Contest: Phase 1
Problem: ChildStuntedness2

Problem Statement

Stunting (shortness for age) affects more than one in four children worldwide. Wasting (under-weightedness) and stunting in early childhood is associated wi
early death, higher burden of disease, compromised physical capacities, and diminished cognitive development. Stunting and wasting in the first two years o
attainment and reduced economic productivity. This can reduce the productivity of an entire generation. Furthermore, stunting between 12 and 36 months is
school grades in middle childhood, and both height and head circumference at 2 years were shown to be inversely associated with educational attainment.

The ability to predict, at birth and early in childhood, whether a child is on an appropriate growth trajectory will help initiate preventive or therapeutic interv
development outcomes as determined by school performance and a thriving child- and adulthood.

Our goal is to determine a combination of early measures that would be a good predictor for recumbent length (length of child measured while child is lying d
pursuit of this goal, we have collected time series measurements of child growth, and family trait data (mother's age, mother's height, number of previous pr
We would like you to use this data to predict a child's weight, recumbent length, weight, and head circumference in the attached dataset where values have b

You may download the learning data set from here. The format for the data in the data set is a csv with details provided below:

Data Description

Column	Variable	Type	Label/Description	DV for prediction
1	UID	int	Unique child ID	No
2	AGEDAYS	float	Age sinc birth at examination (days)	No
3	GAGEDAYS	float	Gestational age at examination (days)	No
4	SEX	int	Sex, 1 = Male, 2 = Female	No
5	MUACCM	float	Mid upper-arm circumference (cm)	No
6	SFTMM	float	Skinfold thickness (mm)	No
7	BFED	int	Child breast fed at time of visist	No
8	WEAN	int	Child being weaned at time of visit	No
9	GAGEBRTH	float	Gestational age at birth in days	No
10	MAGE	float	Maternal age at examination (years)	No
11	MHTCM	float	Maternal height (cm)	No
12	MPARITY	int	Maternal parity	No
13	FHTCM	float	Fathers height (cm)	No
14	WTKG	float	Weight (kg)	Yes
15	LENCM	float	Recumbent length (cm)	Yes
16	HCIRCM	float	Head circumference (cm)	Yes

Each child is designated

- a unique id [column 1],
- sex [column 4],
- observed mid upper-arm circumference [column 5] and
- skinfold thickness [column 6],
- mother's age [column 10],
- mother's height [column 11],
- number of mother's previous pregnancies [column 12],
- mother's breast-feeding practices [column 7 and 8],
- and father's height[column 13], and
- multiple growth measurements:
  - weight [column 14],
  - recumbent length [column 15],
  - head circumference [column 16],

during early childhood growth (with the time variable provided as Age since birth in days [column 2] and age since conception in days [column 3] . The value '
measured and is therefore not available.

An example of measurements for a single child is given below:

UID	AGEDAYS	GAGEDAYS	SEX	MUACCM	SFTMM	BFED	WEAN	GAGEBRTH	MAGE	MHTCM	MPARI
550	-1.356576074	-1.274154148	2	.	.	.	.	1.614604045	0.112627355	-0.127853527	4
550	-0.865922259	-0.783913419	2	1.497903433	0.894389325	1	0	1.614604045	0.138831421	-0.127853527	4
550	-0.622766386	-0.540962261	2	1.776302755	2.593698371	.	.	1.614604045	0.159518841	-0.127853527	4
550	-0.014876703	0.066415633	2	1.915502416	3.040884962	1	1	1.614604045	0.211926973	-0.127853527	4
550	0.22827917	0.309366791	2	1.358703772	2.593698371	1	1	1.614604045	0.233993555	-0.127853527	4

For each prediction (*wi*, *li* and *ci*), where at least one of the DV values is missing, the error from the true Weight, Recumbent length and Head circumference wi

$$e_i = (w_i - w_{0i},\ l_i - l_{0i},\ c_i - c_{0i})S^{-1}(w_i - w_{0i},\ l_i - l_{0i},\ c_i - c_{0i})'$$

where S^{-1} is the inverse of the sample covariance matrix calculated on the complete dataset.

```
inverseS[0][0] = 11.90869495;   inverseS[0][1] = -7.523165469;  inverseS[0][2] = -4.11222794;
inverseS[1][0] = -7.523165469;  inverseS[1][1] = 13.5665806;    inverseS[1][2] = -4.742982596;
inverseS[2][0] = -4.11222794;   inverseS[2][1] = -4.742982596;  inverseS[2][2] = 8.669060303;
```

Scores will be calculated as a generalized R2 measure of fit. This is calculated as follows. The total sum of errors for the submission will be calculated as *SSE* =

A baseline sum of squared error will be calculated by predicting the sample means for each measurement, where at least one of the DV values is missing, that set,

$$e_{0i} = (\overline{w} - w_{0i},\ \overline{l} - l_{0i},\ \overline{c} - c_{0i})S^{-1}(\overline{w} - w_{0i},\ \overline{l} - l_{0i},\ \overline{c} - c_{0i})'$$

SSE0 = SUM(*e0i*)

Then the submission score will be *Score* = 1000000 * MAX(1 - *SSE*/*SSE0*, 0).

In the string[] **trainingData**, each string states a record of some measurement, and has 16 tokens, comma-separated, in the same order as described above i are presented as "." strings. You can assume that in **trainingData** all DV values are present. The format of **testingData** is almost the same as the **trainingDat** also replaced by "." strings, therefore your task will be to predict them. Replacement goes in the following way:

```
N = number of time points for an ID
X = random between 0 and N/2 inclusive
Y = random between X and N inclusive
foreach time point W(1..N) for an ID
  if W <= X then all three DV values present
  else if W <= Y then 'c' is replaced by "."
  else all three DV values are replaced by "."
```

The data with same IDs are consecutive and ordered by Agedays (time point). The returned string[] should contain the corresponding predictions for weight, this particular order, comma-separated, for each time point, in the same order as it is in **testingData**. The length of the return array equals to the number of r

NOTE: All data values are normalized between -6 and 6 as part of data obfuscation requirements.

Notes on Data Set Generation

- The full data set contains approximately 20,000 lines, covering almost 2000 ID values.
- The full data set is divided into 35% for example tests, 20% for provisional tests, and 45% for system tests. All data belonging to the same ID is placed in th
- For each test, approximately 66% of the data (from that segment) is selected for training, and the remainder for testing. Only for system tests, then we dro
- For provisional tests, all example data is also added to the training set.
- For system tests, all example and approximately 80% of provisional data is also added to the training set.

Definition

Class:	ChildStuntedness2
Method:	predict
Parameters:	String[], String[]
Returns:	String[]
Method signature:	String[] predict(String[] training, String[] testing)
(be sure your method is public)	

Notes

- The time limit is 5 minutes. The memory limit is 2048 megabytes.
- The compilation time limit is 30 seconds. You can find information about compilers that we use and compilation options here.
- Code snippets for calculate score and generate test case.
- There are 10 example test cases and 100 full submission (provisional) test cases.

Examples

0)

```
Seed: 1
```

1)

```
Seed: 2
```

2)

```
Seed: 3
```

3)

```
Seed: 4
```

4)

```
Seed: 5
```

5)

```
Seed: 6
```

6)

```
Seed: 7
```

7)

```
Seed: 8
```

8)

```
Seed: 9
```

9)

```
Seed: 10
```