ANALISIS PENYAKIT KARDIOVASKULAR MENGGUNAKAN METODE KORELASI PEARSON, SPEARMAN DAN KENDALL

Yudi Eko Windarto
Departemen Teknik Komputer Fakultas Teknik Universitas Diponegoro
Jl. Prof Sudharto, SH – Tembalang Semarang
Email: yudi@live.undip.ac.id

ABSTRACT

Cardiovascular disease is a disease that is affected by the heart and blood vessels. There are several main risk factors that cause cardiovascular disease. Risk factors for cardiovascular disease include high blood pressure, high cholesterol, diabetes, being overweight or obese, age, sex, smoking, and alcohol. The more risk factors you have, the greater the chance of causing cardiovascular disease. In this research, it was developed using the Spearman, Pearson and Kendall correlation methods to analyze data on cardiovascular disease patients. The results showed there was correlation between blood pressure (ap_hi and ap_lo), age, and cholesterol had a strong relationship with cardiovascular disease. Glucose and cholesterol levels also have a strong relationship between one another.

***Keyword** : Cardiovascular Disease, Factor, Research and Correlation*

PENDAHULUAN

Penyakit kardiovaskular merupakan salah satu penyakit yang dipengaruhi oleh jantung dan pembuluh darah (D'Agostino et al., 2013). Terdapat beberapa faktor risiko utama yang menyebabkan penyakit kardiovaskular (Muggli et al., 2017). Faktor resiko penyakit kardiovaskular antara lain : tekanan darah tinggi, kolesterol tinggi, diabetes, kelebihan berat badan atau obesitas, usia, jenis kelamin, merokok, dan alkohol (Chen et al., 2017). Semakin banyak faktor risiko penyakit yang dimiliki pasien, semakin besar peluang untuk menyebabkan penyakit kardiovaskular (Hajar, 2016).

Penelitian terdahulu yang pernah dilakukan salah satunya membahas tentang diagnosis penyakit jantung coroner. Diagnosis penyakit jantung coroner tersebut menggunakan metode *Deep Neural Networks* menghasilkan nilai akurasi sebesar 83.67% (Miao & Miao, 2018).

Penelitian selanjutnya dengan menggunakan metode *machine learning* untuk memprediksi penyakit jantung (Jaymin & Tejal, 2016). Sistem yang dibuat tersebut menggunakan perangkat lunak WEKA.

Penelitian tentang prediksi penyakit jantung menggunakan *multiple linear regression model* (Polaraju et al.,

2017). Hasil pada penelitian tersebut bergantung pada jumlah atribut yang ada pada data dan metode yang digunakan.

Pada penelitian ini dikembangkan menggunakan metode Korelasi Spearman, Pearson dan Kendall untuk menganalisis data pasien penyakit kardiovaskular.

METODE

1. Korelasi

Analisis korelasi merupakan metode statistika yang digunakan dalam menentukan suatu besaran yang menyatakan adanya hubungan kuat pada suatu variabel dengan variabel yang lain (Uma & Roger, 2016). Apabila semakin tinggi nilai korelasi, semakin tinggi pula keeratan hubungan diantara kedua variabel. Apabila terdapat angka korelasi mendekati nilai satu, maka korelasi dari dua variabel akan semakin Kuat. Sebaliknya, jika angka korelasi mendekati nol maka korelasi dua variabel semakin lemah (Morris, 2020).

2. Korelasi Pearson

Korelasi Pearson adalah salah satu dari pengujian korelasi yang digunakan dalam mengetahui derajat keeratan hubungan dua variabel yang memiliki interval atau rasio, berdistribusi normal, serta mengembalikan nilai koefisien korelasi dengan rentang nilai antara -1, 0 dan 1 (Zhang et al., 2020). Nilai positif adalah nilai 1, nilai -1 merupakan nilai negatif, dan nilai 0 merupakan nilai yang tidak terdapat korelasi (Fu et al., 2020). Rumus dalam menentukan Korelasi Pearson ditunjukkan sebagai berikut.

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right)\right]^{1/2}} \quad (1)$$

3. Korelasi Spearman

Korelasi Spearman dipakai untuk pengukuran korelasi pada statistik *nonparametric* atau skala ordinal (Bin, Ruodu, dan Yuming, 2019). Korelasi tersebut merupakan ukuran korelasi yang dihubungkan oleh kedua variabel diukur sekurang kurangnya dalam skala ordinal sehingga obyek - obyek penelitiannya dapat diranking dalam dua rangkaian berurut (Andréas, dan Alfonso, 2020). Formula untuk korelasi Spearman ditunjukkan pada persamaan 2.

$$r = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{[\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2]^{\frac{1}{2}}} \quad (2)$$

4. Korelasi Kendall Tau

Jika terdapat data (X_i, Y_i) , i = 1,2, … , n di mana X dan Y setidaknya berskala ordinal. Sehingga pada setiap pasangan nilai (X_i, Y_i) dan (X_j, Y_j) untuk $i \neq j$

dapat didefinisikan pasangan nilai adalah sebagai berikut (Nugroho et al., 2008):

i). Pasangan (X_i, Y_i) dan (X_j, Y_j) konkordan , jika $(X_i - X_j)(Y_i - Y_j) > 0$ artinya adalah jika $X_i > X_j$ maka $Y_i > Y_j$ atau jika $X_i < X_j$ maka $Y_i < Y_j$ sehingga $(X - \bar{X})$ dan $(Y - \bar{Y})$ memiliki tanda yang sama, yaitu sama-sama positif atau sama-sama negatif dengan hasil kali yang selalu positif

ii). Pasangan (X_i, Y_i) dan (X_j, Y_j) diskordan , jika $(X_i - X_j)(Y_i - Y_j) < 0$ artinya adalah jika $X_i > X_j$ maka $Y_i < Y_j$ atau jika $X_i < X_j$ maka $Y_i > Y_j$ sehingga $(X - \bar{X})$ dan $(Y - \bar{Y})$ memiliki memiliki tanda yang berlawanan dengan hasil kali yang negative.

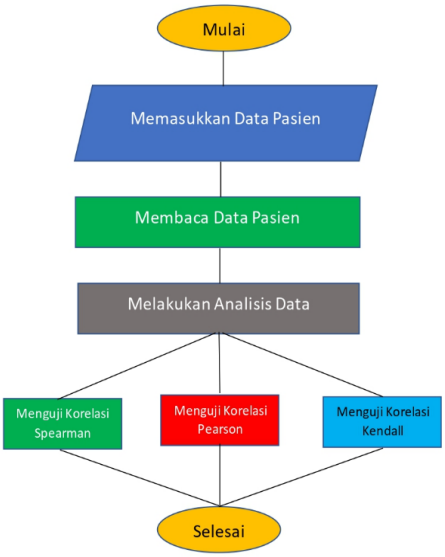
Rumus Korelasi Kendall dapat dihitung seperti persamaan 3 (Alexis & Dean, 2019).

$$\tau = \frac{2(C-D)}{n(n-1)} \quad (3)$$

HASIL DAN PEMBAHASAN

Hasil dan Pembahasan penelitian ini terdapat pada beberapa tahapan, diantaranya adalah sebagai berikut:

1. Memasukkan data pasien
2. Membaca data pasien
3. Melakukan analisa data
   a) Melakukan uji korelasi data menggunakan metode Spearman
   b) Melakukan uji korelasi data menggunakan metode Pearson
   c) Melakukan uji korelasi data menggunakan metode Kendall
   d) Tahapan penelitian ini ditunjukkan pada gambar 1.



Gambar 1 Tahapan Penelitian

4. Memasukkan data pasien

   Tahapan pertama yaitu mengunduh data cardiovascular berdasarkan data yang diperoleh dari dataset kaggle :

https://www.kaggle.com/datasets. Selanjutnya memasukkan data pasien menggunakan aplikasi Jupyter Nootebook.

5. Membaca Data Pasien

Tahapan yang kedua adalah membaca data pasien yang ada pada cardiovascular.csv. Data pasien cardiovascular ditunjukkan pada gambar 2.

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	17474	1	156	56.0	100	60	1	1	0	0	0	0
8	21914	1	151	67.0	120	80	2	2	0	0	0	0
9	22113	1	157	93.0	130	80	3	1	0	0	1	0
12	22584	2	178	95.0	130	90	3	3	0	0	1	1
13	17668	1	158	71.0	110	70	1	1	0	0	1	0
14	19834	1	164	68.0	110	60	1	1	0	0	0	0
15	22530	1	169	80.0	120	80	1	1	0	0	1	0
16	18815	2	173	60.0	120	80	1	1	0	0	1	0
18	14791	2	165	60.0	120	80	1	1	0	0	0	0
21	19809	1	158	78.0	110	70	1	1	0	0	1	0
23	14532	2	181	95.0	130	90	1	1	1	1	1	0
24	16782	2	172	112.0	120	80	1	1	0	0	0	1
25	21296	1	170	75.0	130	70	1	1	0	0	0	0

Gambar 2 Data Pasien

6. Informasi Data

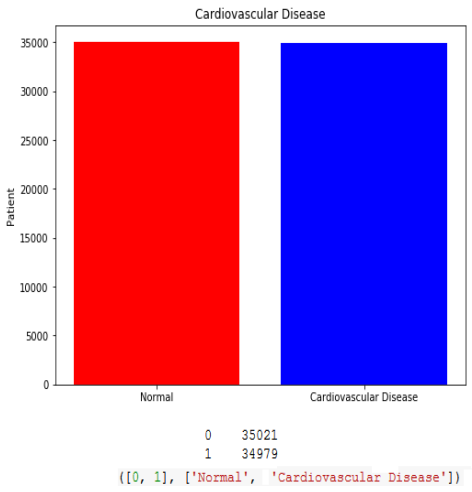
Pada data tersebut terdapat informasi 70.000 data pasien dan 13 kolom yang terdiri dari *id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active, cardio* ditunjukkan pada gambar 3.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
id              70000 non-null int64
age             70000 non-null int64
gender          70000 non-null int64
height          70000 non-null int64
weight          70000 non-null float64
ap_hi           70000 non-null int64
ap_lo           70000 non-null int64
cholesterol     70000 non-null int64
gluc            70000 non-null int64
smoke           70000 non-null int64
alco            70000 non-null int64
active          70000 non-null int64
cardio          70000 non-null int64
```

Gambar 3 Informasi Data Pasien

7. Melakukan Analisa Data

Pada tahapan ini, melakukan perbandingan jumlah pasien normal dengan pasien yang memiliki penyakit kardiovaskular. Dari 70.000 terdapat 35.021 jumlah pasien normal dan 34.979 data penderita kardiovaskular. Gambar 4 menunjukkan perbandingan antara pasien normal dengan pasien penderita penyakit kardiovaskular.

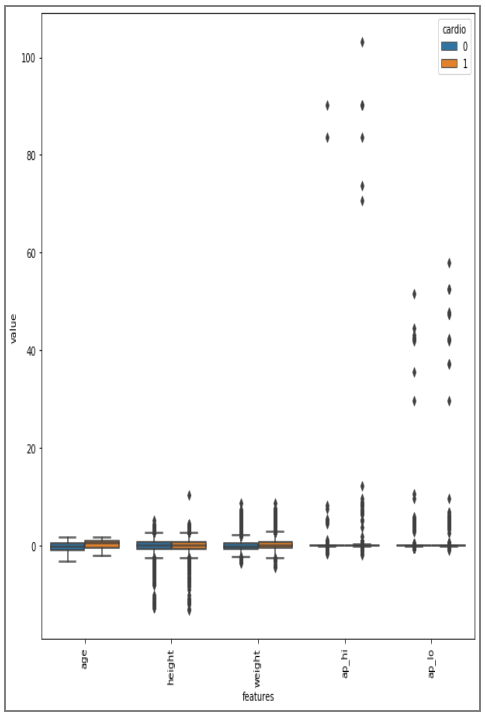


Gambar 4 Perbandingan Pasien Normal dengan Pasien Penderita Kardiovaskular

8. Mendeteksi Pencilan

Pada kolom *age, height, weight, ap_hi, ap_lo* ada kemungkinan memiliki pencilan. Sehingga untuk membandingkannya pada skala yang sama, perlu melakukan normalisasi data terlebih dahulu.

Setelah itu perlu meleburkan data untuk menampilkan *Multibox Graph Plot* seperti gambar 5.



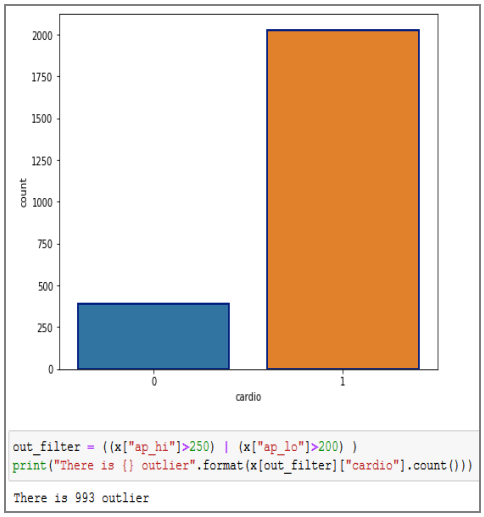
Gambar 5 *Multibox Graph Plot*

Berdasarkan gambar 5 diatas terdapat beberapa pencilan dalam dataset. Sehingga peneliti harus menentukan batas bawah dan batas atas yang ditunjukkan pada gambar 6.

	ap_hi	ap_lo
lower_bound	90.0	65.0
upper_bound	170.0	105.0

Gambar 6 Batas Bawah dan Batas Atas

Penyakit kardiovaskular terdapat 993 data pencilan dari *ap_hi* dan *ap_lo* di tunjukkan pada gambar 7.

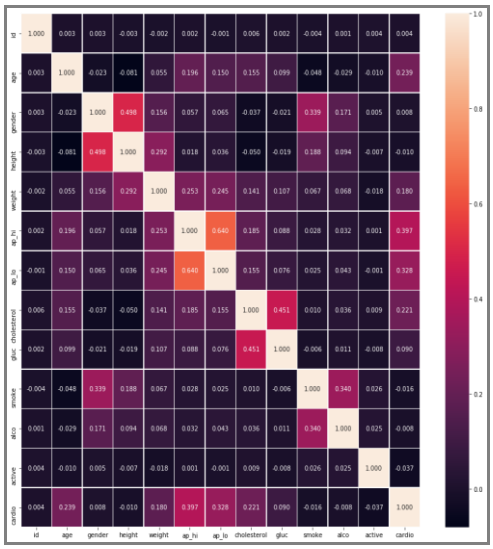


Gambar 7 Jumlah Data Pencilan

Karena *ap_hi* dan *ap_lo* merupakan tekanan darah, sehingga peneliti memutuskan untuk menghilangkan data yang tidak wajar secara medis dari dataset. Peneliti mengurangi beberapa data untuk mendapat nilai baru yang lebih ideal.

Gambar 8 menunjukkan bahwa terdapat korelasi antara kolesterol, tekanan darah *(ap_hi* dan *ap_low)* dan usia memiliki hubungan yang kuat dengan penyakit kardiovaskular.

Kadar gula darah dan kolesterol juga memiliki hubungan yang kuat di antara satu dengan yang lain dengan nilai 0,451.



Gambar 8 Korelasi Data

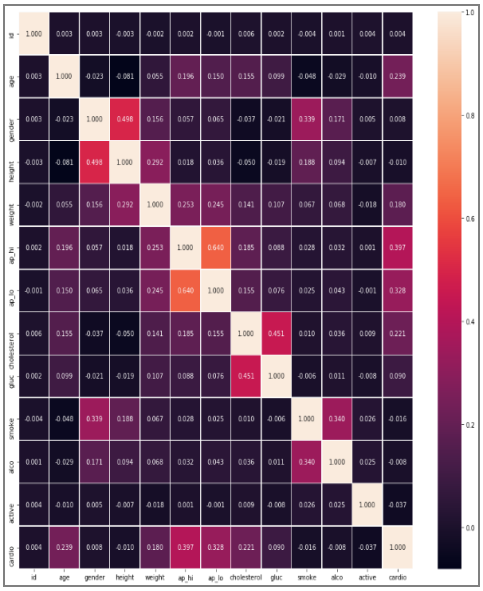
Berdasarkan pada gambar 8 tersebut ternyata yang berkorelasi dengan penyakit kardiovaskular adalah:

1. Tekanan darah sistole (ap_hi) dengan nilai = 0.397
2. Tekanan darah diastole (ap_lo) dengan nilai = 0.328
3. Usia dengan nilai = 0.239
4. Kolesterol dengan nilai = 0,221
5. Berat badan = 0.180
6. Kadar gula darah = 0.090

9. Korelasi Pearson

Pada pengujian menggunakan metode Pearson dihasilkan nilai korelasi dengan penyakit kardiovaskular adalah tekanan darah sistole (ap_hi) dengan nilai = 0.397, tekanan darah diastole (ap_lo) dengan nilai = 0.328, usia dengan nilai = 0.239, kolesterol dengan nilai = 0,221, berat badan dengan nilai = 0.180 dan kadar gula darah dengan nilai = 0.090.

Hasil pengujian korelasi menggunakan metode Pearson ditunjukkan pada gambar 9.

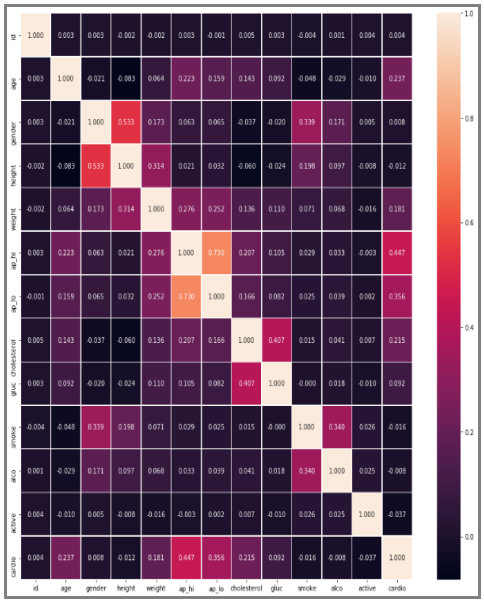


Gambar 9 Korelasi Pearson

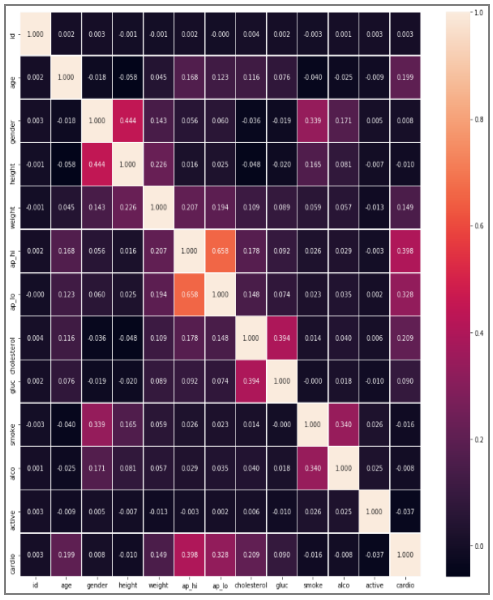
10. Korelasi Spearman

Berdasarkan pengujian menggunakan metode Spearman dihasilkan nilai korelasi dengan penyakit kardiovaskular adalah tekanan darah sistole (ap_hi) dengan nilai = 0.447, tekanan darah diastole (ap_lo) dengan nilai = 0.356, usia dengan nilai = 0.237, kolesterol dengan nilai = 0,215, berat badan dengan nilai = 0.181 dan kadar gula darah dengan nilai = 0.092.

Hasil pengujian korelasi menggunakan metode Spearman ditunjukkan pada gambar 10.



Gambar 10 Korelasi Spearman



Gambar 11 Korelasi Kendall

11. Korelasi Kendall

Setelah dilakukan pengujian menggunakan metode kendall dihasilkan nilai korelasi dengan penyakit kardiovaskular yaitu tekanan darah sistole (ap_hi) dengan nilai = 0.398, tekanan darah diastole (ap_lo) dengan nilai = 0.328, kolesterol dengan nilai = 0.209, usia dengan nilai = 0.199, berat badan dengan nilai = 0.149 dan kadar gula darah dengan nilai = 0.090.

Hasil pengujian korelasi menggunakan metode Kendall ditunjukkan pada gambar 11.

Hasil pengujian menggunakan metode Kendall memiliki perbedaan pada urutan korelasi di bandingan pada metode korelasi Pearson dan Spearman. Pada uji korelasi Pearson dan Spearman memiliki urutan sebagai berikut: tekanan darah sistole (ap_hi), tekanan darah diastole (ap_lo), usia, kolesterol, berat badan, dan kadar gula darah. Sebaliknya pada uji korelasi menggunakan metode kendall memiliki urutan sebagai berikut: tekanan darah sistole (ap_hi), tekanan darah diastole (ap_lo), kolesterol, usia, berat badan, dan kadar gula darah.

DAFTAR PUSTAKA

Alexis, D., & Jean, D, F. (2019). A classification point-of-view about conditional Kendall's tau. *Computational Statistics & Data Analysis*, *135*, 70–94. https://doi.org/10.1016/j.csda.2019.01.013.

Andréas, H., & Alfonso, V. (2020). Spearman rank correlation of the bivariate Student and scale mixtures of normal distributions. *Journal of Multivariate Analysis*, *179*, 1–11. https://doi.org/10.1016/j.jmva.2020.104650.

Bin, W., Ruodu, W., & Yuming, W. (2019). Compatible matrices of Spearman's rank correlation. *Statistics and Probability Letters*, *151*, 67–72. https://doi.org/10.1016/j.spl.2019.03.015

Chen, W. W., Gao, R. L., Liu, L. S., Zhu, M. L., Wang, W., Wang, Y. J., Hu, S. S. (2017). China cardiovascular diseases report 2015: A summary. *Journal of Geriatric Cardiology*, *14*(1), 1–10. https://doi.org/10.11909/j.issn.1671-5411.2017.01.012.

D'Agostino, R. B., Pencina, M. J., Massaro, J. M., & Coady, S. (2013). Cardiovascular disease risk assessment: Insights from Framingham. *Global Heart*, *8*(1), 11–23. https://doi.org/10.1016/j.gheart.2013.01.001.

Fu, T., Tang, X., Cai, Z., Zuo, Y., Tang, Y., & Zhao, X. (2020). Correlation research of phase angle variation and coating performance by means of Pearson's correlation coefficient. *Progress in Organic Coatings*, *139*(October 2019), 105459. https://doi.org/10.1016/j.porgcoat.2019.105459.

Hajar, R. (2016). Framingham contribution to cardiovascular disease. *Heart Views*, *17*(2), 78. https://doi.org/10.4103/1995-705x.185130.

Jaymin, P., Tejal, U., S. (2016). *Heart Disease Prediction Using Machine learning and Data Mining Technique*. *7*, 129–137. https://doi.org/10.090592/IJCSC.2016.018.

Miao, K. H., & Miao, J. H. (2018). Coronary heart disease diagnosis using deep neural networks. *International Journal of Advanced Computer Science and Applications*, *9*(10), 1–8. https://doi.org/10.14569/IJACSA.2018.091001.

Morris, A. (2020). A more scientific approach to applied economics: Reconstructing statistical, analytical significance, and correlation analysis. *Economic Analysis and Policy*, *66*, 315-324.

Muggli, F., Rabufetti, A., Simonetti, Gd., Bianchetti, Mg., Gallino, A. (2017). *CARDIOVASCULAR DISEASE RISK FACTORS AMONG MALE YOUTHS IN SOUTHERN SWITZERLAND: A TRANSVERSAL STUDY*. 2017.

Polaraju, K., Durga Prasad, D., & Tech Scholar, M. (2017). Prediction of Heart Disease using Multiple Linear Regression Model. *International Journal of Engineering Development and Research*, *5*(4), 2321–9939.

Uma, S., & Roger, B. (2016). Research Methods For Business: A Skill Building Approach. In *Encyclopedia of Quality of Life and Well-Being Research* (7th ed.). https://doi.org/10.1007/978-94-007-0753-5_102084.

Zhang, Y., Li, Y., Song, J., Chen, X., Lu, Y., & Wang, W. (2020). Pearson correlation coefficient of current derivatives based pilot protection scheme for long-distance LCC-HVDC transmission lines. *International Journal of Electrical Power and Energy Systems*, *116*(September 2019), 105526. https://doi.org/10.1016/j.ijepes.2019.105526.