recall: The proportion of all observations in the actual class that were accurately predicted to be positive. Class 0 recall is 1.00, whereas Class 1 recall is 0.67.
The precision and recall weighted average is known as the F1-score. Class 0's F1-score is 0.93, whereas Class 1's is 0.80.



Overall Interpretation: For Class 0 (negative class), the Lasso and Ridge logistic regression models exhibit good accuracy, precision, and recall; however, the recall for Class 1 (positive class) is lower, suggesting some misclassification in the positive class prediction. For the positive class, the Lasso model seems to perform marginally better than the Ridge model in terms of accuracy and recall.

AUC: 0.92: An AUC of 0.92 indicates that there is a 92% chance that the classifier will be able to distinguish between positive and negative classes correctly. This is a high value, suggesting that the model has excellent performance in distinguishing between the two classes.

Iterations 9

Logit Regression Results

Dep. Variable:	BinaryScore	No. Observations:	156
Model:	Logit	Df Residuals:	149
Method:	MLE	Df Model:	6
Date:	Fri, 07 Jun 2024	Pseudo R-squ.:	0.5955
Time:	00:51:15	Log-Likelihood:	-40.161
converged:	True	LL-Null:	-99.296
Covariance Type:	nonrobust	LLR p-value:	3.759e-23

	coef	std err	z	P>|z|	[0.025	0.975]
const	-21.8282	4.639	-4.705	0.000	-30.921	-12.735
GDP per capita	3.4464	1.642	2.099	0.036	0.228	6.665
Social support	6.2111	2.491	2.493	0.013	1.329	11.093
Healthy life expectancy	5.7332	3.047	1.881	0.060	-0.240	11.706
Freedom to make life choices	10.9483	3.388	3.232	0.001	4.309	17.588
Perceptions of corruption	-4.7857	4.683	-1.022	0.307	-13.964	4.392
Generosity	1.4245	3.395	0.420	0.675	-5.229	8.078

Pseudo R-squared: 0.5955

- This indicates that approximately 59.55% of the variability in the binary outcome is explained by the model. While pseudo R-squared values are not directly comparable to R-squared values in linear regression, this is a relatively good fit for a logistic regression model.

The log-likelihood measures the fit of the model. A higher (less negative) log-likelihood value indicates a better fit. In this context, comparing it to the LL-Null (-99.296) shows a significant improvement.

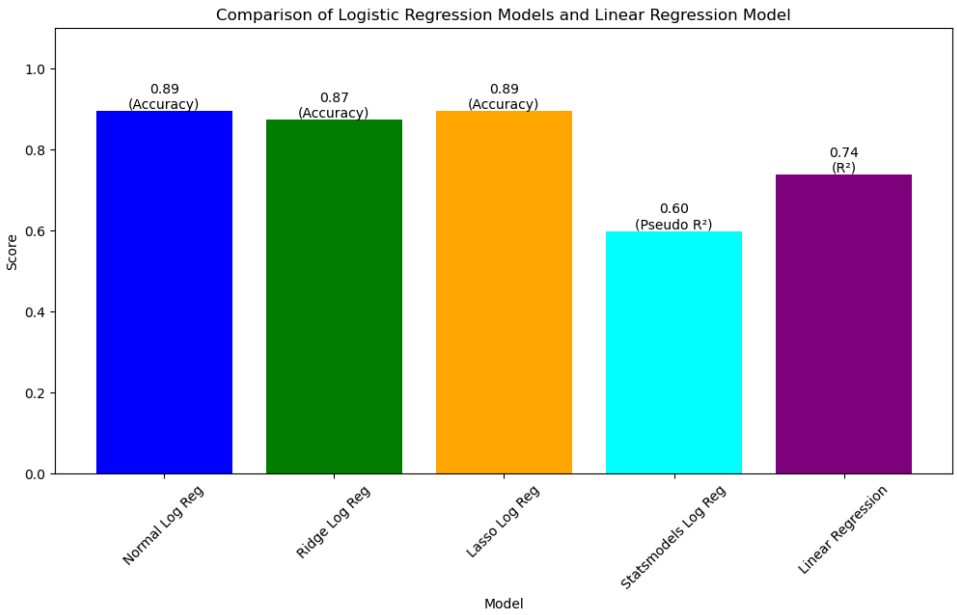
LLR p-value: 3.759e-23

- The likelihood ratio test p-value is extremely small, indicating that the model as a whole is statistically significant

Social support: 6.2111

- Positive and statistically significant (p = 0.013), suggesting that higher social support significantly increases the log-odds of the positive outcome.

Statistically Significant Predictors: GDP per capita, social support, and freedom to make life choices are significant predictors of the binary outcome. Their positive coefficients indicate that as these predictors increase, the likelihood of the positive outcome also increases.



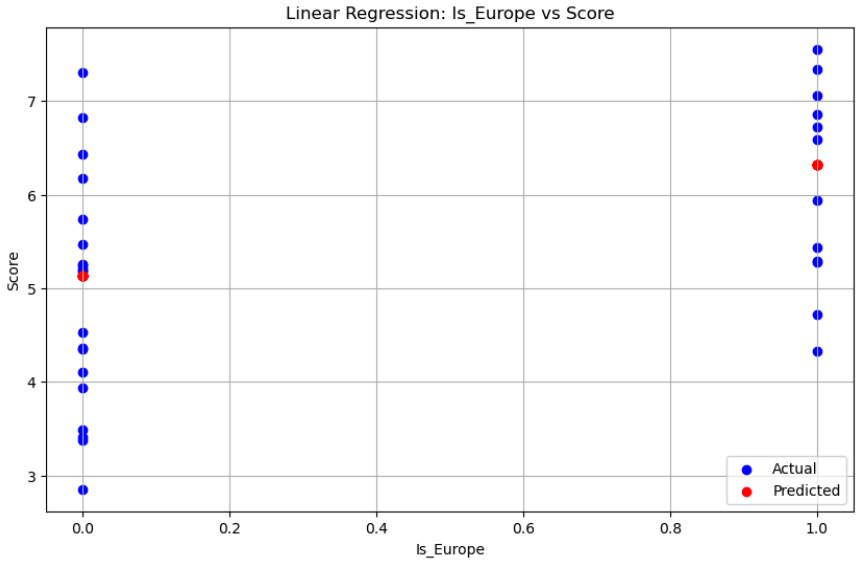
Logistic Regression Intercept: -7.054138917215498
Logistic Regression Coefficients:
GDP per capita: 1.6267428370389814

Social support: ***1.480992204134107: A one-unit increase in Social support is associated with an increase in the log-odds of being happy by 1.48, accepting H1.***
Healthy life expectancy: 1.396284986920191
Freedom to make life choices: 0.25172957535536006
Perceptions of corruption: 0.36306356224808306

Lasso Coefficients:
GDP per capita: 1.1251083192593825
Social support: 0.8735533594961641: A one-unit increase in Social support is associated with an increase in the happiness score by approximately 0.87 units. This positive coefficient indicates that higher Social support is associated with higher happiness scores, accepting H1.
Healthy life expectancy: 1.5622147427709918
Freedom to make life choices: 0.26251801680768827
Perceptions of corruption: 0.842877219980052

<u>ANALYSIS 3</u>: Linear Regression: Happiness Levels in relation to Continent

Hypotheses:
H0(2): There is no significant relationship between happiness scores and living in a country in Europe.
H1(1): There is a significant relationship between happiness scores and living in a country in Europe. Living in Europe is associated with an increase in happiness scores compared to not living in Europe.



Mean Squared Error (MSE): 1.3162814595959718: The average squared difference between the expected and actual values is measured by this metric. The MSE in Lasso, Ridge, and the enhanced linear model is 1.0444, whereas it is roughly 1.3163 in the linear model. Better model

performance is indicated by a lower MSE.

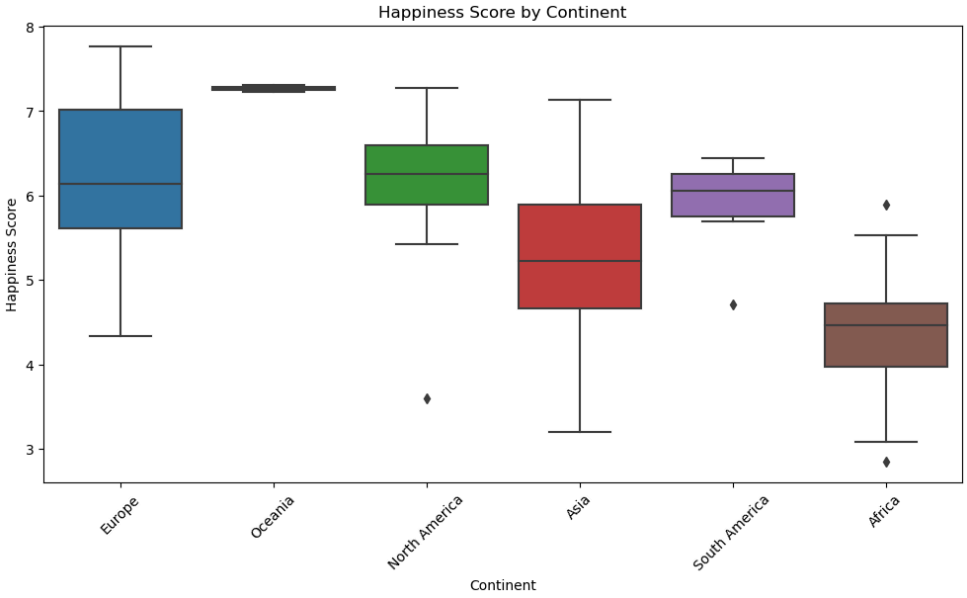
Root Mean Squared Error (RMSE): 1.1472931009972873: The average magnitude of the mistakes in the projected values is measured by RMSE, which is the square root of the MSE. The RMSE of the linear model is roughly 1.4773, but the RMSE of the other models is equal to the MSE. Once more, lower values are preferred.

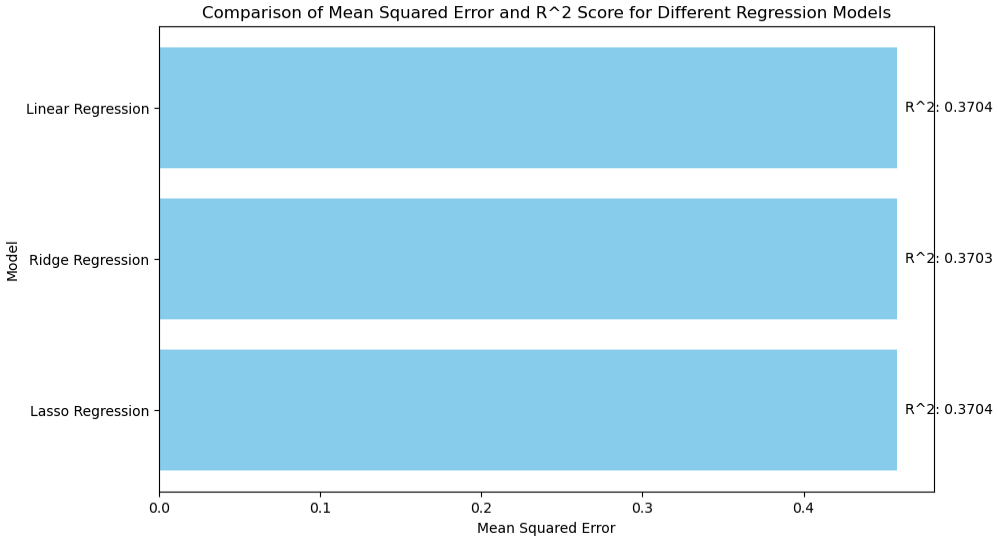
R² Score: 0.17215126960468774: The percentage of the dependent variable's variance that can be predicted from the independent variables is expressed as the R2 score, sometimes referred to as the coefficient of determination. The dependent variable's variation is explained by the independent variable to the extent that the dependent variable's R2 score in the linear model is roughly 0.1722.

Linear Regression Evaluation Metrics:
Mean Squared Error (MSE): 1.3162814595959718
Root Mean Squared Error (RMSE): 1.1472931009972873
R² Score: 0.17215126960468774
Coefficient of Is_Europe: 1.1895829787234042

Lasso Regression Evaluation Metrics:
Best Alpha: 0.01
Coefficient of Is_Europe: 1.135058156028369
Lasso Regression Train R^2 Score: 0.22762486421017003
Lasso Regression Test R^2 Score: 0.1739768490419391
Mean Squared Error: 1.313378783928249
R^2 Score on Testing Set: 0.1739768490419391

Ridge Regression Evaluation Metrics:
Best Alpha: 1
Best Cross-Validation R^2 Score: 0.18884394154913045
R^2 Score on Testing Set: 0.17386396862586762
Coefficient of Is_Europe: 1.1394782608695655
Mean Squared Error: 1.3135582640595471
R^2 Score on Testing Set: 0.17386396862586762


Happiness Score by Continent
Happiness Score
Continent
Europe
Oceania
North America
Asia
South America
Africa
8
7
6
5
4
3


Comparison of Mean Squared Error and R^2 Score for Different Regression Models
Model
Linear Regression
Ridge Regression
Lasso Regression
R^2: 0.3704
R^2: 0.3703
R^2: 0.3704
0.0
0.1
0.2
0.3
0.4
Mean Squared Error

Of the three models, the linear regression model has the largest MSE and RMSE, indicating a greater average squared difference between the actual and projected values.
Similar R2 values for each model suggest that they account for 17–18% of the variance in the target variable.
The Is_Europe predictor variable's coefficients are rather close in all models, indicating that living in Europe affects projected scores in all models in a comparable way.

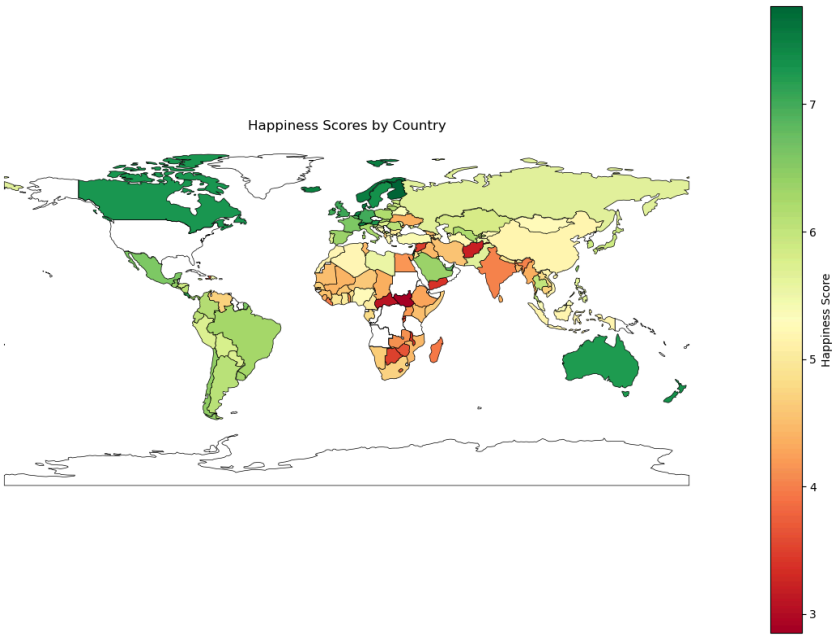
- For Linear Regression: ***The coefficient of Is_Europe is 1.190, indicating that, on average, living in a European country is associated with an increase in happiness scores by 1.190 units compared to not living in Europe, accepting H1.***
- For Lasso Regression: ***The coefficient of Is_Europe is 1.135, suggesting a similar interpretation as above.***

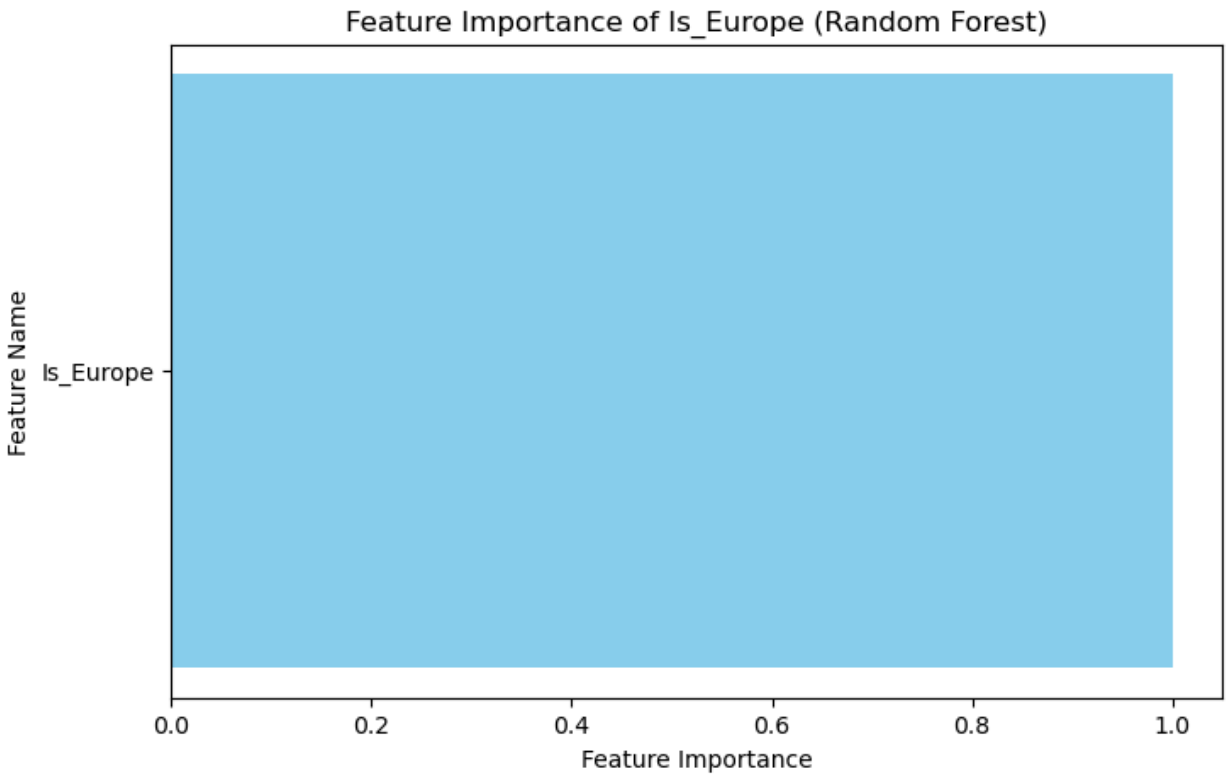
- For Ridge Regression: ***The coefficient of Is_Europe is 1.139, which also aligns with the interpretation provided by Linear and Lasso Regression.***

Since the coefficients are positive and not equal to zero, this suggests that there is evidence to support the alternative hypothesis (H1). Living in Europe is associated with a statistically significant increase in happiness scores compared to not living in Europe

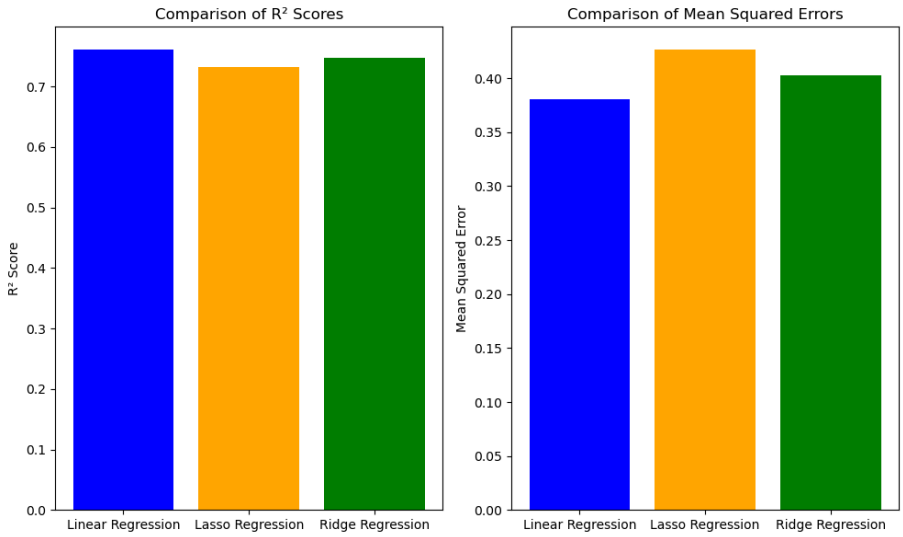
Random Forest Modeling:
R^2 Score on Testing Set (Random Forest): 0.17579930048754733
Feature Importance (Is_Europe): 1.0

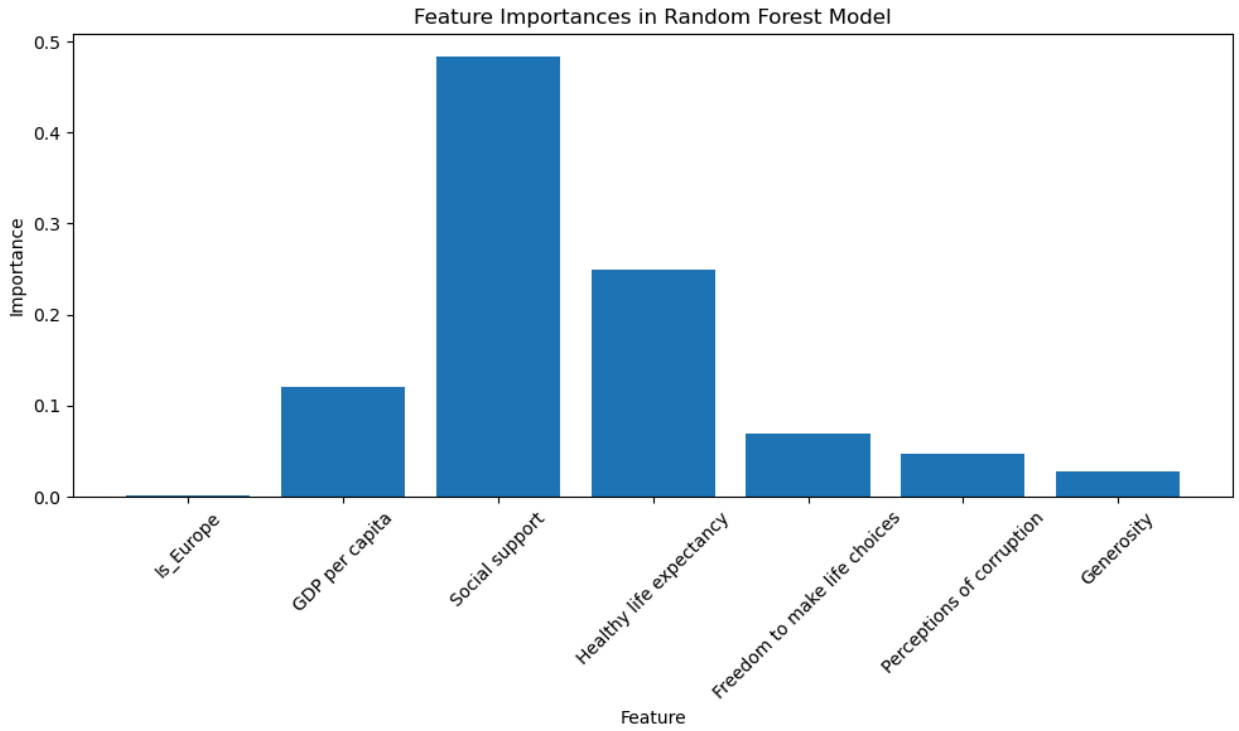
The Random Forest model's feature importance of 1.0 for Is_Europe supports H1(1), indicating that there is a significant relationship between happiness scores and living in a country in Europe. Furthermore, it suggests that living in Europe is strongly associated with an increase in happiness scores compared to not living in Europe, as Is_Europe is the most important feature in predicting happiness scores according to the Random Forest model.

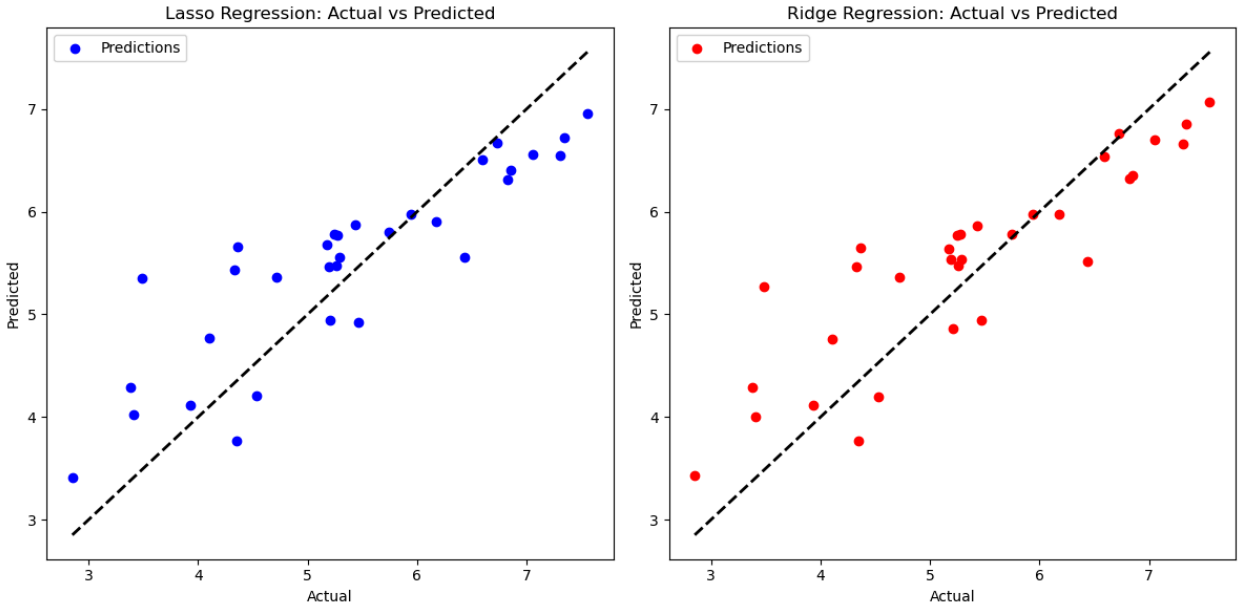




ANALYSIS 3: EXTENSION: OTHER FACTORS INCLUDED







Mean Squared Error (MSE): 0.38041227086962087

Root Mean Squared Error (RMSE): 0.6167757054794075

R^2 Score: 0.7607473590314962

Coefficient of Is_Europe: 0.23540731665880985: note that coefficient for being in europe decreased to 0.235 when other factors were included in the random forest model. Since the coefficient is positive and not equal to zero, this suggests that there is evidence to support the alternative hypothesis (H1). Living in Europe is associated with a statistically significant increase in happiness scores compared to not living in Europe.

Coefficient of GDP per capita: 0.7083550339710856

Coefficient of Social support: 1.0785913651072891

Coefficient of Healthy life expectancy: 0.9508426954623314

Coefficient of Freedom to make life choices: 1.6156442736882448

Coefficient of Perceptions of corruption: 0.2811869157470659

Coefficient of Generosity: 0.5871626032326751

Lasso Regression:

Mean Squared Error (MSE): 0.4263051085060912

R^2 Score: 0.731883982513793

Ridge Regression:

Mean Squared Error (MSE): 0.40252136125019444

R^2 Score: 0.7468422916400803

Random Forest Modeling:

R^2 Score on Testing Set (Random Forest with multiple variables): 0.7474060993836442

Using the Random Forest model, the R2 score shows what much of the variance in the dependent variable (happy score) can be predicted from the independent factors.

An R2 of roughly 0.747 indicates that the predictors in the model account for 74.7% of the variance in the happiness score. This shows that the Random Forest model fits the data well and reveals a high correlation between the predictors and the happiness score.

Feature Importances: [0.00204356 0.12089449 0.48371644 0.24985969 0.06996415 0.0464603 0.02706136]

The coefficient for Is_Europe is approximately 0.235. This coefficient represents the average change in the happiness score for countries in Europe compared to countries not in Europe, holding all other predictors constant. Since coefficient is positive, H0 is rejected.

Mean Squared Error: 0.40162490559062575: The average squared difference between the actual and anticipated values is measured by the Mean Squared Error, or MSE.
The squared differences between the actual happiness scores and the scores predicted by the Random Forest model are, on average, 0.402 units, as indicated by an MSE of roughly 0.402. Better model performance is indicated by lower MSE values.