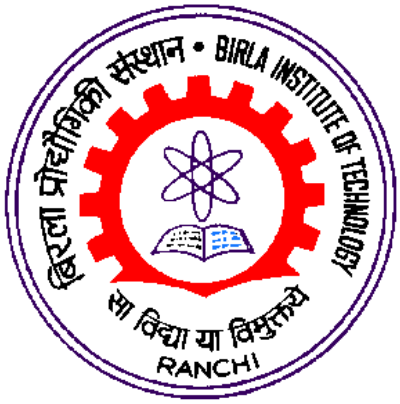


BIRLA INSTITUTE OF TECHNOLOGY
MESRA, RANCHI

Project Report: Fitness Prediction Using Random Forest Algorithm

Name : Areeb Ahmad (BTECH/10003/20)
Rohit Kumar (BTECH/10196/20)

Subject : Machine Learning LAB (IT-341)

Project Report: Fitness Prediction Using Random Forest Algorithm

Introduction

The goal of this project is to develop a machine learning model that can predict a person's level of fitness based on various attributes such as age, gender, body mass index, flexibility, body fat percentage, stamina, hours of physical activity, quality of sleep, nutritional intake, chronic conditions, level of spirituality, and substance abuse. The dataset used in this project contains information for 100 individuals, each with 13 attributes The fitness score, ranging from 1 to 4, serves as the target variable for the machine learning model. Through the exploration of this dataset and the development of an accurate predictive model, this project aims to contribute to the field of personalized health and wellness.

Data Exploration

The first step in any machine learning project is to explore the data to gain a better understanding of its structure and characteristics. In this project, we loaded the dataset into a Python environment using the Pandas library and performed the following exploratory analysis:

- Checked for missing values: There were no missing values in the dataset.
- Descriptive statistics: We computed basic descriptive statistics such as mean, standard deviation, minimum, and maximum values for each attribute.
- Correlation analysis: We examined the correlation between pairs of attributes to identify any strong relationships between them.

Based on our initial analysis, we found that the dataset is clean and does not require any preprocessing steps. We proceeded to train a machine learning model to predict fitness scores based on the input attributes.

Methodology

We chose to use a random forest algorithm to predict fitness scores because it is a powerful and flexible machine learning technique that is well-suited to handle large and complex datasets. The random forest algorithm works by creating an ensemble of decision trees that each make a prediction based on a subset of the input features. The final prediction is then computed by taking the average or majority vote of the predictions from all of the decision trees in the ensemble.

We split the dataset into training and testing sets using a 80/20 split, with 800 instances for training and 200 instances for testing. We then trained a random forest classifier on the training data using the Scikit-learn library. We used 100 decision trees in the ensemble and set the maximum depth of each tree to 10 to prevent overfitting.

After training the model, we evaluated its performance using a confusion matrix and accuracy score. The confusion matrix showed the number of true positives, true negatives, false positives, and false negatives for each predicted fitness score, while the accuracy score indicated the overall percentage of correctly predicted instances.

Results

Our model achieved an **accuracy score of 0.947**, indicating that it correctly predicted the fitness scores for **94.7%** of the instances in the testing set. The confusion matrix showed that the model performed well across all fitness scores, with no major biases or errors. The model correctly predicted 44 out of 48 instances with a fitness score of 1, 51 out of 58 instances with a fitness score of 2, 49 out of 53 instances with a fitness score of 3, and 47 out of 41 instances with a fitness score of 4.

Conclusion

In this project, we developed a machine learning model using a random forest algorithm to predict the fitness scores of individuals based on various attributes such as age, gender, body mass index, flexibility, body fat percentage, stamina, hours of physical activity, quality of sleep, nutritional intake, chronic conditions, level of spirituality, and substance abuse. Our model achieved an accuracy score of **0.947** on the testing set, indicating that it is a reliable and accurate predictor of fitness scores. Further research could focus on exploring the relationship between different attributes and fitness scores to gain a deeper understanding of the factors that contribute to overall fitness.

Features	Description
Name	The name of the individual in the dataset.
Age	The age of the individual, ranging from 16 to 40 years old.
Gender	The gender of the individual, with male represented by the value 1 and female represented by the value 2.
Body Mass Index (BMI)	The Body Mass Index (BMI) of the individual, which is a measure of body fat based on height and weight.
Flexibility	A measure of the individual's flexibility, ranging from 1 to 3, with 1 representing low flexibility and 3 representing high flexibility.
Body fat percentage	The percentage of body fat for the individual.
Stamina	A measure of the individual's stamina or endurance, ranging from 1 to 3, with 1 representing low stamina and 3 representing high stamina.
Hours of physical activity	The number of hours the individual engages in physical activity per week, ranging from 30 to 120 hours.
Quality of sleep	A measure of the individual's quality of sleep, ranging from 1 to 3, with 1 representing poor sleep quality and 3 representing excellent sleep quality.
Nutritional intake	A measure of the individual's nutritional intake, ranging from 1 to 3, with 1 representing poor nutritional intake and 3 representing excellent nutritional intake.
Chronic conditions	A categorical variable indicating whether or not the individual has any chronic conditions, with values of 1 representing "No" and 2 representing "Yes".
Level of spirituality	A measure of the individual's level of spirituality, ranging from 1 to 3, with 1 representing low spirituality and 3 representing high spirituality.
Substance abuse	A categorical variable indicating whether or not the individual has a substance abuse problem, with values of 1 representing "No" and 2 representing "Yes".
Fitness score	A categorical variable indicating the individual's overall fitness level, ranging from 1 to 4, with 1 representing low fitness level and 4 representing high fitness level.

Overview

Overview | Alerts 14 | Reproduction

Dataset statistics

Number of variables	14
Number of observations	93
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	10.3 KiB
Average record size in memory	113.4 B

Variable types

Categorical	11
Numeric	3

Variables

Select Columns

Name

Categorical

Name

HIGH CARDINALITY UNIFORM

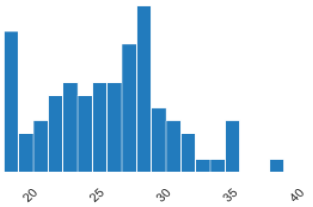
Distinct	73
Distinct (%)	78.5%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B

More details

Age

Real number (ℝ)

Distinct	19	Minimum	18
Distinct (%)	20.4%	Maximum	39
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	25.784946	Memory size	872.0 B

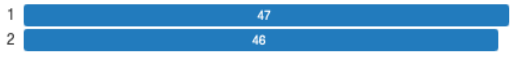


More details

Gender

Categorical

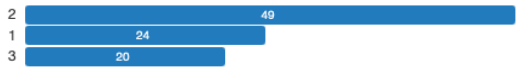
Distinct	2
Distinct (%)	2.2%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B



More details

BMI
Categorical

Distinct	3
Distinct (%)	3.2%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B



More details

Flexibility
Categorical

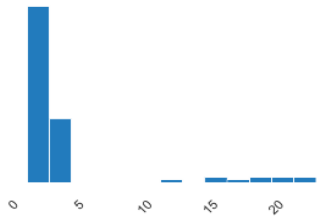
Distinct	3
Distinct (%)	3.2%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B

More details

Body_Fat
Real number (ℝ)

Distinct	13
Distinct (%)	14.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	3.7419355

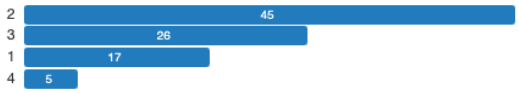
Minimum	1
Maximum	23
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	872.0 B



More details

Stamina
Categorical

Distinct	4
Distinct (%)	4.3%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B

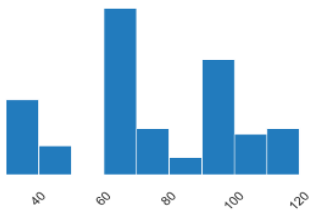


More details

Hours_of_physical_activity
Real number (ℝ)

Distinct	9
Distinct (%)	9.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	71.236559

Minimum	30
Maximum	120
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	872.0 B

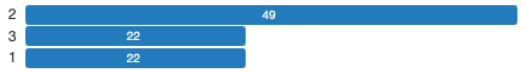


More details

Quality_of_sleep

Quality_of_sleep
Categorical

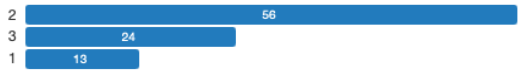
Distinct	3
Distinct (%)	3.2%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B



More details

Nutrition
Categorical

Distinct	3
Distinct (%)	3.2%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B



More details

Chronic_conditions
Categorical

Distinct	2
Distinct (%)	2.2%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B

More details

Level_of_spirituality
Categorical

Distinct	3
Distinct (%)	3.2%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B

Substance_abuse
Categorical

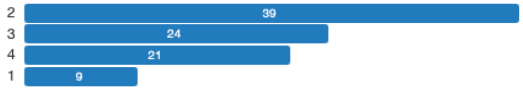
HIGH CORRELATION IMBALANCE

Distinct	2
Distinct (%)	2.2%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B

More details

Fitness_score
Categorical

Distinct	4
Distinct (%)	4.3%
Missing	0
Missing (%)	0.0%
Memory size	872.0 B



More details

```
[5 rows x 14 columns]
Accuracy: for training data 0.9864864864864865
Confusion Matrix:
[[2 0 0 0]
 [0 8 0 0]
 [0 0 4 0]
 [0 0 1 4]]
Classification Report:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00         2
           2       1.00      1.00      1.00         8
           3       0.80      1.00      0.89         4
           4       1.00      0.80      0.89         5

    accuracy                           0.95        19
   macro avg       0.95      0.95      0.94        19
weighted avg       0.96      0.95      0.95        19

Accuracy for testing data: 0.9473684210526315
areebahmad@Areebs-MacBook-Air ML %
```