International Journal of Computer Applications Technology and Research
Volume 9–Issue 12, 327-331, 2020, ISSN:-2319–8656

Optimization of Missing Value Data Imputation Automatic Dependent Surveillance Broadcasting (ADS-B) Based on K-Nearest Neighbor and Genetic Algorithm

Didik Hariyanto
Electrical Engineering Department,
Faculty of Engineering,
Brawijaya University.
Malang, East Java, Indonesia

Sholeh Hadi Pramono
Electrical Engineering Department,
Faculty of Engineering,
Brawijaya University
Malang, East Java, Indonesia

Erni Yudaningtyas
Electrical Engineering Department,
Faculty of Engineering,
Brawijaya University
Malang, East Java, Indonesia

Abstract: The flight navigation equipments technology use still conventional, namely using radar, now slowly starting to switch to Automatic Dependent Surveillance-Broadcast (ADS-B [6]. In this study, using RTL-SDR to detect aircraft and carry out tests through the Monte Carlo alltitude method, latitude, and longitude only [3]. However, in this system there is a problem regarding the missing value in the preprocessed data results / ADS-B flow data. In handling missing values, the KNN method is the most popular, but the weakness in the KNN method, can reduce the performance[9]. So a Genetic Algorithm (GA) is proposed to optimize the k value in the KNN method. The results of this study obtained a better MSE value in the imputation process. Altitude k = 3, with MSE 128668.96, Speed k = 6, with the MSE value = 457.5201, while the k value in the Heading variable k = 61 with MSE = 752.1429. For Lattitude and Longitude, the value of k = 3, MSE 9.16E-05 and k = 2 and MSE 1.68E-05.

Keywords: ADS-B, Missing Value, Imputation, K-Nearest Neighbor, Genetic Algotirhm

1. INTRODUCTION

The Air transportation safety is an important and major factor in the operation of flight services including flight navigation services. Meanwhile, flight navigation services can be provided maximally by airport operators when supported by good airport facilities. Along with technological developments that are increasingly sophisticated day after day, supporting facilities for flight navigation services are growing rapidly. The use of aviation navigation equipment technology, which was initially still conventional, namely radar, is now slowly starting to switch to Automatic Dependent Surveillance-Broadcast (ADS-B) [1].

With the development of communication technology, one way to get ADS-B data is use a Software-Defined Radio (SDR), namely the Mini USB RTL-SDR receiver using a new IC tuner, the R820T2 made by Rafael Micro. SDR technology was first introduced in 1991 by Joseph Mitola. RTL-SDR has a wide frequency range, with frequencies ranging from 25 Mhz - 1750Mhz. So that you can listen to all radio activities in that range and in the form of other data [2].

By integrating RTL-SDR, and dump 1090 as decoder software on the Rasbery Pi3, the ADS-B receiver is relatively more efficient and inexpensive[5]. The Research conducted by Akshay, N et al in 2017, which produced a fairly complete ADS-B flight data, was carried out using the RTL-SDR. The results of this study are information on altitude, position, speed, direction, and other information to ground stations and other aircraft. In this study, using RTL-SDR only to detect aircraft and carry out tests through the Monte Carlo method of latitude, longitude and latitude [3]. However, in the system, there is a problem about missing value in the results of the preprocessing data / ADS-B flow data.[6]

Imputation is filling in the missing value (empty data) with a certain value. [7]The rule of imputation is to get the predicted value as close as possible to the missing value, in other words imputation tries to minimize the value between the missing value and the predicted value of the missing value [4].

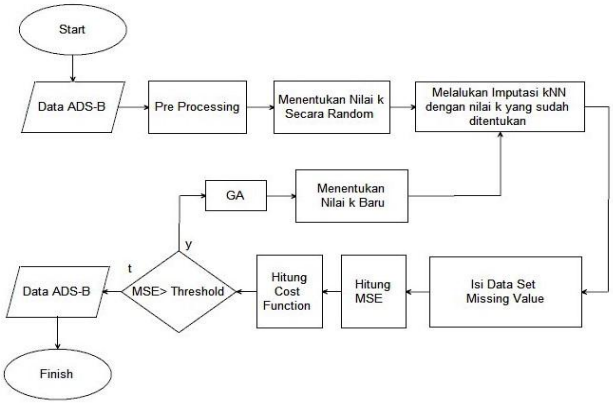
In handling missing data, KNN is the easiest and most popular method. However, this method has several drawbacks, one of which is that incorrect selection of *k* values can reduce classification performance [9]. So that a genetic algorithm is proposed to optimize the k value in the KNN so that it can produce a good estimated value with the smallest possible MSE[8]. This, the classification results will be obtained with high accuracy. This study aims to deal with missing data with imputation techniques using a combination of the KNN and GA (KNN-GA) algorithms.

Based on the previous studies, in this study an alternative solution that can be given in dealing with ADS-B data which includes heading, speed, longitude, altitude, latitude where missing values are found is by means of imputation. Meanwhile, to overcome the weaknesses of the k-NN method, namely by increasing the value of k using Genetic Algorithm (GA). By conducting research on the imputation of missing values, it is hoped that the data will be more accurate and correct, so that the information that will be provided to ATC in carrying out its duties as air traffic guides has the integrity of the information.

2. METHOD

The concept of the method in this study is shown in Figure 1 below.

Figure 1. Research Method



2.1 Data Collection

The data collection mechanism uses a model like Figure 2 below.

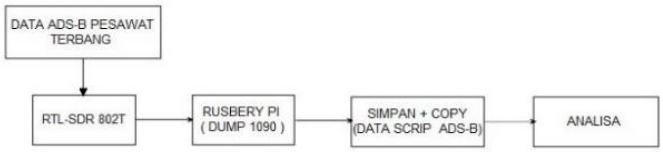


Figure 2. ADS-B Data Retrieval

1) Aircraft ADS-B data is taken by making a groundstation consisting of colinear antennas, RTL-SDR and Rasberry Pi 3 which are integrated with Linux OS and dump1090 decoder whose function is to convert analog data into binary form.
2) After the data is collected and stored in the .csv file, then save the ADS-B data in the memory embedded in the Rasberry Pi 3
3) Copy the ADS-B data file to a PC / laptop for data analysis.
4) Convert ADS-B data to normalize data.

2.2 *k-Nearest Neighbor (*K-NN) Method

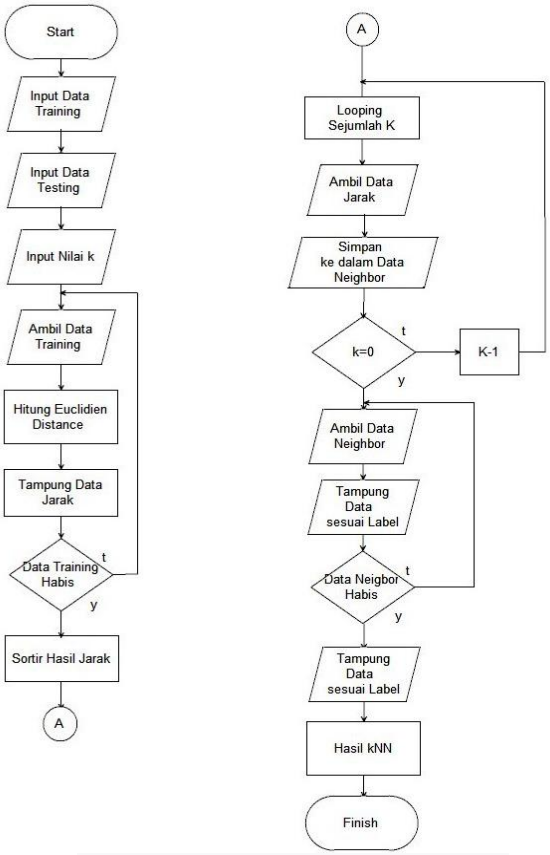


Figure 3. Flow diagram of the KNN Method

From Figure 3 it can be explained that the K-Nearest Neighbor Method Flowchart is as follows:

1) Take the missing value data from the .csv file, and sort the valuable data into training data
2) Take blank data to be used as testing data from the .csv file
3) Determine the initial K value for the KNN algorithm
4) Of the many training data, take the training data one by one to be compared with the relevant testing data
5) Calculating the distance from training data and testing data with the Euclidean Distance formula
6) The results from this distance are accommodated in certain variables so that they can be processed
7) If the training data is still not finished, take new training data and recalculate the value of Euclidean Distance
8) Of these data. If the training data has been taken out, then proceed to the previous process.
9) Sorting distance storage variables from the lowest value to the farthest distance, this is used because basically KNN looks for the data closest to the training data
10) The results of the distance are not taken all but a number of K values are taken. To retrieve this data, it is necessary to loop a predetermined number of K values.
11) Retrieving distance data one by one in sequence.
12) Save the retrieved data into neighbor variables.
13) If the value of k is still not used up, then reduce the value of K by 1 and return the looping value to k. If the value of K has run out then continue to the next process
14) Taking on the neighbors data one by one.
15) Hold the neighbors data according to the data label. Data label is a classification of data or output data.
16) If the neighbors' data has not run out, then take the neighbors data continuously, if it runs out then continue to the next process.
17) Hold labeled data into variables.
18) The results of the KNN are in the first serial number data with a predefined label.

2.3 Genetic Algorithm (GA) Method

The flow diagram work steps of the Genetic Algorithm (GA) are as follows in Figure 4:

1) Determining the GA input value, the GA input is the data before the missing value is found and the data after
2) Missing value. the data value which will later also be a benchmark for fitness values
3) Determine the number of population. The number of population is obtained from the solution per population (chromosome) and the number of input data (gene)
4) Because at the beginning of the iteration the best value is still unknown, the value of the population is a random value with predetermined limits.
5) Determine the number of generations or the number of iterations. The number of early iterations is used for how long the algorithm is executed, the more generations the algorithm will run, but the resulting data can be better.
6) Calculating the fitness value or cost function. This function is used to find out how suitable / good the data is in the population
7) The results of the fitness value are accommodated in a variable. This variable will be filtered again to get the most optimal value
8) From the fitness data, take the best data to be used as a parent (parent). Parent data is taken more than 2 data
9) Parent data is combined with a crossover function. This crossover function divides the parent data in half. The first half of the parent data is taken to be combined with the second half of the parent data.
10) Mutation process is used to prevent the same data on the population. From the crossover data, add the random value.
11) The results of the mutation data and parent data enter the new population
12) Looping to step 5 if the generation is not yet 0
13) Recalculate the fitness value of the population formed from iterations.
14) The data that has the highest fitness value will be taken for the results of the GA algorithm

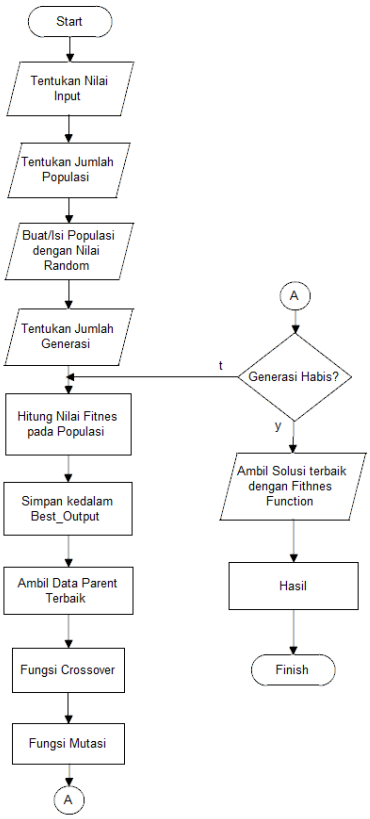


Figure 4. Flowchart of the Genetic Algorithm (GA) Method

3. RESULT AND DISCUSSION

3.1 Data Collection Results

In conducting the research, the researchers collected ADS-B data on civil aircraft at Abdulrahman saleh airport which was taken by making a ground station consisting of collinear antennas, RTL-SDR and Rasberry Pi 3 which have been integrated with Linux OS and dump1090 decoder whose function is to convert analog data into binary form. After the data is collected and stored in the form of a .csv file, then storing the ADS-B data in the memory embedded in the Rasberry Pi 3. Copying the ADS-B data file to a laptop for data analysis. Convert ADS-B data to normalize data. Tried on ADSB data. The following is the data acquisition result in Figure 7. then in this research a method for filling in the blank data will be presented by trying to use the KNN method and optimizing using the GA method.

Table 1. Missing Value on ADSB data for AbdulrahmanSaleh Airport.

Time	*Fligth*	*Variabel*				
		Altitude	*Speed*	*Heading*	*Lattitude*	*Longitude*
Missing value						
12:25:10 AM	Sjy3131	34975			-7.80826	112.96259
12:25:12 AM	Sjy3131	?	?	?	?	?
12:25:12 AM	Sjy3131	35000	?	?	?	?
12:25:13 AM	Sjy3131	?				
12:25:15 AM	Sjy3131	?	459	126		
12:25:16 AM	Sjy3131	34975			-7.8157	112.97274
12:25:16 AM	Sjy3131		459	126		
12:25:17 AM	Sjy3131		?	?	?	?
12:25:17 AM	Sjy3131	35000	?	?	?	?
12:25:18 AM	Sjy3131		458	126		
12:25:19 AM	Sjy3131	35000			-7.81947	112.97791
12:25:20 AM	Sjy3131	35000			-7.82073	112.97961
12:25:20 AM	Sjy3131		458	126		

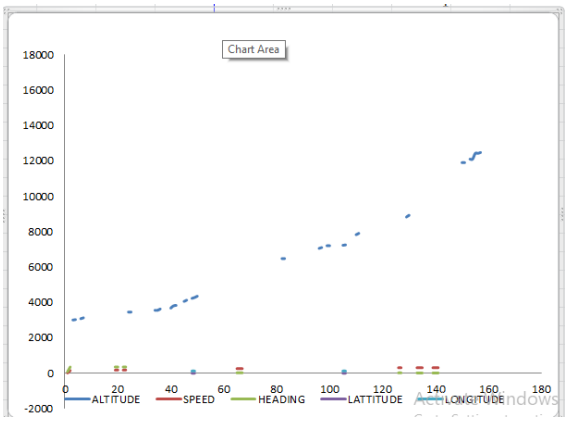
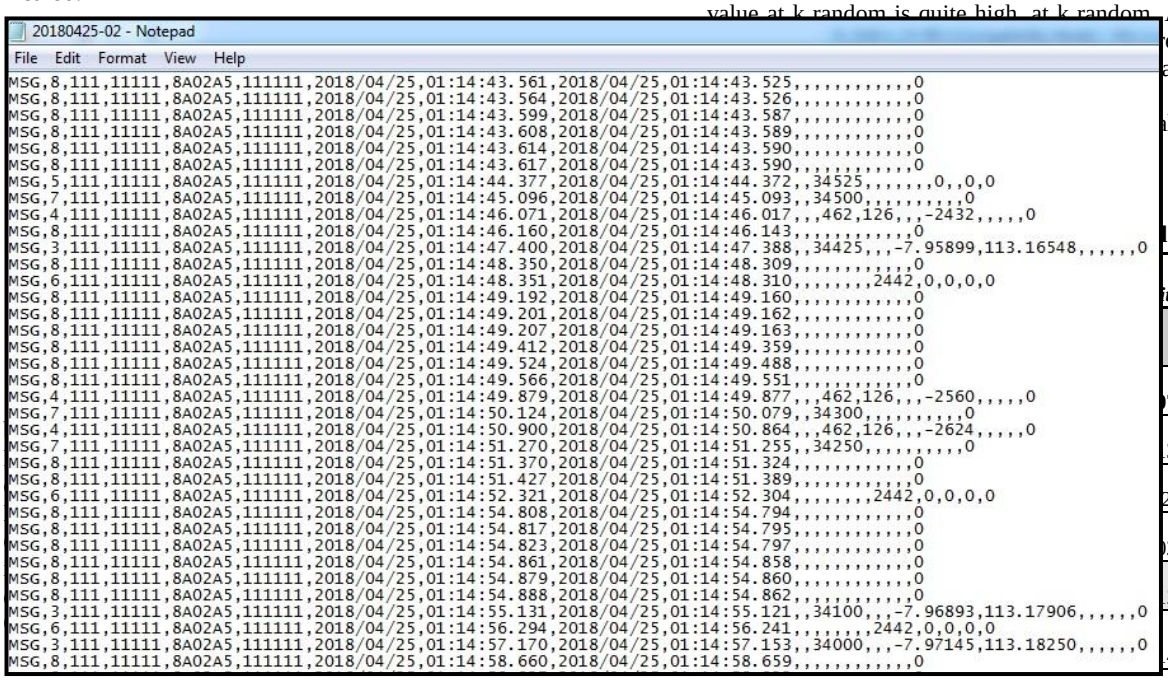


Figure. 6. Graphic display of Sriwijaya Air aircraft data which is still missing value

3.2 Imputation Results with the KNN Method

In Table 2 shows the MSE value of imputation results with the KNN method using the Euclidean distance at k = 2. Experiments were carried out on five variables with different percentages of missing. From this table, it can be seen that for each k random given the resulting MSE value tends to increase. The Altitude variable at the value of k = 2 results in a small MSE. Meanwhile, the best MSE heading variable is obtained at k = 2 with MSE 772.1429469. From the results of the table above, there is an increase in the MSE value at k random is quite high, at k random. An increase in ed at all the k also seemed to given. In each lue, the higher



itude	*Longitude*
07486	6.36099E-05
.3283	6.35987E-05
2224	8.69229E-05
02803	0.000143168
.26593	0.000145379
.48593	0.000185453

Table 3 is the best k search result with the GA method. In the Altitude variable, 10 iterations were carried out and the value of k = 3 was obtained, with an MSE value of 128668.96. At Variable Speed, 10 iterations were also carried out so that the value of k = 9 was obtained with MSE 457.52. In the Heading variable, 70 iterations were carried out resulting in k = 61 with the best MSE value of 752.1429. Meanwhile, for the Lattitude 10 and Longitude variables, 50 iterations were carried out to produce an value of k = 3 and MSE at Lattitude = 9.16E-05 and MSE Longitude 1.68E-05.

Table 3. Table of best k search results with the GA method

Variabel	*Altitude*	*Speed*	*Heading*	*Lattitude*	*Longitude*
iterasi	10	10	70	10	50
k - best	***k:3***	***k:6***	***k:61***	***k:3***	***k:2***
MSE	128668.95	457.52	752.1429	9.16E-05	1.68E-05

3.4 Results of imputation using the KKN-GA method

After completing the experiment with the KNN algorithm and the results have been obtained, then the experiment is continued with optimization with the Genetic Algorithm (GA). In this experiment, the results were quite good when filling in data that experienced missing value (mv). The experiment was carried out on the Sriwijaya Air aircraft type by taking data from 10-20 seconds.

The comparison of the MSE values generated by the two methods using the Euclidein distance measure can be seen in table 4. The MSE value generated by the KNN-GA method in each missing group is entirely better than the MSE value generated by the KNN method. With the KNN algorithm the best MSE results are small. It is the same with the KNN-GA where the best MSE value is obtained with a small one. From the comparison of parameter combinations that give the best results for both methods, it can be seen that there is a fluctuation pattern of the same MSE value increase and the best MSE value is produced by a small MSE value.

For a clearer comparison of the performance of the KNN and the KNN-GA, it can be seen in Figure 7. Comparison of the MSE produced by the KNN method (blue line) with the optimization result (red line) is presented in the figure. For the same missing k values, the GA optimization results all succeeded in giving better values than the experiment with the KNN method.

Table 4 Comparison Table of KNN and KNN-GA

Variable	*K- best*	*KNN*	*k -best*	*KNN-GA*
		MSE		*MSE*
Altitude	2	561219.7664	3	128668,96
Speed	5	392.6019214	6	457.5201.
Heading	2	772.1429469	61	752.1429.
Lattitude	10	0.000126593	3	9.16E-05
Longitude	12	1.68E-05	2	1.68E-05

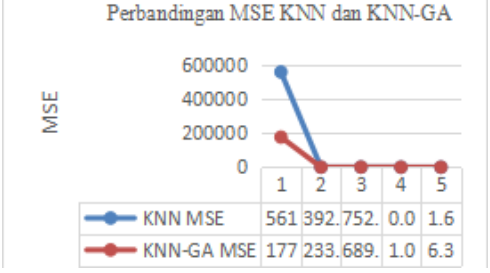


Figure 7. Comparison Graph of MSE KNN and KNN GA.

5.4. Results of Aircraft Data Imputation at Abdurrahman Saleh Airport

From the results of experiments on aircraft data at Abdurahman Saleh Airport, the KNN-GA, GA method has a good performance. This is shown by its ability to minimize MSE. If the imputation value for each observation is considered, the imputation result for each variable in the missing observation is relatively different from the actual value. There is a big difference and some are close to the actual value.

Table 5 presents some of the results of the data imputation of medium aircraft at 1 hour of data collection. It can be seen from the table that the imputation value in some experiments is different from the actual value in the complete data (the part with a question mark "?"). For example, the variable Lattitude in the second experiment. The resulting imputation value is -7.91262 while the actual value of the variable is -7.8157. However, there are imputation results that are relatively close to the actual data value such as the value of the Speed variable in the 10341th experiment. The resulting imputation value was 455, not much different from the actual value of 454. The full imputation results can be seen in the Appendix. Table 5 Results of the Abdurrahman Saleh Airport Aircraft Data Imputation.

Table 5. Comparison Table of actual data and Imputasion

	Aktual Data					Imputation Results				
No	*Altitude*	*Speed*	*Heading*	*Lattitude*	*Longitude*	Altitude	Speed	Heading	Lattitude	Longitude
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
20	34975	?	?	-7.80826	112.96259	34975	458	126	-7.80826	112.96259
21	?	?	?	?	?	35000	458	126	-7.91262	112.96327
22	35000	?	?	?	?	35000	458	126	-7.91262	112.96327
23	?	?	?	?	?	35000	458	126	-7.91262	112.96327
24	?	459	126	?	?	35000	459	126	-7.91262	112.96327
25	34975			-7.8157	112.97274	34975	459	126	-7.91262	112.96327
26	?	459	126	?	?	35000	459	126	-7.91262	112.96327
27	?	?	?	?	?	34975	459	126	-7.81519	112.97203
28	35000	?	?	?	?	35000	459	126	-7.8157	112.97274
29	?	458	126	?	?	34975	459	126	-7.8157	112.97274
30	35000	?	?	-7.81947	112.97791	35000	459	126	-7.81668	112.97274
31	35000	?	?	-7.82073	112.97961	35000	459	126	-7.81668	112.97274
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
10335	31750			-7.94261	113.0975	31750	455	126	-7.94261	113.0975
10336		455	124			31725	455	124	-7.9438	113.09922
10337	31750			-7.94307	113.0982	31750	455	126	-7.94307	113.0982
10338	31725			-7.9438	113.09922	31725	455	126	-7.9438	113.09922
10339		455	124			31700	455	124	-7.9447	113.09922
10340		454	125			31700	454	125	-7.9447	113.09922
10341	31725					31725	455	344	-7.9447	113.09922
10342	31725			-7.94493	113.10084	31725	455	344	-7.94493	113.10084
10343		454	125			31700	454	125	-7.94717	113.10378
10344		454	125			31700	454	125	-7.94717	113.10378

4. CONCLUSION

A. Conclusion

From the research that has been done, several conclusions can be drawn.

1. The KNN method in this study is used to fill in data that has missing value. As with the properties of K-NN, this method will increase in accuracy if the training data or old case patterns that are owned are increasingly varied. In addition, this method has strong consistency, by looking for cases by calculating the closeness between new cases and old cases based on their k values. In accordance with the experimental results of k values randomly at k = 2, 5, 2, 10, and 2. For the better In the finding process for k on the KNN, optimization was carried out using the GA method.
2. Optimization using the GA method is superior, this is indicated by the smaller MSE value compared to the KNN. GA is better from the two methods.

3. Through GA Optimization, get variable information from the method through MSE so that the imputation of the missing value data for aircraft is better.

B Advice

Several things that can be developed for further research in the same scope include:

1. Assessing the application of the KNN-GA to different data structures through data collection and other types of data
2. Do a combination of other methods from the data cases that experience missing value.
3. Use a distance measure other than Euclidean Distance.
4. Applying optimization methods other than the GA method and making comparisons with other imputation methods

6. REFERENCES

[1] ICAO. Guidance Material on Issues to be Considered in ATC Multi-Sensor Fusion Processing Including the Integration of ADS-B Data. 2008.

[2] Eko Marpanji, Kadarisman, Gamantyo H, Aplikasi platform komputasi software defined radio (SDR) untuk digital spektrum analizer .ITS. ISSN 2337-3539.2010.

[3] Akshay N et, Shruthi R, Sushmitha K N, Vanitha R, Dr. Rekha K R .*"Live detection with Mode-S Trasponder Using RTL-SDR"*.2017. Karnataka, India

[4] Little, Roderick, J. A & Rubin, Donal B. "Statistical Analysis With Missing Data". California. 1987.

[5] C. F. Suharno, M. A. Fauzi, and R. S. Perdana,"Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors (K-NN) dan Chi-Square," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput. Vol 1 No 10 2017*, Jul. 2017.

[6] Aprilio Adrianus Yoza, Ginardi Hari & Purwananto Yudhi. . Manajemen Kualitas Aliran Data A *utomatic Dependent Surveillance-Broadcast* (ADS-B) Banyak Titik Dengan Pohon Keputusan, Jurnal Teknik Pomits, Vol. 4, No. 2. 2014

[7] Hiroshi De Silva and A. Shehan Perera, "Missing Data Imputation using Evolutionary K-Nearest Neighbor Algorithm for Gene Expression Data", Proceedings of 6 th International Conference on Advances in ICT for Emerging Regions, 2017.

[8] Pang Nin Tan, Michael Steinbach, & Vipin Kumar. . *"Introduction to Data Mining 5th Ed. Boston: Pearson Education Inc.*" 2006

[9] Steffi Pauli & fazat Nur Azizah. "Imputation of missing value using dynamic Bayesian network for multivariate time series data. IEEE. 2017

.