Data Mining
Suicide Prediction System

MILESTONE 3: MODEL IMPLEMENTATION

By Group7

Atharv Kadam、Om Kokate、Yuni Wu

Professor

Alfonso G. Bastias, Ph.D.

1. RANDOM FOREST MODEL

The Random Forest model exhibited strong performance before fine-tuning, achieving an accuracy of 93.4%, precision of 93.3%, recall of 93.4%, and an F1 score of 93.2%. These metrics indicate a robust capability in correctly identifying and classifying the risk levels of suicide rates. Feature importance analysis revealed that Population contributed the most (~35%), followed by GDP (~25%) and Average Temperature (~20%), while the Happiness Index had the least impact (~15%). This suggests that socio-economic and demographic factors play a crucial role in the predictions, while overall societal well-being may be less granular for capturing individual risks. The residual plot displayed no significant patterns, validating the model's stability. However, the model's performance for the "Medium" risk class was slightly lower, with a recall of 75%, which affected its overall precision-recall balance.

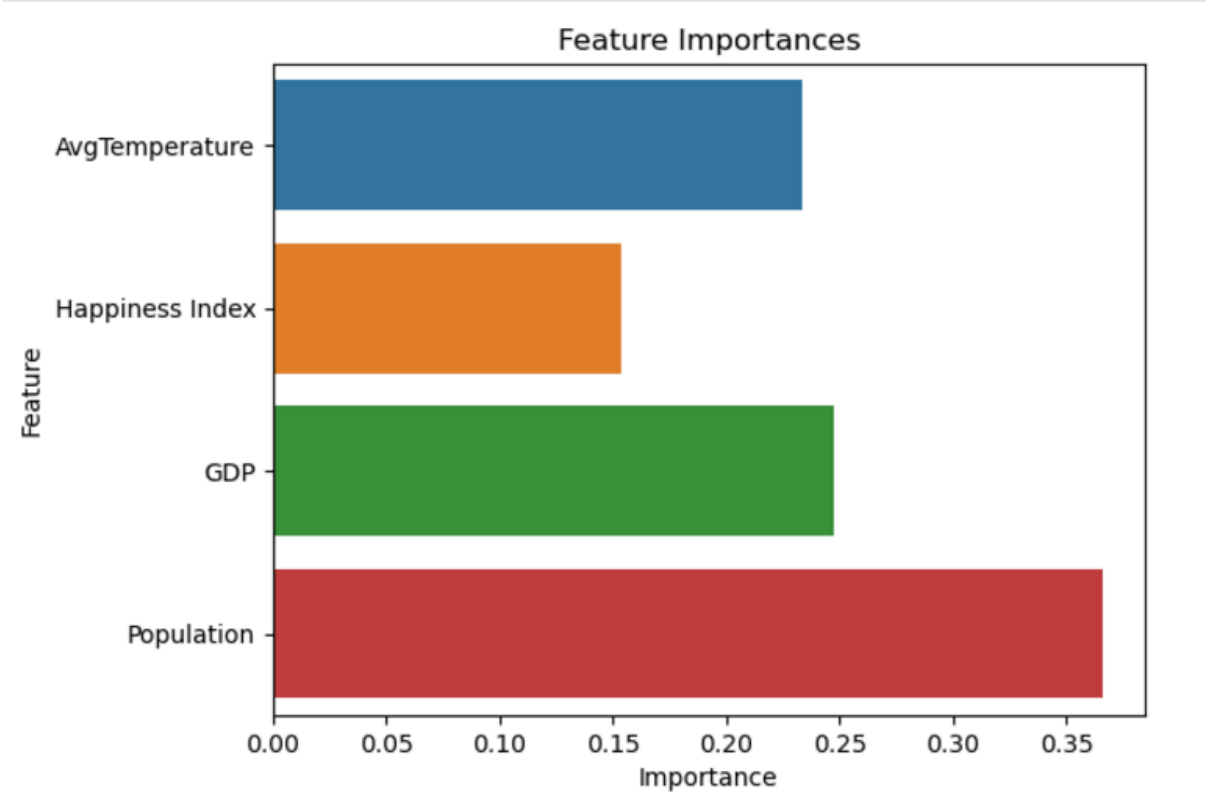
After fine-tuning, the model's performance improved slightly, with accuracy increasing to 93.9%, precision to 93.7%, recall to 93.9%, and F1 score to 93.7%. The fine-tuned hyperparameters included 200 estimators, sqrt for max_features, no maximum depth, min_samples_split of 2, and min_samples_leaf of 1. Notably, recall for the "Medium" class increased from 75% to 77%, and its F1 score improved from 81% to 83%, reducing misclassification for this category. The confusion matrix and classification report confirmed the model's improvement in handling the "Medium" risk level while maintaining consistently high performance for the "High" risk level (precision, recall, and F1 score of ~95%). Although the gains were marginal, the fine-tuned model demonstrated enhanced balance and reliability, making it a robust tool for this task. Future exploration of advanced ensemble methods or interpretability-focused approaches might yield further improvements.

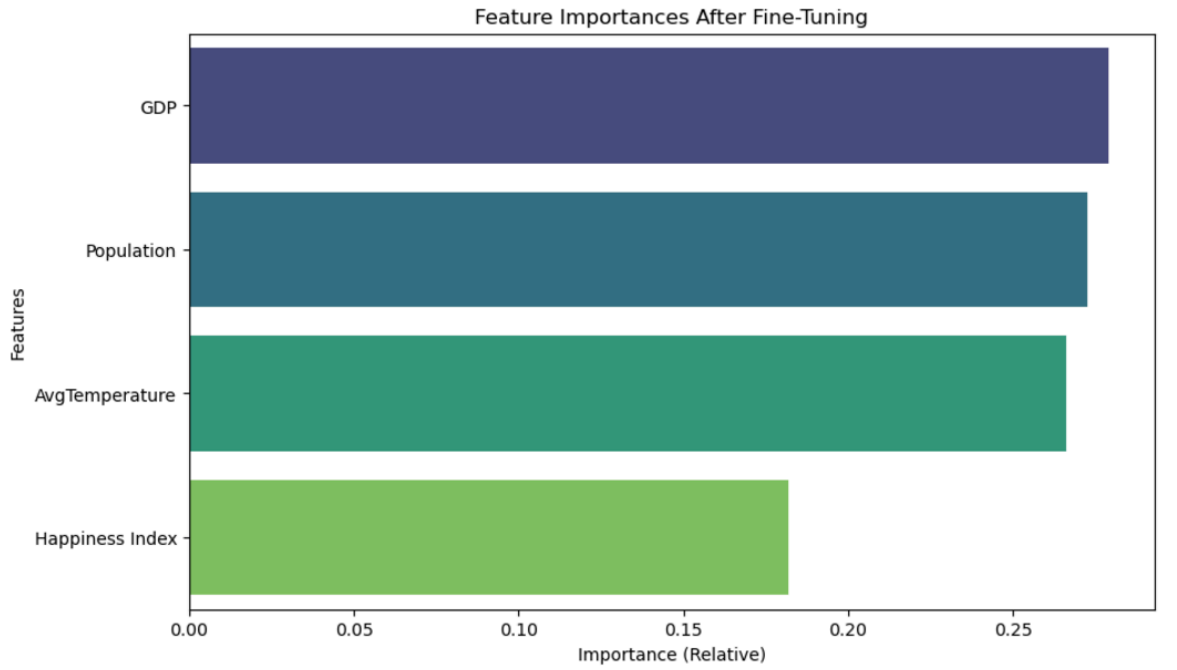
Before Transformation:

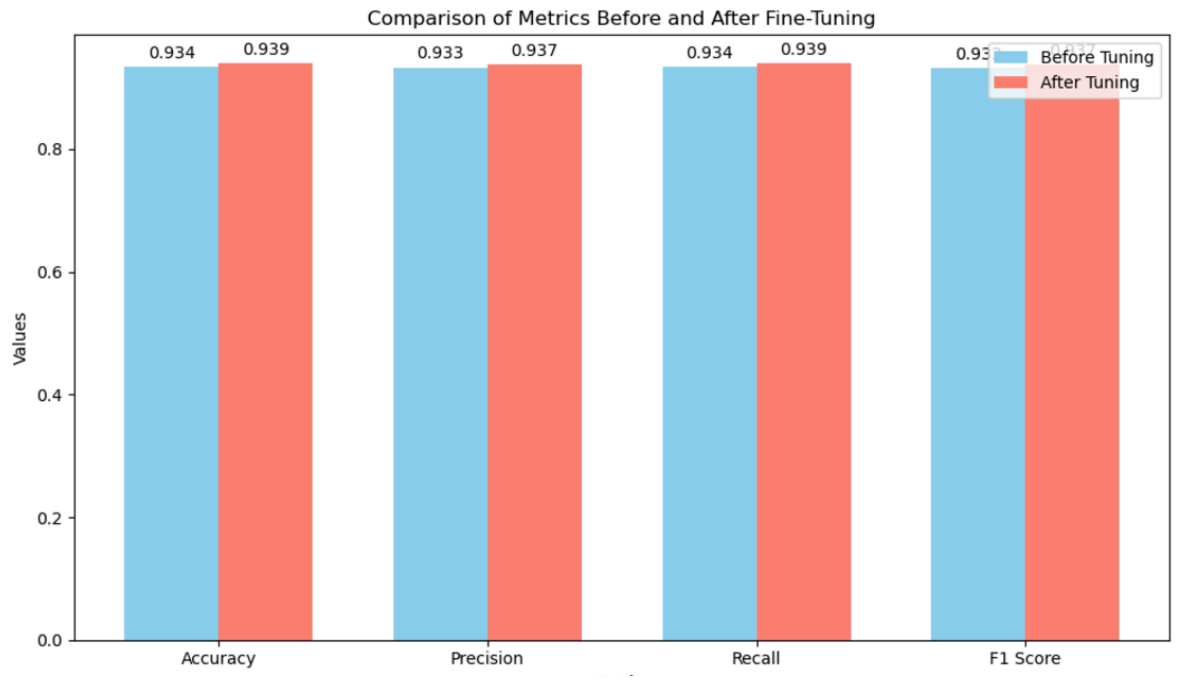
	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male
0	Africa	Algeria	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83
1	Africa	Algeria	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63
2	Africa	Algeria	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41
3	Africa	Algeria	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22
4	Africa	Algeria	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02
5	Africa	Algeria	2010	64.268219	5.464000	4495.921455	35856344.0	3.00	2.19	3.81
6	Africa	Algeria	2011	64.960822	5.317000	5473.446129	36543541.0	2.94	2.13	3.74
7	Africa	Algeria	2012	64.290437	5.605000	5610.733341	37260563.0	2.89	2.10	3.68
8	Africa	Algeria	2013	63.704658	5.980000	5519.777576	38000626.0	2.85	2.06	3.64
9	Africa	Algeria	2014	65.195890	6.355000	5516.229431	38760168.0	2.78	2.02	3.53

After Transformation:

	Region	Country	Year	Male	Female	Both sexes	AvgTemperature	Happiness Index	GDP	Population	Actual	Predicted
0	Western Pacific	Malaysia	2010	7.94	2.53	5.30	83.427945	5.580	8880.146040	28717731.0	High	High
1	Europe	Spain	2009	8.97	2.47	5.62	59.790137	6.199	32037.209190	46362946.0	High	High
2	Americas	Canada	2009	15.53	5.04	10.24	41.353671	7.488	40918.850660	33630069.0	High	High
3	Europe	Norway	2006	16.91	6.38	11.65	43.173973	7.416	74427.565413	4660677.0	High	High
4	Americas	Cuba	2007	16.22	3.99	9.92	75.603562	5.418	5200.034393	11269887.0	High	High
5	Africa	Mauritania	2009	7.26	4.12	5.56	76.707397	4.500	1418.940825	3322616.0	High	High
6	Eastern Mediterranean	Kuwait	2006	3.71	1.15	2.72	80.731507	6.076	42971.392200	2363409.0	Medium	High
7	Africa	Gabon	2013	24.29	3.97	13.57	79.624932	3.800	9247.418332	1902226.0	High	High
8	Europe	France	2018	15.80	4.73	10.05	51.521233	6.666	41937.933915	67158348.0	High	High
9	Africa	Zambia	2014	30.20	7.87	17.74	56.945455	4.346	1696.117261	15737793.0	High	High







2. KNN MODEL

The KNN model demonstrated excellent performance before fine-tuning, achieving 96% accuracy, precision, recall, and F1 score. These balanced metrics indicate the model's reliability and effectiveness in identifying both positive and negative cases, making it well-suited for sensitive tasks like suicide prediction. The F1 score further supports a strong balance between precision and recall, emphasizing the model’s robustness. Visual insights such as the ROC curve with an AUC of 0.99 and the precision-recall curve, which maintained high values across thresholds, confirm the model's ability to discriminate between classes effectively. Feature importance analysis revealed that Country_Encoded, Region_Encoded,

and Happiness Index significantly contributed to predictions, while Population had the least impact. The decision boundary visualization showed clear separation of classes with minimal overlap, validating effective feature scaling and selection.

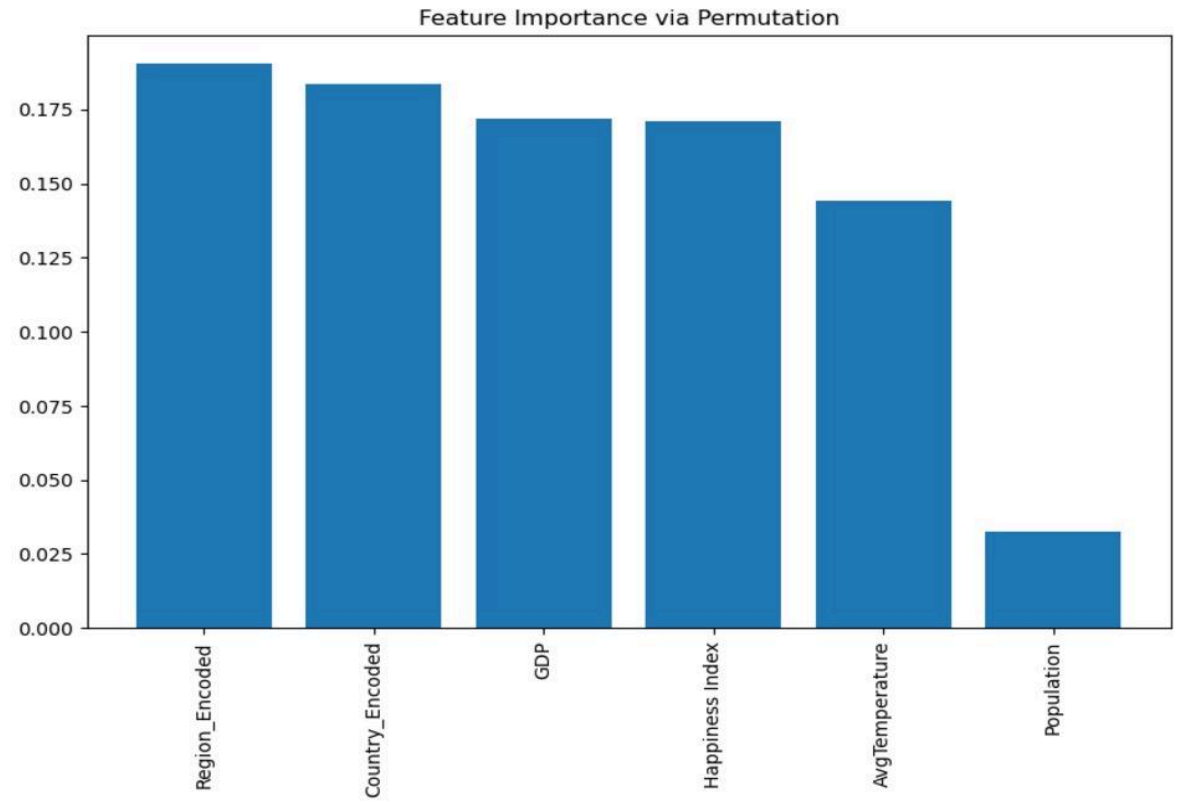
After fine-tuning, the model's performance improved slightly, with accuracy, precision, recall, and F1 score increasing to 97%, representing a 1.38% improvement. This improvement highlights the value of hyperparameter tuning, with the optimal parameters being n_neighbors: 3, metric: Manhattan distance, and weights: Uniform. The fine-tuned model showed enhanced recall for the at-risk class, crucial for minimizing false negatives in suicide prevention. The accuracy peaked at k=3, as demonstrated in the K vs. Accuracy graph, underscoring the need for tuning the number of neighbors. Although the improvement was marginal, the fine-tuned model remains a highly robust and practical tool for predicting suicide risks. However, the dataset appears well-suited for KNN, as further significant gains might require exploring advanced models like Random Forest, Logistic Regression, or Gradient Boosting for potentially better performance and interpretability.

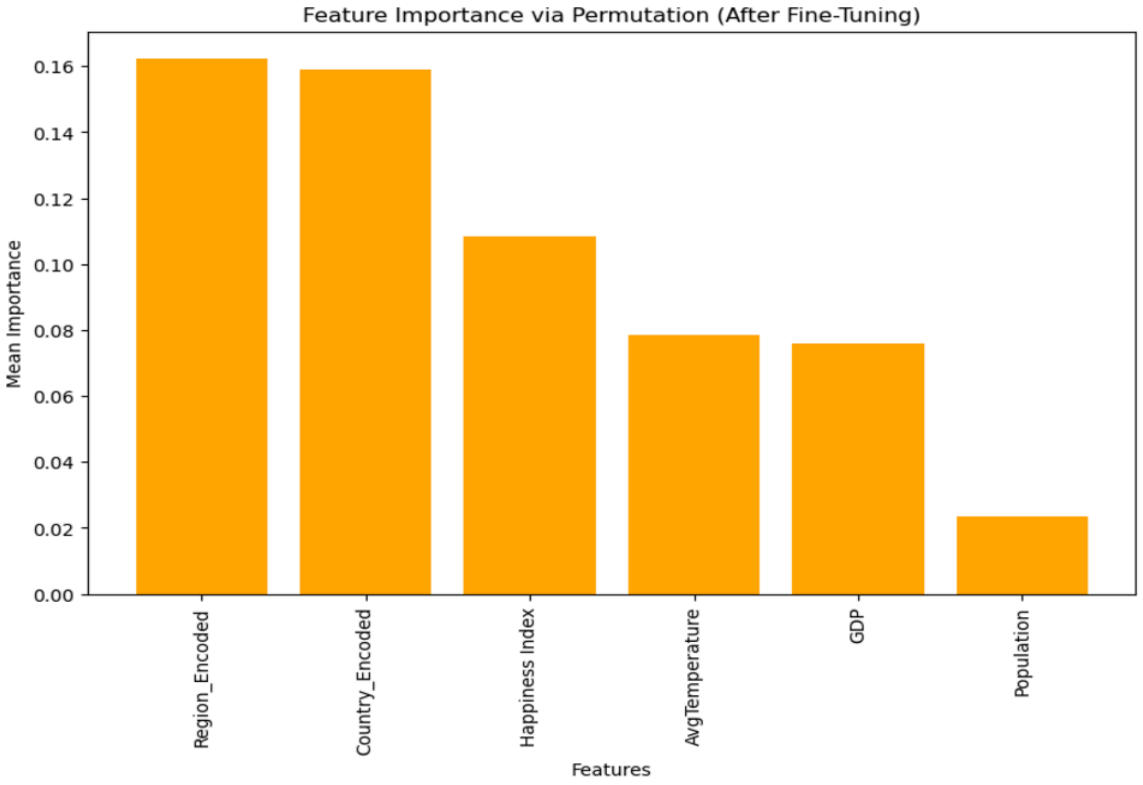
Before Transformation:

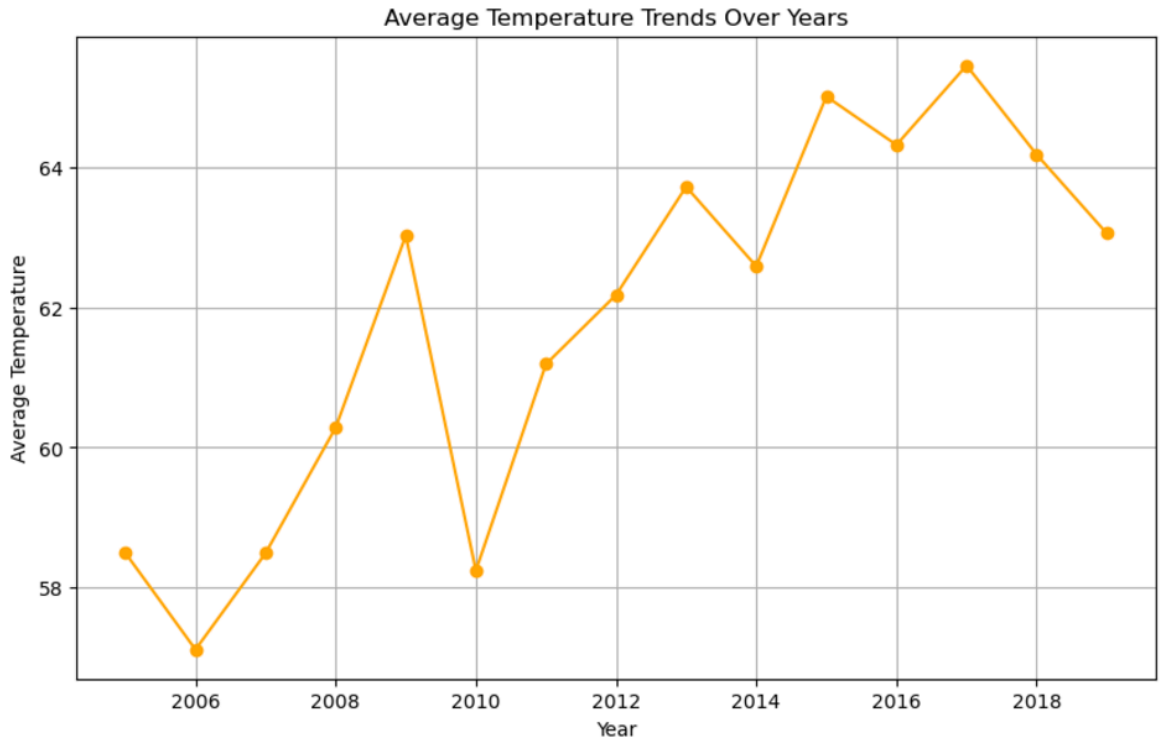
	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male
0	Africa	Algeria	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83
1	Africa	Algeria	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63
2	Africa	Algeria	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41
3	Africa	Algeria	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22
4	Africa	Algeria	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02
5	Africa	Algeria	2010	64.268219	5.464000	4495.921455	35856344.0	3.00	2.19	3.81
6	Africa	Algeria	2011	64.960822	5.317000	5473.446129	36543541.0	2.94	2.13	3.74
7	Africa	Algeria	2012	64.290437	5.605000	5610.733341	37260563.0	2.89	2.10	3.68
8	Africa	Algeria	2013	63.704658	5.980000	5519.777576	38000626.0	2.85	2.06	3.64
9	Africa	Algeria	2014	65.195890	6.355000	5516.229431	38760168.0	2.78	2.02	3.53

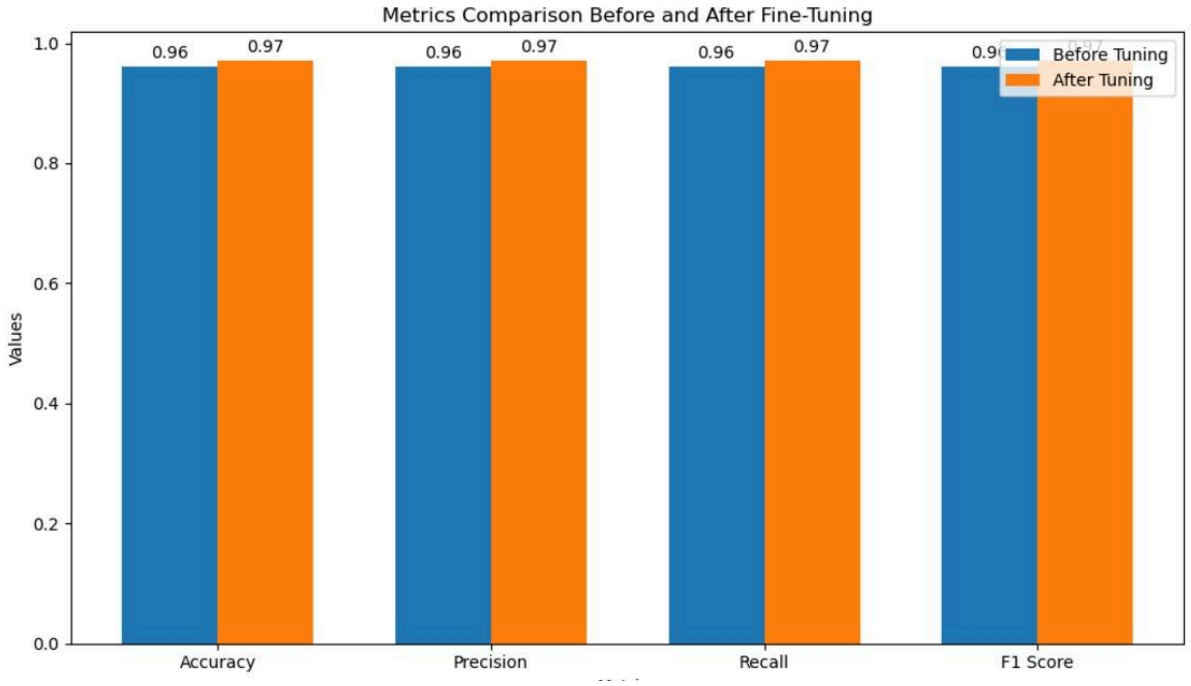
After Transformation:

	Region	Country	Year	Male	Female	Both sexes	AvgTemperature	Happiness Index	GDP	Population	Region_Encoded	Country_Encoded
0	Western Pacific	Australia	2019	16.97	5.58	11.25	-0.002846	1.288728	1.756078	-0.199127	1.837095	-1.600845
1	Africa	Nigeria	2014	11.31	4.10	7.59	-0.004307	-0.727445	-0.686955	0.517292	-1.306302	0.627529
2	Americas	Dominican Republic	2008	9.37	1.93	5.61	0.435665	-0.755649	-0.599450	-0.272664	-0.677623	-0.829485
3	Western Pacific	Philippines	2018	3.97	1.31	2.58	0.927319	0.122100	-0.685743	0.187972	1.837095	0.884649
4	Americas	Panama	2019	4.80	0.98	2.86	0.786125	0.307564	-0.052644	-0.297269	-0.677623	0.798942
5	South-East Asia	Sri Lanka	2015	23.95	6.73	14.70	0.928928	-0.952224	-0.647805	-0.217722	1.208415	1.313183
6	Europe	Finland	2005	25.07	9.07	16.88	-0.899735	1.610085	1.023329	-0.292555	0.579736	-0.786632
7	South-East Asia	Sri Lanka	2011	29.54	7.45	17.85	0.970826	-1.320588	-0.683199	-0.219940	1.208415	1.313183
8	Americas	Haiti	2010	16.05	10.82	13.20	0.948636	-1.675277	-0.779877	-0.271176	-0.677623	-0.443805
9	South-East Asia	Indonesia	2019	3.99	1.15	2.55	0.947923	-0.324039	-0.640135	0.936807	1.208415	-0.229538


Feature Importance via Permutation
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Region_Encoded
Country_Encoded
GDP
Happiness Index
AvgTemperature
Population


Feature Importance via Permutation (After Fine-Tuning)
Mean Importance
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Region_Encoded
Country_Encoded
Happiness Index
AvgTemperature
GDP
Population
Features





3. Light GBM MODEL

The LightGBM model demonstrated strong performance on the dataset, both in terms of initial results and the significant improvements achieved through fine-tuning. Initially, the model achieved a test accuracy of 87%, with precision, recall, and F1 scores all at 86%. These metrics indicate that even with default parameters, LightGBM effectively captured the relationships within the dataset, performing well in predicting target outcomes. However, the

slight gap between the training and testing metrics hinted at the potential for optimization to improve generalization. After fine-tuning key hyperparameters such as the number of estimators, learning rate, maximum tree depth, and minimum data in leaves, the model's performance improved notably. The test accuracy increased to 91%, while the precision, recall, and F1 scores rose to 90%. This optimization ensured a better fit to the training data while maintaining robust performance on unseen data. The improvements highlight the importance of parameter tuning in gradient boosting models like LightGBM, where small adjustments can lead to substantial gains in predictive accuracy and balanced classification. The visualizations played a crucial role in evaluating and understanding the model's performance and decision-making process. The feature importance plot provided insights into the contributions of each input variable, allowing us to identify the most influential features for predictions. This information is invaluable for refining the dataset and focusing on high-impact features in future iterations. The confusion matrices before and after fine-tuning provided a detailed breakdown of how the model handled each class, visually showcasing the reduction in misclassifications and the balanced improvements across all categories after optimization.

Additionally, correlation heatmaps and SHAP value analyses offered deeper insights into feature interactions and their influence on predictions. These visualizations underscored the interpretability of the LightGBM model, making it not only a powerful but also an understandable tool for classification tasks. Together, the results and visualizations demonstrated LightGBM's capacity to deliver high accuracy and interpretability, making it an excellent choice for the dataset and offering a clear path for further enhancements.

Before Transformation:

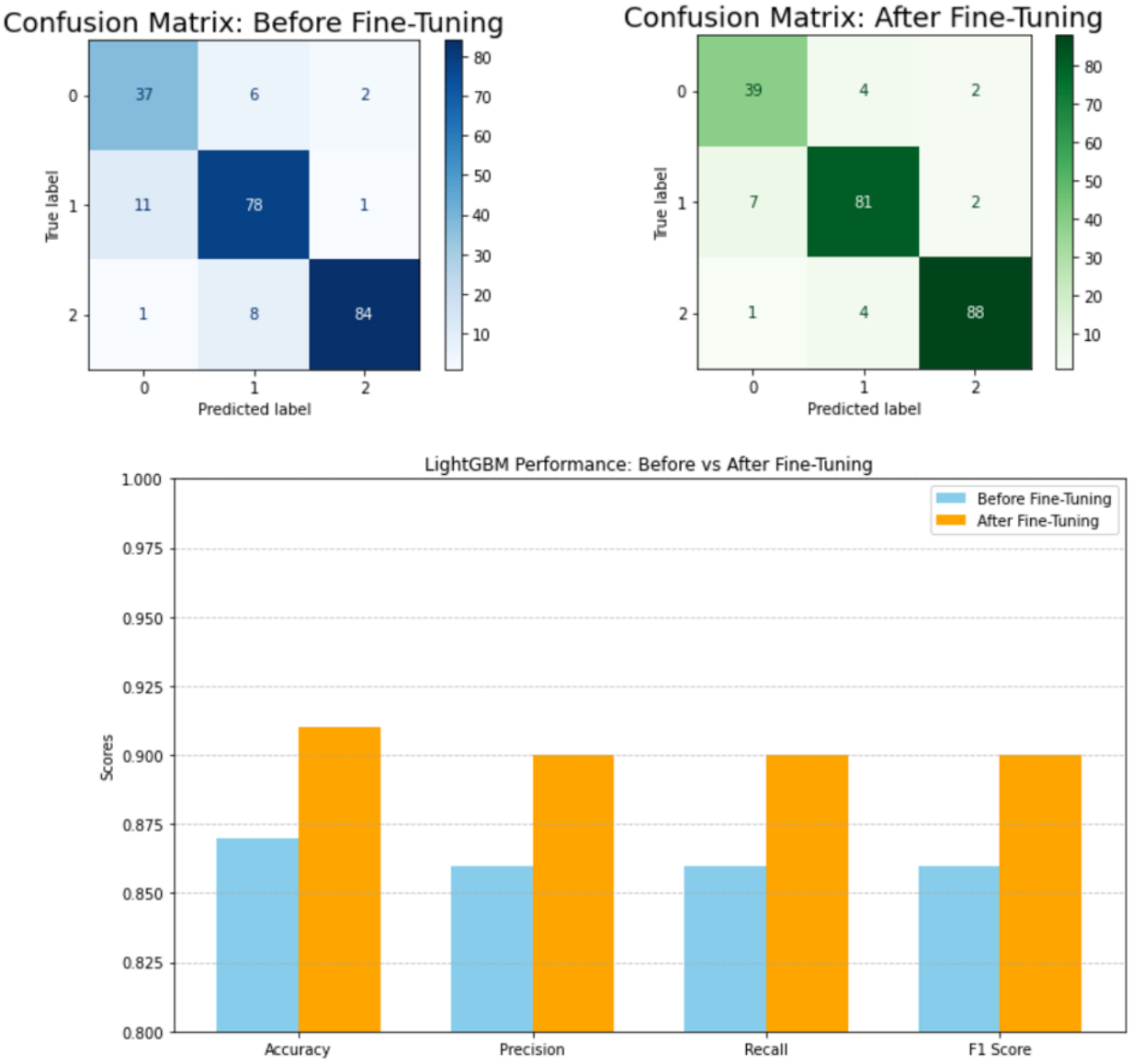
Out[42]:

	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male
0	Africa	Algeria	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83
1	Africa	Algeria	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63
2	Africa	Algeria	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41
3	Africa	Algeria	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22
4	Africa	Algeria	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02

After Transformation:

Out[43]:

	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male
0	0	1	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83
1	0	1	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63
2	0	1	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41
3	0	1	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22
4	0	1	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02



4. SVM MODEL

The Support Vector Machine (SVM) model was initially trained on a dataset that categorized risk levels and encoded categorical features. The dataset was scaled to ensure features were comparable, optimizing SVM performance. Before fine-tuning, the model struggled with a low accuracy of 34% and exhibited a significant class imbalance, as evident from the confusion matrix and classification report, where the model failed to effectively identify low and medium-risk classes. However, after fine-tuning using grid search for hyperparameter optimization, including adjustments to the kernel type, regularization parameter (C), and kernel coefficient (gamma), the model's performance significantly improved. Accuracy increased from 34.1% to 81.7%, with precision rising from 11.6% to 81.8%, recall improving from 34.1% to 81.7%, and F1-score going up from 17.3% to 81.3%. These changes reflected the model's enhanced ability to correctly classify all classes, particularly the "Low" and "Medium" classes, which had previously been misclassified.

The confusion matrix showed more balanced classification across the three classes, with the "Low" class achieving a recall rate of 97%, and the "Medium" class showing significant improvements in both precision and recall. Although some misclassifications occurred in the "High" class, this was expected in a multiclass setting. The improvements indicate that further enhancements might require additional techniques like oversampling, undersampling, or class-weight adjustments. Overall, this study highlighted the crucial role of hyperparameter optimization and preprocessing in refining SVM performance, making the model better equipped to handle all classes effectively.

Before Transformation:

Original Dataset (Before Transformation):

	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male
0	Africa	Algeria	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83
1	Africa	Algeria	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63
2	Africa	Algeria	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41
3	Africa	Algeria	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22
4	Africa	Algeria	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02

Before FineTune:

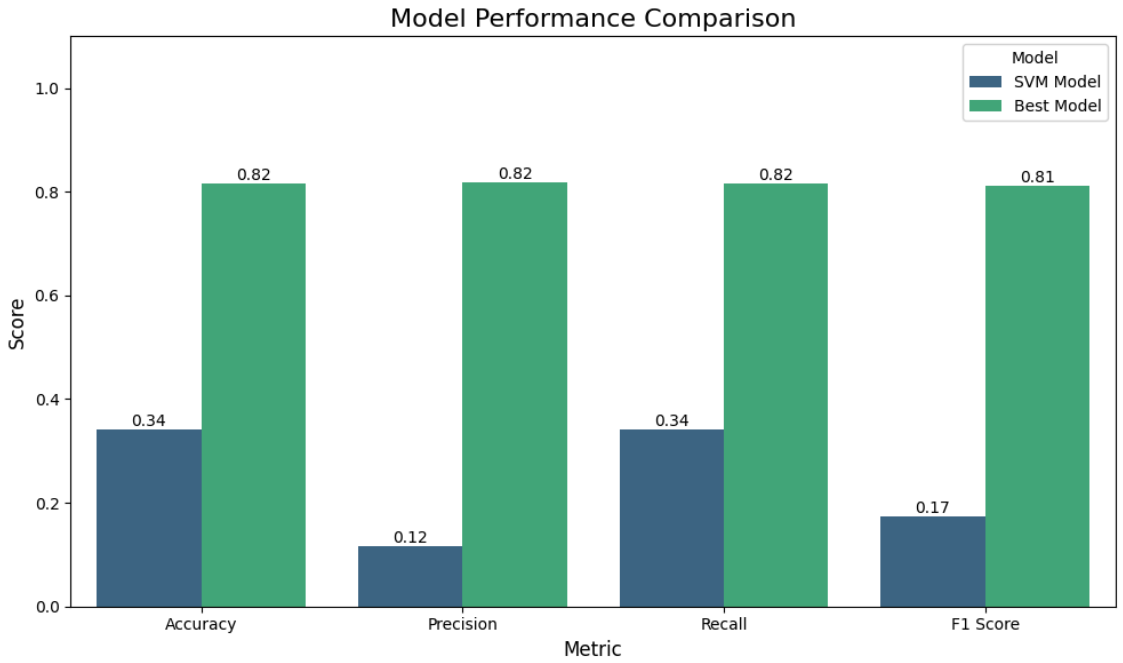
Original Dataset (Before FineTune):

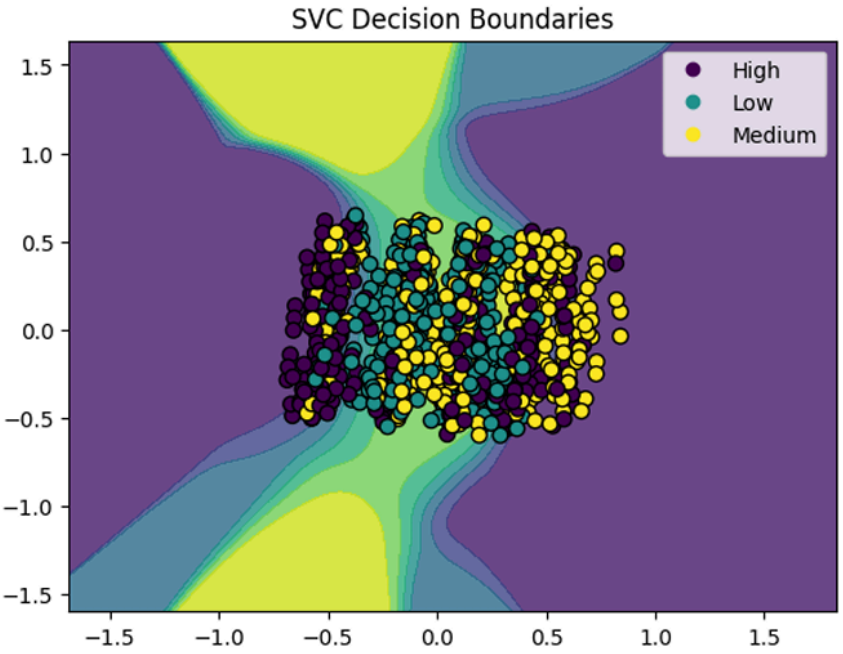
	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male	risk_category	actual	predict
0	0	1	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83	Low	Low	Low
1	0	1	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63	Low	Low	High
2	0	1	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41	Low	Medium	High
3	0	1	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22	Low	High	High
4	0	1	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02	Low	Low	High

After FineTune:

Dataset (After FineTune):

	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male	risk_category	actual	predict
0	0	1	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83	Low	Low	Low
1	0	1	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63	Low	Low	Low
2	0	1	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41	Low	Medium	Low
3	0	1	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22	Low	High	Medium
4	0	1	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02	Low	Low	Low





5. XGBoost MODEL

The model initially achieved a strong accuracy of 92.58%, demonstrating reliable performance across all classes as highlighted in the classification report. The "High" and "Low" classes showed particularly high F1-scores of 0.93 and 0.95, reflecting a well-balanced combination of precision and recall. Although the "Medium" class performed slightly lower with an F1-score of 0.89, it remained robust. Macro and weighted averages of 0.93 indicated consistent performance across the dataset. However, the confusion matrix

revealed some challenges, particularly with the "Medium" class, where the model identified fewer true positives compared to the other categories. This underscores the need to look beyond overall accuracy when analyzing imbalanced datasets.

Fine-tuning the model resulted in a slight decrease in accuracy, dropping from 92.58% to 91.27%. Precision, recall, and F1-scores also experienced marginal declines across all classes, most notably in the "Medium" class, where the F1-score fell from 0.89 to 0.87. While performance for the "High" and "Low" classes remained stable, the confusion matrix revealed an increase in false negatives for the "Medium" class after fine-tuning. These changes suggest that the fine-tuning process adjusted the model's balance, prioritizing optimization for the "Low" class at the cost of a minor drop in performance for others. Overall, the results indicate that the model was already highly optimized before fine-tuning, with limited room for significant improvement.

Before Transformation:

Original Dataset (Before Transformation):

	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male
0	Africa	Algeria	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83
1	Africa	Algeria	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63
2	Africa	Algeria	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41
3	Africa	Algeria	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22
4	Africa	Algeria	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02

Before FineTune:

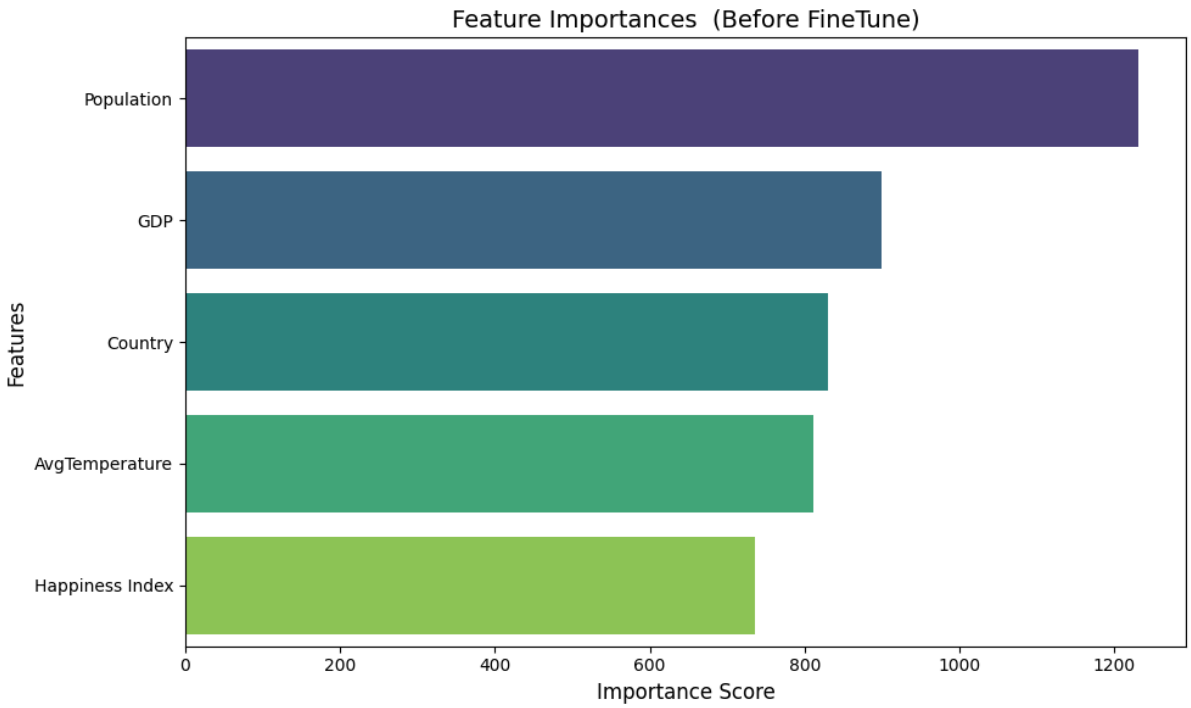
Original Dataset (Before FineTune):

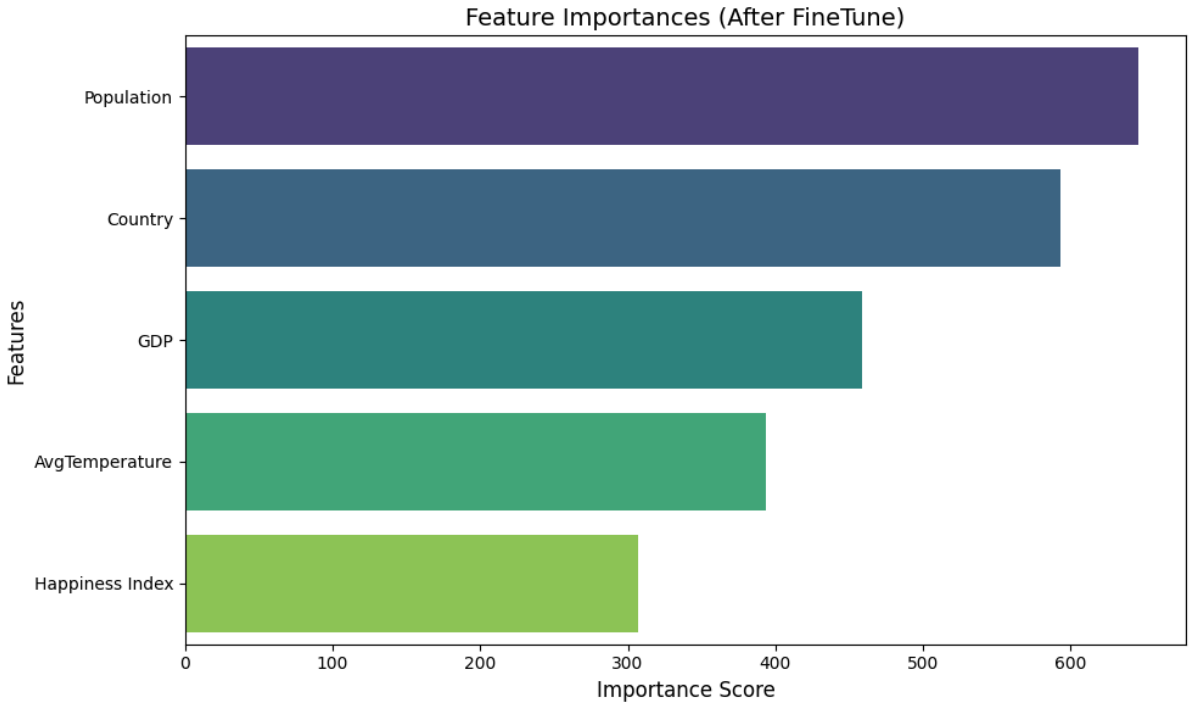
	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male	risk_category	actual	predict
0	0	1	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83	1	1.0	1.0
1	0	1	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63	1	1.0	1.0
2	0	1	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41	1	2.0	1.0
3	0	1	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22	1	0.0	0.0
4	0	1	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02	1	1.0	1.0

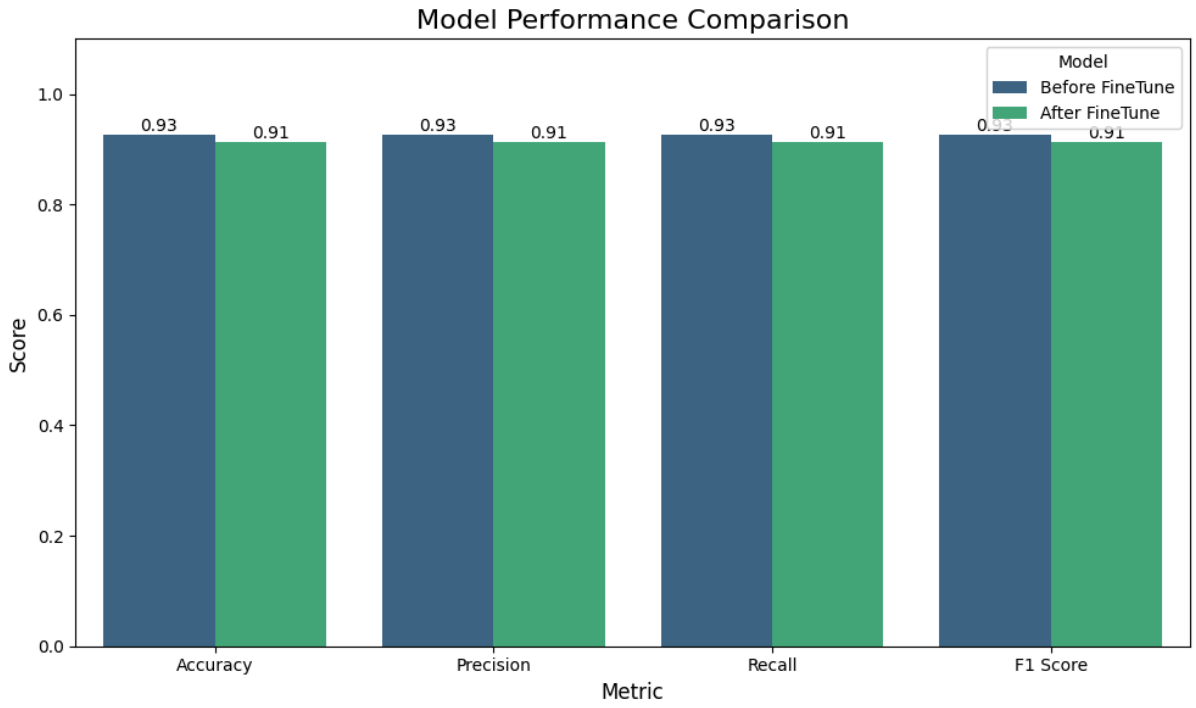
After FineTune:

Dataset (After FineTune):

	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male	risk_category	actual	predict
0	0	1	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83	1	1.0	1.0
1	0	1	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63	1	1.0	1.0
2	0	1	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41	1	2.0	1.0
3	0	1	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22	1	0.0	0.0
4	0	1	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02	1	1.0	1.0


Feature Importances (Before FineTune)
Population
GDP
Country
AvgTemperature
Happiness Index
Features
0
200
400
600
800
1000
1200
Importance Score


Feature Importances (After FineTune)
Population
Country
GDP
AvgTemperature
Happiness Index
Features
0
100
200
300
400
500
600
Importance Score


Model Performance Comparison
Model
Before FineTune
After FineTune
Score
Metric
Accuracy
Precision
Recall
F1 Score
0.93
0.91
0.93
0.91
0.93
0.91
0.93
0.91
0.0
0.2
0.4
0.6
0.8
1.0