Feature Selection : High Correlation

Common Dimensionality Reduction Techniques

- Missing value ratio
- Low variance
- High correlation
- Backward feature elimination
- Forward feature selection

Common Dimensionality Reduction Techniques

- Missing value ratio
- Low variance
- High correlation
- Backward feature elimination
- Forward feature selection

Common Dimensionality Reduction Techniques

- Missing value ratio
- Low variance
- High correlation
- Backward feature elimination
- Forward feature selection

Feature Selection : High Correlation

ID	Workout_time(minutes)	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	20	121	M	Yes	Fit
2	40	230	M	No	Fit
3	60	342	F	No	Unfit
4	5	70	M	Yes	Fit
5	48	278	F	Yes	Unfit
6	26	146	M	Yes	Fit
7	29	168	F	No	Unfit
8	45	231	F	Yes	Fit
9	30	150	M	No	Fit
10	35	190	F	No	Fit

Feature Selection : High Correlation

ID	Workout_time(minutes)	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	20	121	M	Yes	Fit
2	40	230	M	No	Fit
3	60	342	F	No	Unfit
4	5	70	M	Yes	Fit
5	48	278	F	Yes	Unfit
6	26	146	M	Yes	Fit
7	29	168	F	No	Unfit
8	45	231	F	Yes	Fit
9	30	150	M	No	Fit
10	35	190	F	No	Fit

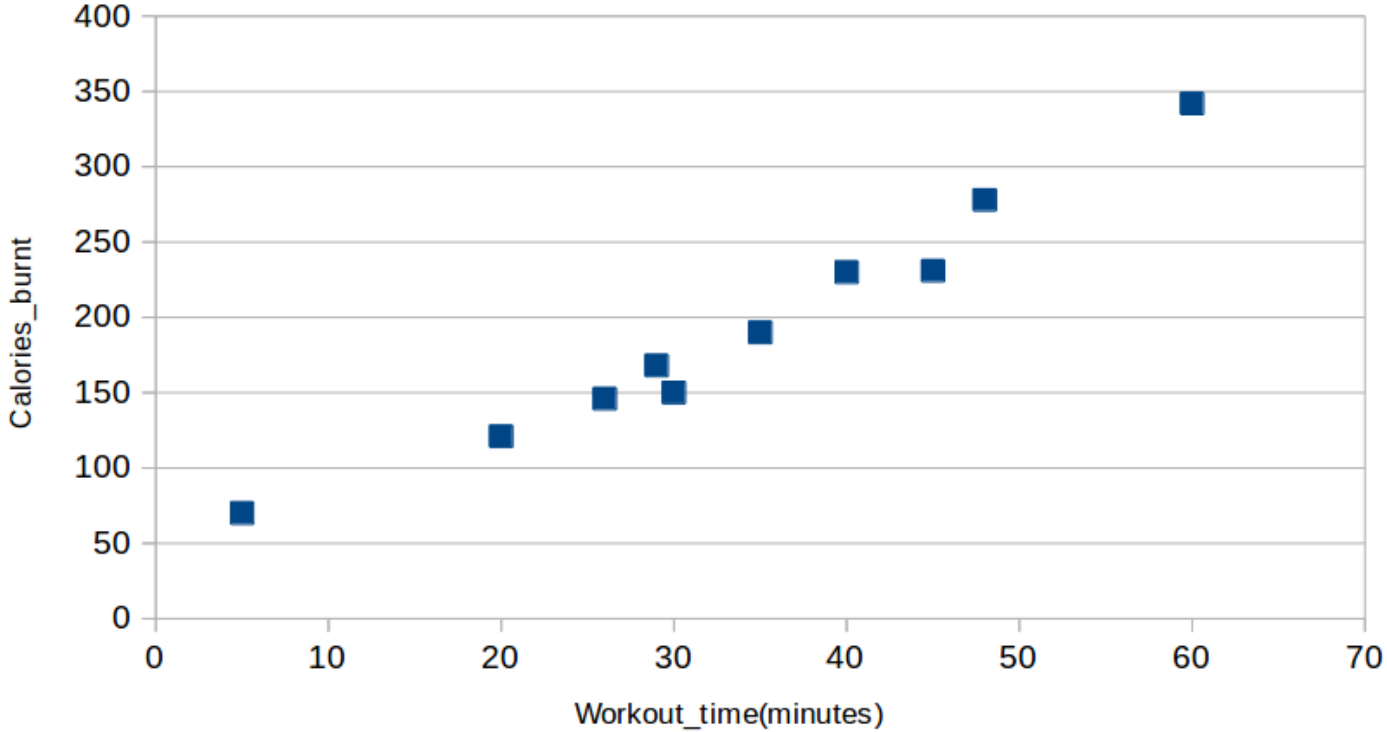
Feature Selection : High Correlation

ID	Workout_time(minutes)	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	20	121	M	Yes	Fit
2	40	230	M	No	Fit
3	60	342	F	No	Unfit
4	5	70	M	Yes	Fit
5	48	278	F	Yes	Unfit
6	26	146	M	Yes	Fit
7	29	168	F	No	Unfit
8	45	231	F	Yes	Fit
9	30	150	M	No	Fit
10	35	190	F	No	Fit

Feature Selection : High Correlation

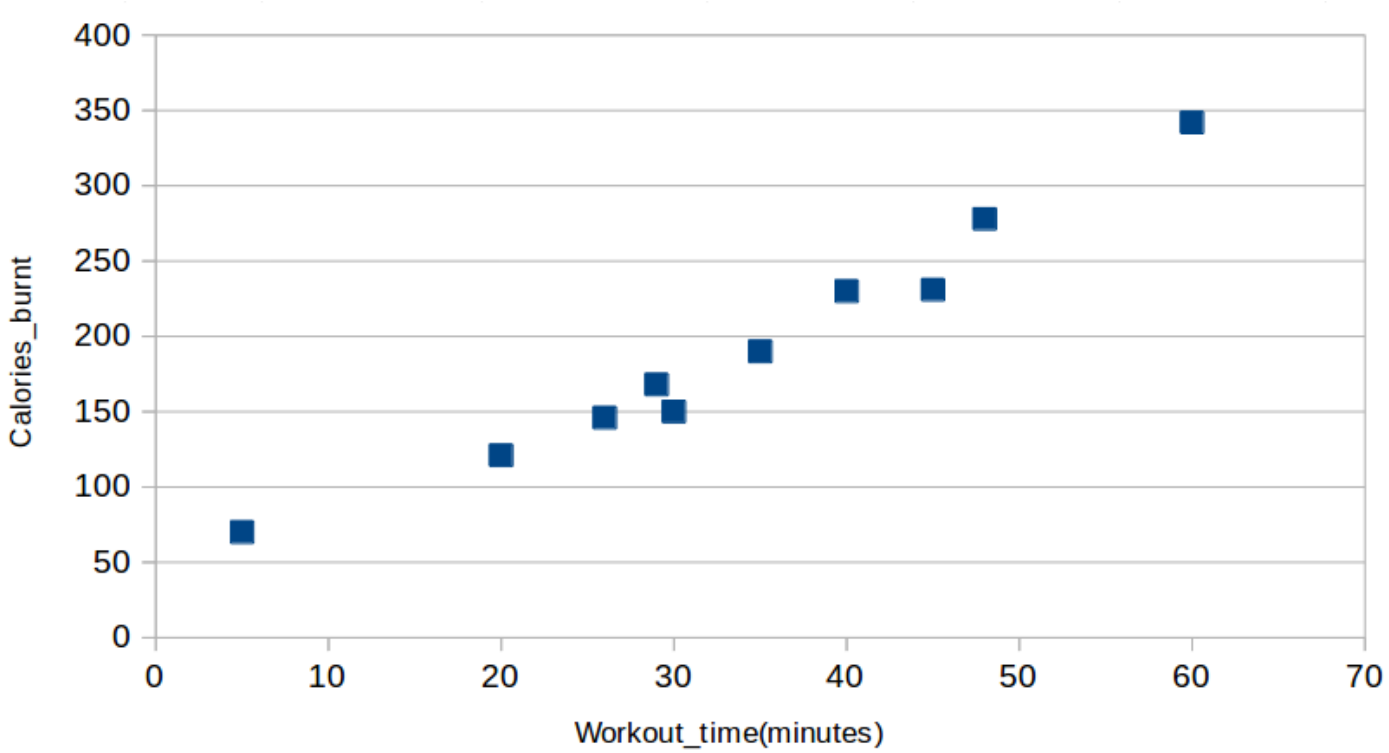
ID	Workout_time(minutes)	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	20	121	M	Yes	Fit
2	40	230	M	No	Fit
3	60	342	F	No	Unfit
4	5	70	M	Yes	Fit
5	48	278	F	Yes	Unfit
6	26	146	M	Yes	Fit
7	29	168	F	No	Unfit
8	45	231	F	Yes	Fit
9	30	150	M	No	Fit
10	35	190	F	No	Fit

Feature Selection : High Correlation



Analytics Vidhya
Learn everything about analytics

Feature Selection : High Correlation



Pearson Correlation = 0.9819

Analytics Vidhya
Learn everything about analytics

Feature Selection : High Correlation

Correlation

- Determines the relationship between two variables
- Ranges between -1 to 1
- Higher magnitude represents stronger relationship

Feature Selection : High Correlation

High Correlation  Multicollinearity

Feature Selection : High Correlation

- Calculate the correlation between all the independent variables
- Drop the variable, if the correlation crosses a certain threshold (generally 0.5-0.6)
- **Guideline** : Drop the variables which have less correlation with the target variable

Thank
You!

Analytics Vidhya
Learn everything about analytics