Name : Raghad al-thobaiti

Student ID : 44105548

Machine learning project

The dataset I worked with is called the World Happiness Report, and it was downloaded from the Kaggle database. It includes data about happiness scores and factors that contribute to happiness for countries around the world between 2015 and 2019.

The original link to the dataset is: https://www.kaggle.com/datasets/unsdsn/world-happiness. The data contains important attributes like GDP, Health, Freedom, Generosity, and Trust.

This is both a regression and classification project. I first predicted the happiness score (regression), and then I created a function to classify the score into Low, Medium, and High happiness levels.

There are 6 main attributes (features), and the total number of samples after combining all years is 782 rows.

The statistical properties show that GDP and Health had the highest impact on happiness based on feature importance from the Random Forest model.

There were missing values in the dataset, and I handled them by removing rows with nulls using the dropna() function.

I visualized the results using scatter plots, feature importance bar charts, decision trees, and confusion matrices.

I didn’t normalize or standardize the data because models like Decision Tree and Random Forest don’t require it.

I renamed columns, created a new column for happiness level, dropped nulls, and split the data into training and testing sets.

I divided the data using an 80/20 split with the train_test_split() function from sklearn.

I applied all the models studied in class: Linear Regression, Decision Tree, Random Forest, KNN for regression, and Naive Bayes, SVM, and ANN for classification.

The best regression model was Random Forest, and the best classification model was ANN. KNN performed the worst due to its sensitivity to data distribution.

I evaluated models using MSE and R^2 for regression, and accuracy and classification report for classification models.

I used matplotlib and seaborn libraries for visualizations to make the results clearer and more professional.

I chose this dataset because it is meaningful and reflects real-world situations. It's important for understanding global well-being.

My best model, ANN, was able to classify happiness levels accurately using multiple input features.

One insight I discovered is that GDP and Health are consistently the strongest predictors of happiness across different countries.

To perform classification, I created a custom function called label_happiness(). This function takes the numeric happiness score and returns a label: 'Low', 'Medium', or 'High'.

The function was written with if-elif-else statements to define the thresholds for each class: below 5.5 is Low, between 5.5 and 6.5 is Medium, and above 6.5 is High.

GitHub link to my project and code: https://github.com/yourusername/World-Happiness-Project