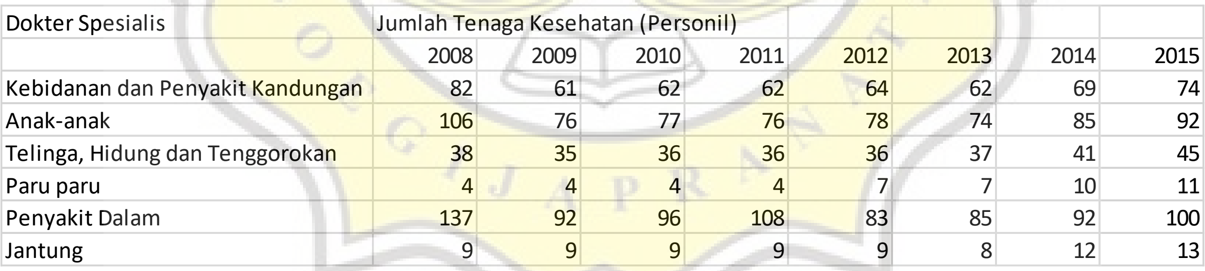
CHAPTER 4
ANALYSIS AND DESIGN

4.1. Data Acquisition and Preprocessing

Dataset required for this research is acquired from Central Bureau of Statistics (BPS). The dataset is in the form of time series data containing the number of physicians in Semarang city from 2008 to 2015. There are 20 types of physicians included in this dataset, which is further detailed in the next section. There are difficulties acquiring local data to support this study. Most data presented in BPS are either incomplete or missing completely. This dataset is the best I could muster to satisfy the data requirement.

There are 20 types of physicians included in the dataset, as follows (1) Midwifery, (2) Pediatric, (3) Otolaryngologist, (4) Pulmonologist, (5) Internist, (6) Cardiologist, (7) Surgeon, (8) Anesthesic, (9) Skin and Genital Specialist, (10) Dentist, (11) Psychiatrist, (12) Neurologist, (13) Radiology, (14) Andrology, (15) Ophthalmologist, (16) Clinic Pathology / Anatomy, (17) Medic Rehabilitation / PRU, (18) Obstetric and Gynecology, (19) Nutritionist, (20) Forensic. Each contains the number of physicians in their respective field of expertise.

Table 1: Data Sample Table



Dokter Spesialis	Jumlah Tenaga Kesehatan (Personil)							
	2008	2009	2010	2011	2012	2013	2014	2015
Kebidanan dan Penyakit Kandungan	82	61	62	62	64	62	69	74
Anak-anak	106	76	77	76	78	74	85	92
Telinga, Hidung dan Tenggorokan	38	35	36	36	36	37	41	45
Paru paru	4	4	4	4	7	7	10	11
Penyakit Dalam	137	92	96	108	83	85	92	100
Jantung	9	9	9	9	9	8	12	13

The original dataset is already complete and does not require further cleaning to be used effectively. The author transposed the dataset using Microsoft Excel, since having year data as columns is not ideal to be used in panda dataframe due to columns being the searchable index and not the rows. By transposing the dataset, the column represents field of expertise, and the rows represents years. I also converted the original data format from characters to integers, to make them computable.

Table 2: Preprocessed Data Sample Table

Tahun	Kebidanan dan Penyakit Kandungan	Anak-anak	Telinga, Hidung dan Tenggorokan	Paru paru
2008	82	106	38	4
2009	61	76	35	4
2010	62	77	36	4
2011	62	76	36	4
2012	64	78	36	7
2013	62	74	37	7
2014	69	85	41	10
2015	74	92	45	11

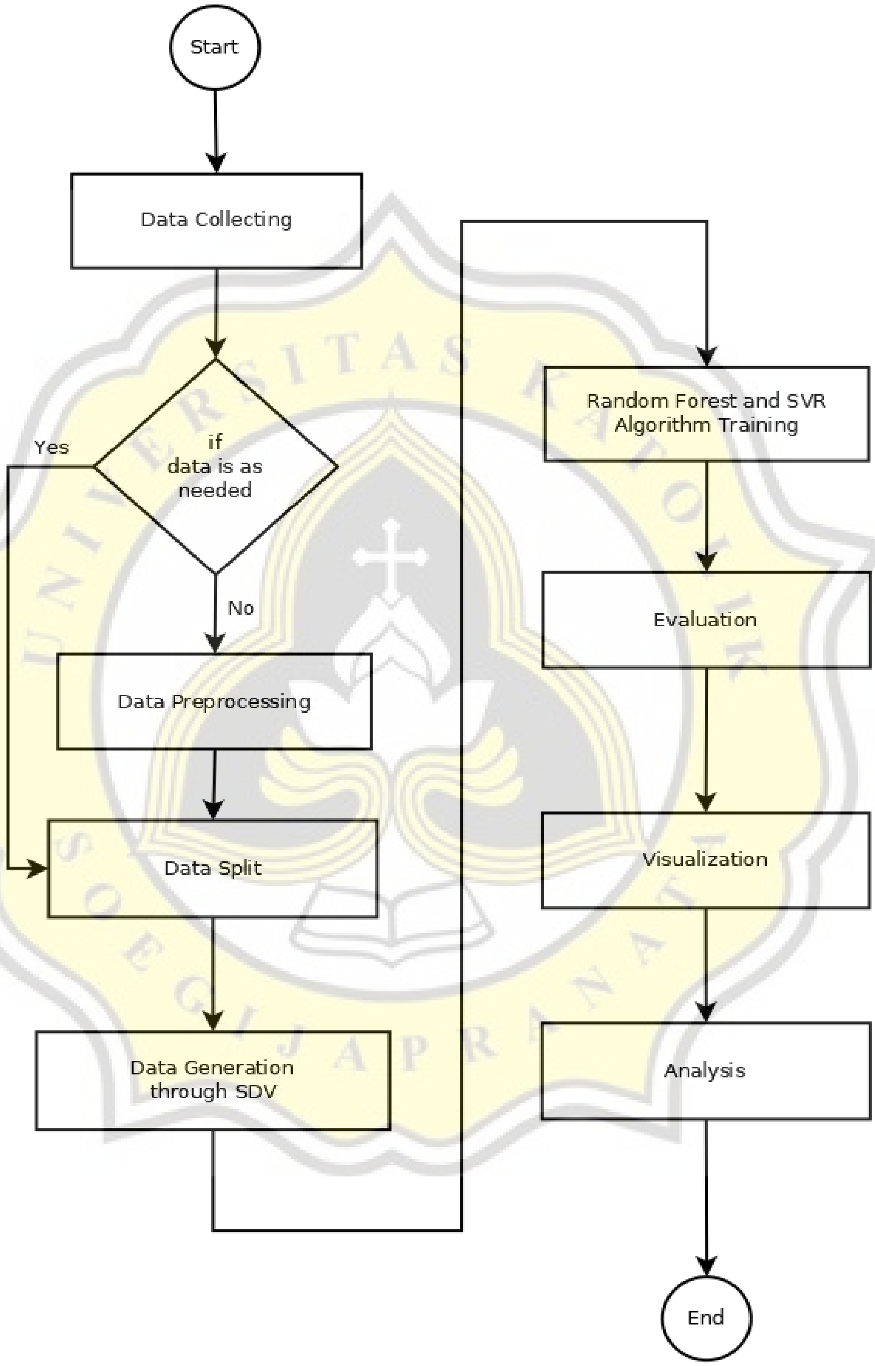
4.2. Workflow

Testing subjects in this research are SDV and prediction algorithms. As mentioned before, SDV is used to generate data based on existing data. The next subject is prediction algorithms. Support Vector Regression and Random Forest Regression are both commonly used algorithms to solve regression problems.

The original dataset acts as the control parameter in this research. The original data is split into training and test data. SDV generates data based on this training data. Test data are reserved for evaluation. SDV generates data in sequences. Each sequence generates data equal to the number of training data fitted into the algorithm. Five sequences, for example, generates the amount of training data multiplied by five. To find the optimal number of data to generate, the author tested the SDV accuracy using five different sequence numbers: 1, 5, 10, 15, 20, 25.

According to Dobbin and Simon [13], the optimal number of training data for MSE is between 60% to 80%, however, for smaller datasets the number can go closer to 100%. The author picked 80% split proportion to maximize training data availability for the original dataset. This decision is taken to maximize the performance of the original dataset, as its data amount is already small.

Decisive declarators of whether SDV can be used as substitute or not can be explored by comparing error value between SDV and the original data. Both SDV and original training data are fitted into Support Vector Regression and Random Forest. The prediction results are evaluated using MAPE and MSE. These error value across different generated data amount are visualized in the form of accuracy tables and figures.

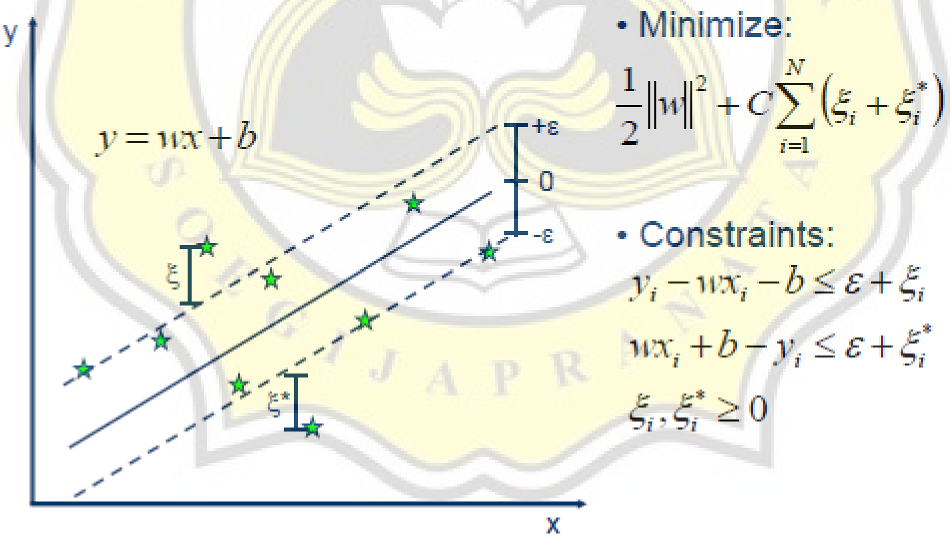


Gambar 1: Research Workflow

There are four things to visualize: SDV generated data stats, SDV generated data accuracy, original data accuracy, and error values graphs of RF and SVR. SDV generated dataset stats are described in a plot visualizing their number of data, deviation standard, and mean across different sequences. SDV datasets and original dataset accuracy are visualized using tables. Comparison between RF and SVR are shown in a graph describing their error value across different numbers of sequences or data.

4.3. Methods and Formula

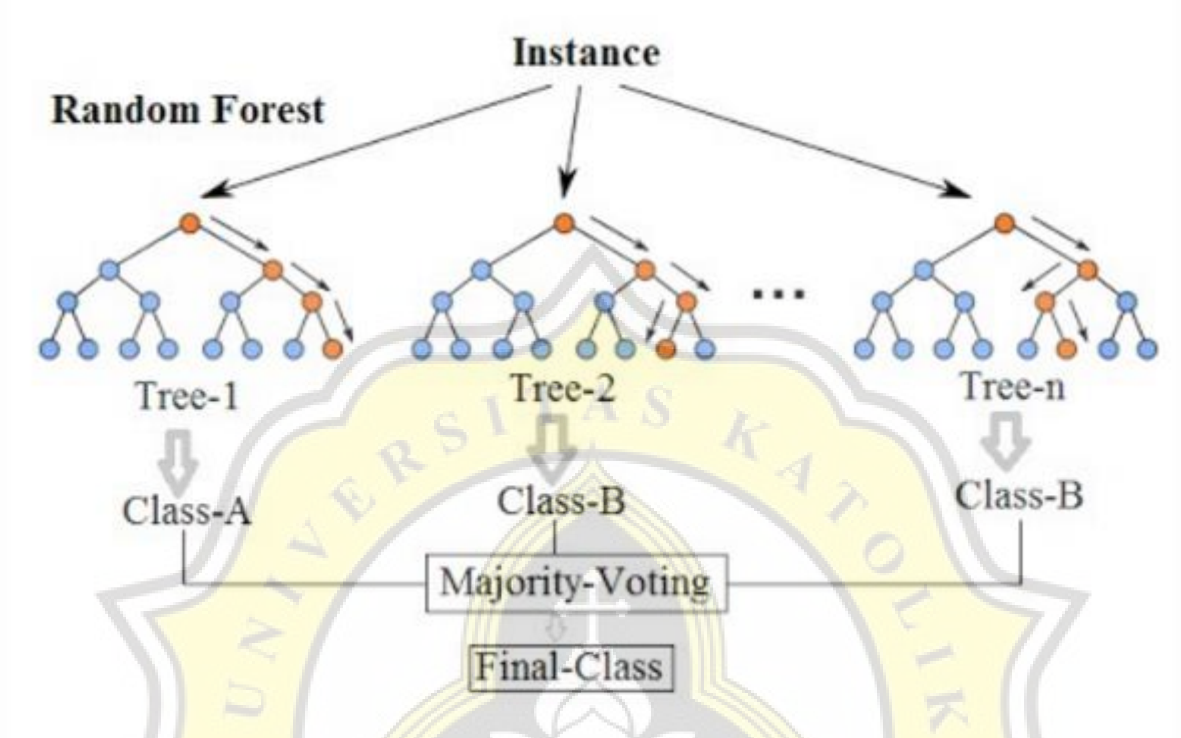
Synthetic data generation technique is implemented in python by utilizing SDV library. Because the dataset is in the form of time series, PAR or Probabilistic Auto Regressive function is used in this context. PAR allows learning multi-type, multivariate time series data to generate new data with the same format and identities as the one it's based on. PAR models is trained using real data, and the number of generated data depends on the number of iterations or sequence parameters in python.



y
$y = wx + b$
+ε
0
-ε
ξ
ξ*
x
• Minimize:
$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$
• Constraints:
$y_i - wx_i - b \leq \varepsilon + \xi_i$
$wx_i + b - y_i \leq \varepsilon + \xi_i^*$
$\xi_i, \xi_i^* \geq 0$

Gambar 2: Support Vector Regression [14]

Support Vector Regression uses hyperplane to compute prediction data based on trained model. This hyperplane is in the form of straight line separator in 2 dimensional time series model. Every elements are calculated in a positional relative to the hyperplane. This way, deviation/variance comes into play to predict following elements.



Gambar 3: Random Forest [15]

Random Forest Regression uses decision trees as predictor. Using sampling and ensemble aggregation method (often called bagging), Random Forest creates multiple decision trees to solve classification problems. The decisive answer is selected by vote. This system effectively hides individual tree mistake by hiding it among the other trees.

Random Forest Regression is especially resistant to overfitting problem that SDV might possess. This is a clear advantage over Support Vector Regression to be used with data generation techniques. This is also the reason author decided to select the algorithm in particular over numerous others.

$$MSE = \sum \frac{(Y' - Y)^2}{n}$$

Gambar 4: MSE [16]

$$\mathrm{MAPE} = \sum_{t=1}^{n} \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right| \times 100\%$$

Gambar 5: MAPE [17]

Evaluation in this research utilizes both MSE and MAPE. Mean Square Error is calculated by seeking the gap value between actual data and predicted data and powering the result by two. The total powered gap are compiled before extracting mean from the compiled data. This scoring system is scale dependent, meaning it should not be used to compare accuracy between two dataset with different scale. However, since my research will compare said data in mixed manner, mse is usable.

MAPE score is similar to MSE but the gaps are not powered by 2. Instead, they are divided by actual data, and multiplied by 100. This method produces a percentage value that is easy to perceive. This scoring system also scale independent, meaning scale is irrelevant when comparing data with this method.