Can we successfully predict a person's happiness level?

Introduction

Our project proposal is worthwhile because in the growing age of social media, there are increased concerns regarding impacts on mental health. There are many factors and habits that play a role in a person's happiness. Previous studies have included a connection with physical health, lack of security of fulfillment of basic needs, and social isolation. Mental health is crucial to society and struggles with it need to be addressed, but proposed solutions or methods for combating mental health struggles need to be rooted in the variables that actually impact individuals the most. Therefore, it is important to attempt to distinguish which factors contain more of a potentially significant role in order to best be able to apply findings to aid people's mental health struggles.

Methods

We found our dataset on Kaggle. The sample size is 500 entries, all being unique samples. There are 10 columns, including age, gender, sleep quality, screen time, days without social media, happiness index, and more. Our dataset is 50% male, 46% female, and 4% other. The participants range from 16 to 49 years old, with a mean age of 32.98 and a standard deviation of 9.96. To prepare our dataset for modeling, we had no missing values so we did not have to do anything in that regard. We did need to encode the categorical variables, and we dropped the user ID column.

We used four models: A Penalized Linear Model, Support Vector Machine, Ensemble Model, and a Neural Network. All models required the same preprocessing steps which are label encoding categorical variables and standardizing features using StandardScaler. Elastic Net (Penalized Linear Model) combines L1 and L2 regularization for feature selection, with hyperparameters tuned over alpha values [0.001, 0.01, 0.1, 0.5, 1, 5, 10] and l1_ratio [0.1, 0.3, 0.5, 0.7, 0.9]. SVMs capture non-linear relationships through kernel transformations with kernels including linear, rbf, and poly. The Ensemble Model used Random Forest for improved robustness. The Neural Network used fully connected layers with ReLU activation and dropout regularization, trained with Adam optimization. All models used 5-fold cross-validation to estimate generalization performance by partitioning training data into 5 equal subsets, with each fold serving as validation while the remaining folds were used for training. Model performance was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 Score to assess prediction accuracy and model reliability.

Results

Our machine learning analysis yielded significant performance differences among the four models evaluated through cross-validation. The table below summarizes the cross-validated performance metrics for the optimal hyperparameter combination of each model. Compared to the other models, the Random Forest model demonstrated the strongest performance during the optimization process. This model achieved the lowest mean squared error (MSE = 0.2495) and highest R-squared value (R^2 = 0.8951) on the training data.

Cross-Validated Performance of Optimized Models on Training Data

Model Type	MSE	R^2	Best Hyperparameters
ElasticNet Regression	0.8799	0.6302	α=0.1 L1 ratio=0.9
Support Vector Machine	0.8544	0.6409	C=0.1 kernel='linear'
Random Forest	0.2495	0.8951	n_estimators=200 max_depth=10 min_samples_leaf=2 min_samples_split=5
Neural Network (PyTorch)	0.9335	0.5960	hidden_dim=32 dropout=0.2 learning_rate=0.001 batch_size=16 n_epochs=300

Based on its cross-validation performance, we selected the Random Forest model with optimal hyperparameters (n_estimators=200, max_depth=10, min_samples_split=5, min_samples_leaf=2) to evaluate on the test set. The model achieved a test MSE of 0.8142, RMSE of 0.9023, R^2 of 0.5976, and MAE of 0.6766. Our model produced significantly degraded performance metrics when predicting on the test set. This decrease in performance is indicative of some overfitting in the model. It suggests that the model is capturing some of the patterns that allow it to generalize to unseen data, but is becoming too specific to the training data. Despite this, the model maintains some explanatory power. Its R^2 value indicates that the model can explain ~60% of the variance in the data.

Discussion

Our Random Forest model was able to handle the data the best, likely due to not be limited to only viewing linear relationships and having protections against overfitting. These advantages of a Random Forest model indicate that our data does not contain a linear relationship, which would makes since for factors contributing to happiness. We anticipated going into this project that the relationship we would be examining would be more complicated than a linear one. A lack of linear relationship would explain why the Random Forest model performed better than the ElasticNet Regression, as that inherently handles only linear data. The size of our sample likely contributes to the Random Forest out performing the Neural Network as Neural Networks usually need rather large datasets to function well. The Support Vector Machine model doesn't have a restriction in terms of needing linear data or a lot of data, but it is very sensitive to multi-class data and parameter tuning which likely caused it to not function as well.

While our project is interesting and handle a decent amount of data, there are some limitation to the conclusions that we can draw from our results. To start with, our data set was only 500 entries, while that is a sizeable sample, it is not necessarily broad enough to apply conclusions to large populations of people. Our data appeared to balance different portions of a population well, however the smaller size of the overall dataset means that it cannot claim to be truly representative. Additionally, we are lacking a few details about how the data was collected which would allow us to make further claims about the happiness index. If the index is a self-reported statistic, then it wouldn't be able to stand in as a distinctly measurable variable as individuals have different perspectives of what a 7/10 happiness looks like.