International Journal of Multidisciplinary Research and Publications

ISSN (Online): 2581-6187

Comparation Logistic Regression and Decision Tree Method to Distribution Type of Works in Jakarta

Handy Noviyarto

Faculty of Computer Science, Mercu Buana University, Indonesia

Abstract— *In the digital era, the data is one of the components that are important in decision making. Data must be processed first so that it can be understood by the recipient data. The results of data processing is called information. In this study, the method used are Logistic Regression and Decision Tree. Both of these methods are included in the classification method. The purpose of this study was to determine the accuracy of the data from implementation of methods logistic regression and decision tree. This research was conducted using the Python programming language and the Visual Studio code.*

Keywords— *logistic regression, decision tree, classification.*

I. Introduction

In the digital era, the data is one of the components that are important in decision making. Data must be processed first so that it can be understood by the recipient data. The results of data processing is called information. Later this information to be used as a benchmark by a person, institution or company in decision making.

The increasing number of population and the existing technology, the amount of data is also growing and the information can be obtained from such data is becoming more diverse. Indonesia is one country with the highest number of inhabitants in the world. Most of the Indonesian population lives on the island of Java and Jakarta as the capital of Indonesia has a high population density.

Jakarta is one of the provinces in Indonesia which is the business center so that residents have a very diverse professions. The diversity of professions people living in Jakarta as well as administration areas were divided into six regions, create jobs data people living in Jakarta must be treated to information about the distribution of occupations by region of residence can be obtained.

Based on these descriptions, will be the classification of areas in Jakarta based on the type of work with methods of classification.

II. Platform Theory

2.1 Definition of Logistic Regression Method

Logistic regression, in statistics, is used to predict the probability of an event occurring by matching the data to the logit curve logistic function. This method is a general linear model used for binomial regression. Like regression analysis in general, this method uses several predictor variables, both numerical and category. For example, the probability that people who suffer a heart attack at a certain time can be predicted from information on age, sex, and body mass index.

Logistic regression is also used extensively in the fields of medicine and social sciences, as well as marketing such as prediction of customers' tendency to buy a product or unsubscribe.

$$\pi(x) = \frac{\exp(g(x))}{1 + \exp(g(x))}$$

Dimana:

π (x) = Proporsi terjadinya sebuah kejadian

$$g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

2.2. Definition of Decision Tree Method

Decision Tree (Decision Tree) is a tree in which each branch shows a choice among a number of alternative choices available, and each leaf shows the selected decision. Decision tree is usually used to obtain information for the purpose of making a decision. The decision tree starts with a root node (starting point) used by the user to take action. From this root node, the user breaks it down according to the decision tree algorithm. The end result is a decision tree with each branch showing possible scenarios of the decision taken and the results.

$$Entropi\ (S) = \sum_{j=1}^{k} -p_j \log_2 p_j$$

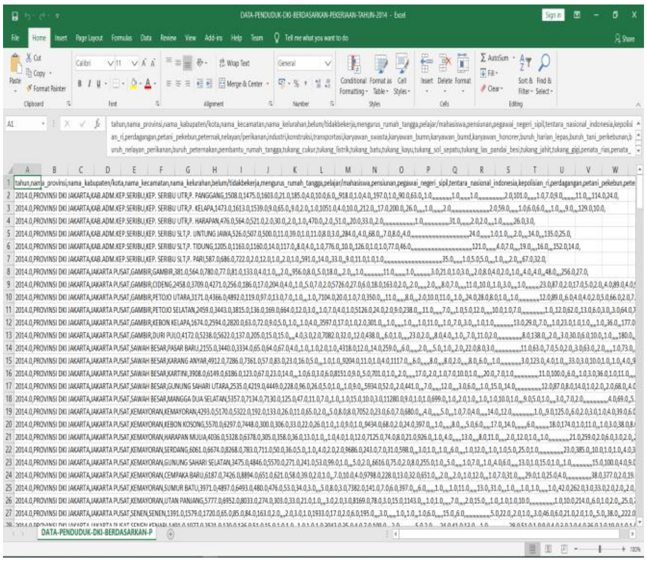
III. Results and Discussion

a) Data

Data were obtained from Jakarta Open Data, Jakarta Open Data is a website that provides the dataset associated with the data contents in Jakarta. The data obtained is the data in 2014. The data has a format (.csv) with a total of 267 rows and 95 columns, and the size of 86 kb. Here is a view of the data to be used for processing.

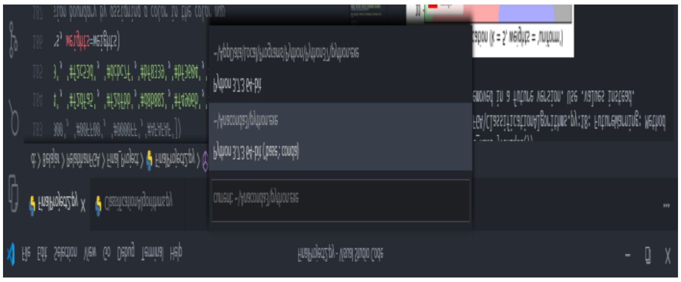
b) Device

Processing is done using the programming language Python 3.7.3 (based on anaconda) and the text editor of Visual Studio Code with python interpreter, as already mentioned.



DATA-PENDUDUK-DKI-BERDASARKAN-PEKERJAAN-TAHUN-2014 - Excel

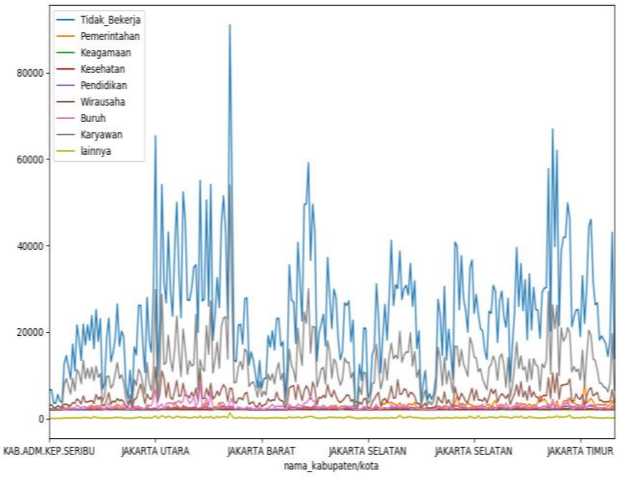
```
tahun,nama_provinsi,nama_kabupaten/kota,nama_kecamatan,nama_kelurahan,belum/tidakbekerja,mengurus_rumah_tangga,pelajar/mahasiswa,pensiunan,pegawai_negeri_sipil,tentara_nasional_indonesia,kepolisian_ri,perdagangan,petani_pekebun,peternak,nelayan/perikanan,industri,konstruksi,transportasi,karyawan_swasta,karyawan_bumn,karyawan_bumd,karyawan_honorer,buruh_harian_lepas,buruh_tani_perkebunan,buruh_nelayan_perikanan,buruh_peternakan,pembantu_rumah_tangga,tukang_cukur,tukang_listrik,tukang_batu,tukang_kayu,tukang_sol_sepatu,tukang_las_pandai_besi,tukang_jahit,tukang_gigi,penata_rias,penata_

1  tahun,nama_provinsi,nama_kabupaten/kota,nama_kecamatan,nama_kelurahan,belum/tidakbekerja,mengurus_rumah_tangga,pelajar/mahasiswa,pensiunan,pegawai_negeri_sipil,tentara_nasional_indonesia,kepolisian_ri,perdagangan,petani_pekebun,pete
2  2014.0,PROVINSI DKI JAKARTA,KAB.ADM.KEP.SERIBU,KEP. SERIBU UTR,P. PANGGANG,1508.0,1475.0,1603.0,21.0,185.0,4.0,10.0,6.0,,918.0,1.0,4.0,,197.0,1.0,,90.0,63.0,,1.0,,,,,,,1.0,,,,1.0,,,,,,,,,2.0,101.0,,,,1.0,7.0,9.0,,,,,11.0,,,114.0,24.0,
3  2014.0,PROVINSI DKI JAKARTA,KAB.ADM.KEP.SERIBU,KEP. SERIBU UTR,P. KELAPA,1473.0,1613.0,1539.0,9.0,65.0,,9.0,2.0,,1.0,1051.0,4.0,10.0,,212.0,,,17.0,200.0,,26.0,,,,1.0,,,,2.0,,,,,,,,,,,,,,2.0,59.0,,,,1.0,6.0,6.0,,,1.0,,,9.0,,,129.0,10.0,
4  2014.0,PROVINSI DKI JAKARTA,KAB.ADM.KEP.SERIBU,KEP. SERIBU UTR,P. HARAPAN,476.0,564.0,521.0,2.0,30.0,,2.0,,1.0,,470.0,,2.0,,51.0,,,20.0,33.0,,2.0,,,,,,,,,,,,1.0,,,,,,,,,,31.0,,,,,2.0,2.0,,,1.0,,,,,26.0,3.0,
5  2014.0,PROVINSI DKI JAKARTA,KAB.ADM.KEP.SERIBU,KEP. SERIBU SLT,P. UNTUNG JAWA,526.0,507.0,500.0,11.0,39.0,1.0,11.0,8.0,3.0,,284.0,,4.0,,68.0,,,7.0,8.0,,4.0,,,,,,,,,,,,,,,,,,,,,24.0,,,,,1.0,1.0,,,2.0,,,14.0,,,135.0,25.0,
6  2014.0,PROVINSI DKI JAKARTA,KAB.ADM.KEP.SERIBU,KEP. SERIBU SLT,P. TIDUNG,1205.0,1163.0,1160.0,14.0,117.0,,8.0,4.0,,1.0,776.0,,10.0,,126.0,1.0,1.0,77.0,46.0,,,,,,,,,,,,,,,,,,,,,121.0,,,,,4.0,7.0,,,,19.0,,,16.0,,,152.0,14.0,
7  2014.0,PROVINSI DKI JAKARTA,KAB.ADM.KEP.SERIBU,KEP. SERIBU SLT,P. PARI,587.0,686.0,722.0,2.0,12.0,,1.0,,2.0,1.0,,591.0,,14.0,,33.0,,,9.0,11.0,,1.0,1.0,,,,,,,,,,,,,,,,,,,,35.0,,,,1.0,5.0,5.0,,,1.0,,,2.0,,,67.0,32.0,
8  2014.0,PROVINSI DKI JAKARTA,JAKARTA PUSAT,GAMBIR,GAMBIR,381.0,564.0,780.0,77.0,81.0,133.0,4.0,1.0,,,2.0,,,956.0,8.0,,5.0,18.0,,,2.0,,,1.0,,,,,,,,11.0,,,,,1.0,,,,,,,3.0,21.0,1.0,3.0,,,2.0,8.0,4.0,2.0,,1.0,,,4.0,,4.0,,48.0,,,256.0,27.0,
9  2014.0,PROVINSI DKI JAKARTA,JAKARTA PUSAT,GAMBIR,CIDENG,2458.0,3709.0,4271.0,256.0,186.0,17.0,204.0,4.0,,1.0,,5.0,7.0,2.0,5726.0,27.0,6.0,18.0,163.0,2.0,,,2.0,,,,2.0,,,8.0,7.0,,,,11.0,,10.0,,1.0,3.0,,,1.0,,,,,,23.0,87.0,2.0,17.0,5.0,2.0,,4.0,89.0,4.0,
10 2014.0,PROVINSI DKI JAKARTA,JAKARTA PUSAT,GAMBIR,PETOJO UTARA,3171.0,4366.0,4892.0,119.0,97.0,13.0,7.0,,1.0,,,1.0,,7104.0,20.0,1.0,7.0,350.0,,,11.0,,,,8.0,,,2.0,10.0,11.0,,,1.0,,,24.0,28.0,8.0,1.0,,,1.0,,,,,,,,12.0,89.0,,6.0,4.0,4.0,2.0,5.0,66.0,2.0,7.
```



c) *Modeling*

Data processing method used is the classification. Classification method has several algorithms for pengolahaannya. Data processing was performed using several modeling ie, Logistic Regression, Decision Tree, K-Nearest Neighbors, Linear Discriminant Analysist, Gaussian Naive Bayes and Support Vector Machine. Many modeling performed to determine the ratio between the model and get the best data accuracy of some types of modeling. The accuracy of the data that will either maximize data processing is done.

d) *Data visualization*



3.1. Process

The process is done consists of several stages of preprocessing, modeling, visualization.

a) preprocessing

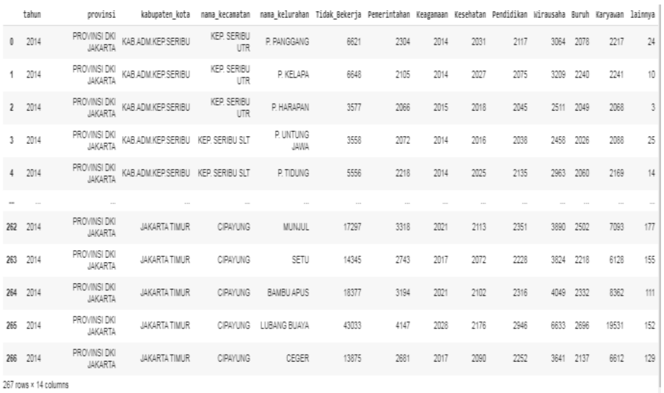
Works are grouped by area of work as shown in the following table.

TABLE 3.1 Public Sector

No.	Field of work	Profession
1	Does not work	Does not work
		Student / Student
		Retired
		Taking care of household
2	Government	Government employees
		Indonesian national army
		⋮
		employees BUMD
		employees Honorer
⋮	⋮	⋮
8	employee	Industry
		Construction
		⋮
		reporter
		General employees
9	more	more

The data have some fields empty because in some regions are not found type of work, so the use of charging a missing value with a value of 0 for each field empty.

Furthermore, after the data for each type of work complete, the merger jobs that have the appropriate work field in Table 3.1.



	tahun	provinsi	kabupaten_kota	nama_kecamatan	nama_kelurahan	Tidak_Bekerja	Pemerintahan	Keagamaan	Kesehatan	Pendidikan	Wirausaha	Buruh	Karyawan	lainnya
0	2014	PROVINSI DKI JAKARTA	KAB.ADM.KEP.SERIBU	KEP. SERIBU UTR	P. PANGGANG	6621	2304	2014	2031	2117	3064	2078	2217	24
1	2014	PROVINSI DKI JAKARTA	KAB.ADM.KEP.SERIBU	KEP. SERIBU UTR	P. KELAPA	6648	2105	2014	2027	2075	3209	2240	2241	10
2	2014	PROVINSI DKI JAKARTA	KAB.ADM.KEP.SERIBU	KEP. SERIBU UTR	P. HARAPAN	3577	2066	2015	2018	2045	2511	2049	2068	3
3	2014	PROVINSI DKI JAKARTA	KAB.ADM.KEP.SERIBU	KEP. SERIBU SLT	P. UNTUNG JAWA	3558	2072	2014	2016	2038	2458	2026	2088	25
4	2014	PROVINSI DKI JAKARTA	KAB.ADM.KEP.SERIBU	KEP. SERIBU SLT	P. TIDUNG	5556	2218	2014	2025	2135	2963	2060	2169	14
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
262	2014	PROVINSI DKI JAKARTA	JAKARTA TIMUR	CIPAYUNG	MUNJUL	17297	3318	2021	2113	2351	3890	2502	7093	177
263	2014	PROVINSI DKI JAKARTA	JAKARTA TIMUR	CIPAYUNG	SETU	14345	2743	2017	2072	2228	3824	2218	6128	155
264	2014	PROVINSI DKI JAKARTA	JAKARTA TIMUR	CIPAYUNG	BAMBU APUS	18377	3194	2021	2102	2316	4049	2332	8362	111
265	2014	PROVINSI DKI JAKARTA	JAKARTA TIMUR	CIPAYUNG	LUBANG BUAYA	43033	4147	2028	2176	2946	6633	2696	19531	152
266	2014	PROVINSI DKI JAKARTA	JAKARTA TIMUR	CIPAYUNG	CEGER	13875	2681	2017	2090	2252	3641	2137	6612	129

267 rows × 14 columns

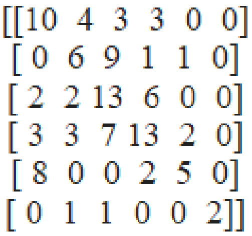
Furthermore, namely the determination of features and label. Features that used that line of work, amounting to 9 features with kabupaten_kota label. Variables feature was added to the variables x and labels addedthe variable y. then performed the percentage split to divide the training data and testing to be used. The percentage of testing the data used by 40% taken at random.

3.2. Implementation Method

The method used are Logistic Regression and Decision Tree.

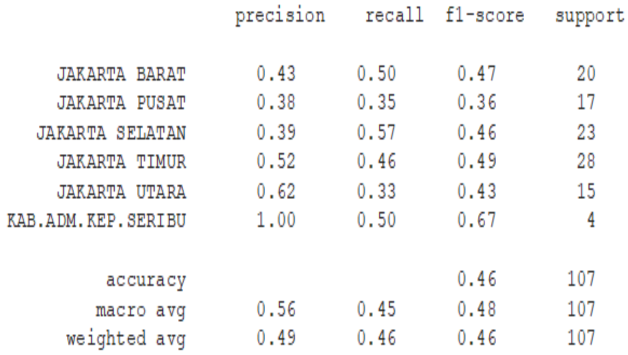
a. *Logistic Regression*

Classifying the research subject based on the probability threshold. The model used uses optimizer = 'lbfgs' to overcome loss for multi-class and use multi-class = 'auto' according to the optimizer used. Next, the fitting process is carried out on the model using the x train and y train parameters. The results of the modeling process using the Logistic Regression method produce an accuracy value of 56% for the training set and 46% with the value of the confusion matrix as follows:



```
[[10  4  3  3  0  0]
 [ 0  6  9  1  1  0]
 [ 2  2 13  6  0  0]
 [ 3  3  7 13  2  0]
 [ 8  0  0  2  5  0]
 [ 0  1  1  0  0  2]]
```

Confusion matrix can be used to generate the evaluation of the following classifications:



```
                   precision    recall  f1-score   support

    JAKARTA BARAT       0.43      0.50      0.47        20
    JAKARTA PUSAT       0.38      0.35      0.36        17
  JAKARTA SELATAN       0.39      0.57      0.46        23
    JAKARTA TIMUR       0.52      0.46      0.49        28
    JAKARTA UTARA       0.62      0.33      0.43        15
KAB.ADM.KEP.SERIBU      1.00      0.50      0.67         4

         accuracy                           0.46       107
        macro avg       0.56      0.45      0.48       107
     weighted avg       0.49      0.46      0.46       107
```

Visualization label comparison of actual and predicted label shown in the following:

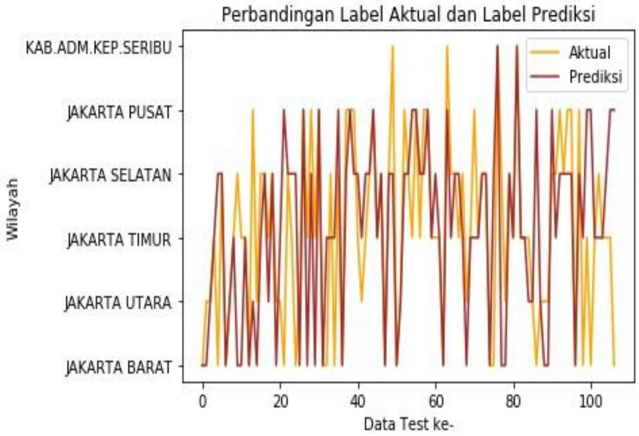
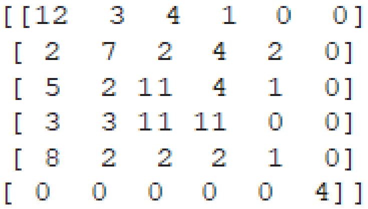


Figure 3.1 Comparison of the Results of Logistic Regression Methods

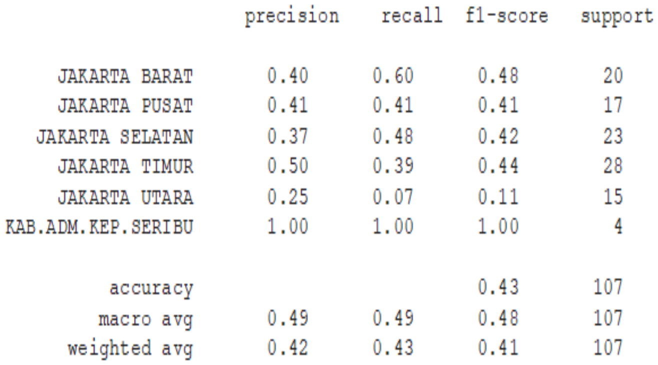
b. *Decision Tree*

Decision tree is a predictive model using a tree structure or hierarchical structure. The concept of the decision tree is to transform data into decision tree and decision rules. The model used a decision tree using the default parameters in python. Furthermore, the process of fitting in the model using the parameters x and y train. The results of the modeling process using the Decision Tree generate value accuracy of 100% for training set and 43% for testing set by the confusion matrix values as follows:



```
[[12  3  4  1  0  0]
 [ 2  7  2  4  2  0]
 [ 5  2 11  4  1  0]
 [ 3  3 11 11  0  0]
 [ 8  2  2  2  1  0]
 [ 0  0  0  0  0  4]]
```

Confusion matrix can be used to generate the evaluation of the following classifications:



```
                   precision    recall  f1-score   support

    JAKARTA BARAT       0.40      0.60      0.48        20
    JAKARTA PUSAT       0.41      0.41      0.41        17
  JAKARTA SELATAN       0.37      0.48      0.42        23
    JAKARTA TIMUR       0.50      0.39      0.44        28
    JAKARTA UTARA       0.25      0.07      0.11        15
KAB.ADM.KEP.SERIBU      1.00      1.00      1.00         4

         accuracy                           0.43       107
        macro avg       0.49      0.49      0.48       107
     weighted avg       0.42      0.43      0.41       107
```

Visualization label comparison of actual and predicted label shown in the following:

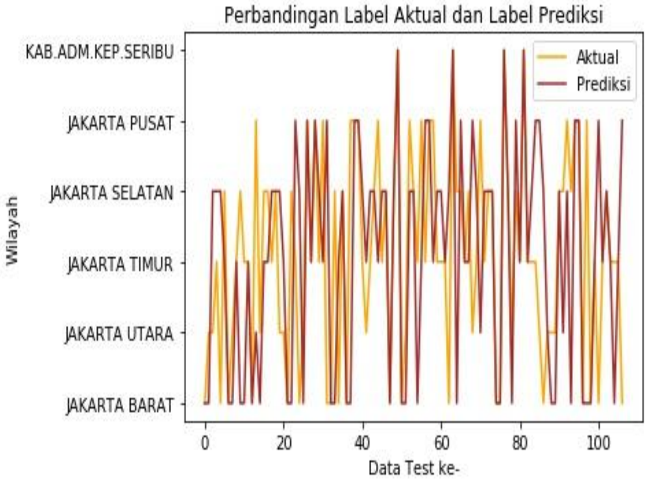


Figure 3.2 Comparison of results on Decision Tree Method

Comparison of Logistic Regression and Decision Tree methods that have been done can be seen in the following table:

TABLE 3.2 Comparison of Accuracy of Method

No.	Method	accuracy	
		training	testing
1	Logistic Regression	56%	46%
2	Decision Tree	100%	43%

IV. CONCLUSION

Based on the results of several methods can be concluded that:

1. The method is good enough Logistic Regression testing which resulted in an accuracy of 46% means that a predictive model has been pretty good.

2. Distribution of the work that most of the field work is not work (yet / Not Working, Taking Care of Household, Student / Students and Pensioners) and is located in West Jakarta.
3. Distribution fewest jobs are more job fields and lies in the territory of all regions.

REFERENCES

[1] Hilbe, Joseph M. (2009). Logistic Regression Models. Chapman & Hall/CRC Press.

[2] Nia Rahma Kurnianda & Yunita Sartika Sari. Analysis and Design of Information System for Self-Journal on Food Based Dietary Assessment Record for Diabetes Patients. International Research Journal of Computer Science (IRJCS). Volume 06 Issue 5. 2018

[3] Ranggadara, Indra & Suhendra "Naive Bayes Algorithm with Chi Square and NGram Feature for Reviewing Laptop Product on Amazon Site.".IRJCS, Vol 7, issue 2.2018

[4] William H. Kruskal and Judith M. Tanur, ed. (1982), "Linear Hypotheses," International Encyclopedia of Statistics. Free Press, v. 1,

[5] Lindley, D.V. (1987). "Regression and correlation analysis," New Palgrave: A Dictionary of Economics, v. 4, pp. 120–23.

[6] Birkes, David and Yadolah Dodge, Alternative Methods of Regression. ISBN 0-471-56881-3

[7] Chatfield, C. (1993) "Calculating Interval Forecasts," Journal of Business and Economic Statistics, 11. pp. 121–135.

[8] Corder, G.W. and Foreman, D.I. (2009).Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach Wiley, ISBN 978-0-470-45461-9

[9] Draper, N.R. and Smith, H. (1998).Applied Regression Analysis Wiley Series in Probability and Statistics

[10] Fox, J. (2017). Applied Regression Analysis, Linear Models and Related Methods. Sage

[11] Hardle, W., Applied Nonparametric Regression (1990), ISBN 0-521-42950-1

[12] Meade, N. and T. Islam (1995) "Prediction Intervals for Growth Curve Forecasts," Journal of Forecasting, 14, pp. 413–430.