MACHINE LEARNING PROJECT FINAL REPORT

TOPIC : Happiness Score of IITJ Students

Aditya Mishra (M20MA201)

Deepanshi Jindal (M20MA202)

Vandita Agarwal (M20MA208)

Date : 20 October, 2022

ABSTRACT

Firstly, a data set of 156 samples and 25 attributes is collected through google form from IITJ students. Then its cleaning and required preprocessing for our project is done using python. The main aim of our project is to predict the happiness score depending upon the remaining 24 attributes. So, we used different techniques to find the correlation between 24 features, and selected the most relevant features. Then simple linear regression is applied on some selected features and multilevel linear regression is applied to the 14 relevant features received after checking the coefficient matrix. After that using principal component analysis, we have found 9 most high variance features and applied multilevel linear regression over those features. After that we again reduced these features to 4 and reduced the number of classes into 4. Then applied support vector machine, neural networks for classification. We finally, compared these techniques using mean square error, accuracy score, and other measures to friend best result.

INTRODUCTION

A data set consisting of 25 attributes was collected by us from IITJ students. Aim of this data is to predict the happiness score of an individual using Machine Learning techniques. After checking the correlation between various features, we have applied multiple linear regression between most correlated features and the happiness score. Also, we have made use of simple linear regression between selected features and the happiness score. Further clustering and neural networking will be shown to get new instincts about our collected data.

DATA SET DETAILS

We have collected the data Sample from the IITJ students by organizing a survey with the help of google form (form link : Happiness score). There were some key points related to the form collection and the concern related to the User of the form. Under this form the last entry was taken as a happiness score in the scale of 1 to 10 , everyone had to rate themselves.

DATA PREPROCESSING

- Converted the Excel Data set into the CSV file by the excel itself.
- Read the data into the Google colab using PYTHON Language.

DATA cleaning :

- We removed three columns , which was unnecessary to our project, and those three columns were Timestamp, Name , and Consent.
- We replaced short column names to the larger one which was provided in the google form.
- We replaced those irrelevant values by 0 values in the happiness score.
- And those who gave us the range of the happiness score, we replaced them by the mid value of that interval .

METHODOLOGY

- ### Principles

  We have already collected data from the IITJ students. And its cleaning and preprocessing is done in python. Then we are ready to use various Machine Learning techniques to our processed data.

- ### Tools and Techniques

  We will work with google colab to design our project. The Machine learning techniques that are already used in our project are simple linear regression and multilevel linear regression. And the techniques that will be used are logistic regression and neural networking, and Support Vector Machine.

- ### Correlation

  Correlation is a statistical measure (expressed as a number) that is used to describe the size and direction of a relationship between two or more variables. A positive value of correlation gives that the two variables are positively related, while a negative value of correlation tells us that the two variables are negatively related, and zero value denotes no relation at all.

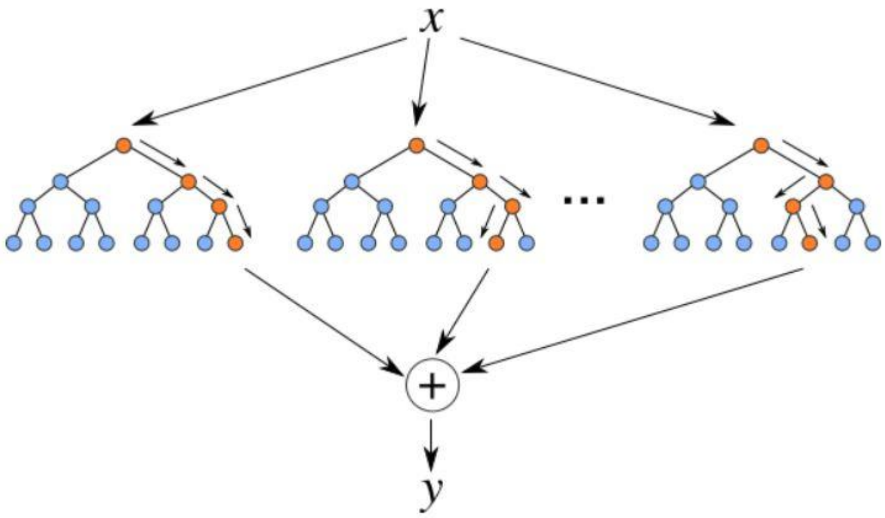
- ### Multiple linear Regression

Multiple linear regression is a supervised machine learning technique that takes several independent variables and one dependent variable and gives a relationship between them. In other words, it has to give a linear relationship between those dependent and independent variables.

- ## Principal Component Analysis

Principal Component Analysis is an unsupervised learning algorithm of feature extraction in machine learning that is used for dimensionality reduction, when a huge number of features are present. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the **Principal Components**. They are directions along whose variance among the scattered data points is the highest.

- ## Random Forest

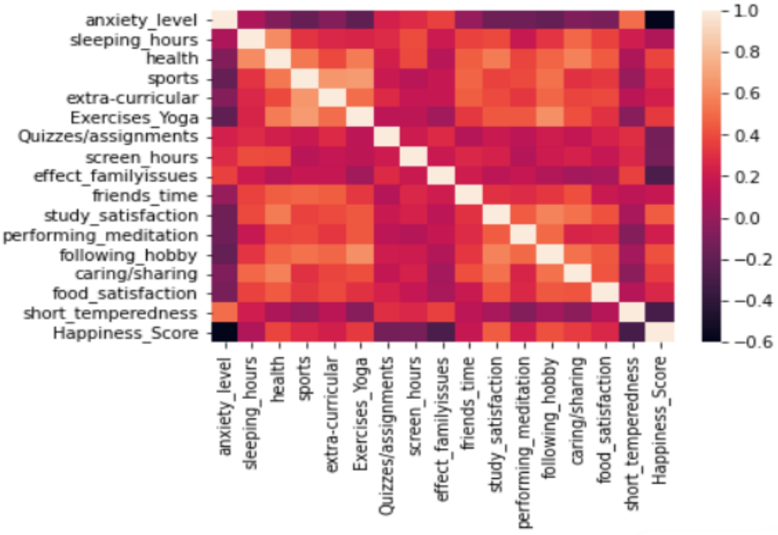


Random Forest Regression is a supervised learning algorithm that uses **ensemble learning** methods for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.The diagram above shows the structure of a Random Forest. You can notice that the trees run in parallel with no interaction amongst them. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

EXPERIMENTS AND RESULTS

- **Correlation between the features :** First of all we plotted a heat map of the data to see the correlation between the features , then we found the correlation matrix. On the basis of this matrix , we removed the most correlated columns, which had more than 0.6 correlation. We removed the following features :
- Health
- Extracurricular
- Exercise_yoga
- Following Hoby

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4436197150>
```



- **Multi-linear Regression :**

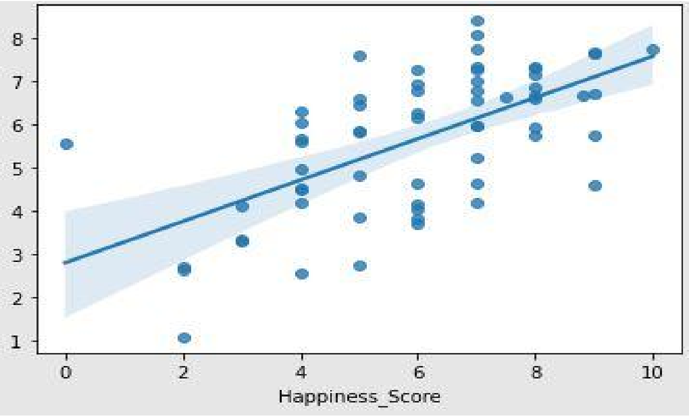
We divided the data into training and testing sets in the 60:40 ratio . Then we decided to work with the following features.

  - Anxiety Level
  - Sleeping hours
  - Quizzes and assignment
  - Screen Hours
  - Effect family issues
  - Friend’s time
  - Performing meditation
  - Food satisfaction

Then found the values of the intercept equals 6.518930097036528 and the value of the coefficients.

We created the multilinear equation with the help of these coefficients by which we can predict the happiness score of our new data point.

Then we plotted the Scattered plot between the Test value and the Prediction value and according to that we calculated the MSE (Mean square error) which appeared to be equal to 2.8828048963148936.



- ## Principal Component Analysis :

  Using principal component analysis, we have extracted the count of the number of features that gives the highest variance among datasets. We have taken all the quantitative features which are mentioned above, and PCA gives 9 most relevant quantitative features.

- ## Random Forest

  After applying the random forest technique on the 9 most relevant quantitative features. we got some results in our focus , those are as fallow ,

  Mean Square Error : 3.5742039662539697

  R_2 error = 0.1693485166452161

- ## Neural Network

  We applied the Neural Network model to our 9 most relevant quantitative features using the given parameters as Epoch = 5000, Hidden layer = 4 , Hidden nodes = 20,20,20,30 respectively , learning rate = 1e-5 , activation function = “tanh” , solver = “sgd”.

Accuracy_score = 0.31746031746031744

R_2 score = -0.184493255342

- ### Support Vector Machine

We applied the Support Vector Machine model to our 9 most relevant quantitative features using the given parameters as kernel is “Radial Basis function kernel”

Accuracy_score = 0.2539682596825

R_2 score = 0.232899606064

And the confusion matrix is concluded in the python code.

- ### Logistic regression

We applied the Logistic regression model to our 9 most relevant quantitative features and get the different results

Accuracy_score = 0.19047619047619047

- ### Major Experiment :

For getting the good accuracy results we only worked with the 4 principal components and we merged the 11 classes into 4 classes only and their results for the different models are given below.

- ### Neural Network

After applying the Neural network technique on the   4 principal components with the predefined parameters as fallow, we got some results in our focus , those are as fallow ,

Accuracy_score = 0.65

R_2 score = -0.2821461609620719

- ### Support Vector Machine

We applied the Support Vector Machine model to our 4 principal components using the given parameters as kernel is “Radial Basis function kernel”

Accuracy_score =0.55555555555

R_2 score = -0.07816836262719695

And the confusion matrix is concluded in the python code.

- ### Logistic regression

  We applied the Logistic regression model to our 4 principal components and get the different results

  Accuracy_score = 0.5714285714

CONCLUSION :

- By the above Experiments and results we conclude that , due to feature reduction using PCA Technique and after applying multiple linear regression on the PC’s (Principle component) features, there was a significant decrease in mean square error as compared to simple multiple linear regression.
- By the above experiment we also concluded that , which features are most positively related to the happiness score (By the results Correlation coefficient between Study satisfaction and happiness score is 0.455165) and which features are most negatively related to the happiness score (By the results Correlation coefficient between anxiety and happiness score is -0.601784) .
- The accuracy is increased after picking the 4 principal components among the 9 principal components and the many more related scores also increased.

FUTURE SCOPE :

We will apply different DL techniques to our data such as Convolution Neural Network. Also, we will try to do a lot of experiments with this collected data for good results. In the end we will check the accuracy of used models using various measures of error.

Google colab link Colab Lonk