CHAPTER 3
RESEARCH METHODOLOGY

3.1. Literature Review

Many studies have been carried out and of course some are related to the problems to be solved, due to the limited knowledge, therefore it is very necessary to gain knowledge from research that has been done before to get a new intuition on how to solve problems in this case predict bus routes. At this stage reading and analyzing research related to what will be researched is carried out and the results of reading will be written in the form of a literature review.

3.2. Data Anlaysis and Preprocessing

This is the very first step when making deep learning models, because this step is highly crucial and will impact the deep learning model performance and result. The data from this stage will be fed into the model to make predictions, so it is important that the data has been processed correctly so that the model can learn and make better predictions. To perform data analysis and processing, there are several stages to be carried out, namely: data selection, data visualization and variable analysis and data preprocessing.

3.2.1. Data Selection and Variable Analysis

At this step, the variables will be analyzed in order to select the variables that relevant for the task and will affect the prediction of bus routes and drop data variable that is not relevant for this task, as already mentioned in the background, data may contain valuable information that can help forecast. Therefore, before continuing, the data needs to be analyzed first to extract hidden information.

A. Data Selection

Data selection is a step to selecting and dropping the used and unused data given from the source, this must be done because not all data variable is related to the problem that want to be solved, so the variable must be dropped, The data variable given from the company is as shown below :

1. scheduleDate : The passenger's departure date time is shown in this variable.
2. Start: The origin of the passenger's departure is shown by this variable.
3. Destination : The passenger's destination is shown in this data variable.
4. Qty : total of passeger per transaction
5. Name : The name of the bus route
6. className : the class of the bus(ECONOMY, AC, *etc*)

from all data variable above the used variable only scheduleDate, start, Destinations and Qty.

A) Data Variable analysis

Indeed, the data variables selected is quite small but with a deeper analysis, some variables have hidden information contained in it which is able to help deep learning training from the data so that predictions from the model can be obtained better.

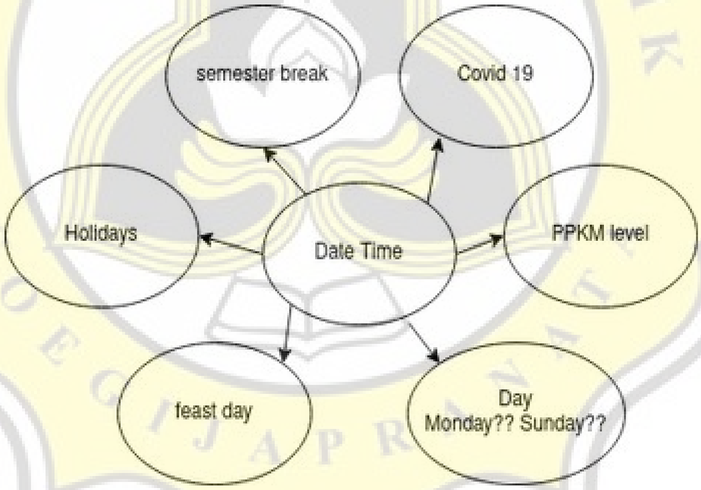

Figure 3.1: Date Time variable analysis

From this figure, it is explained that date time has a lot of information that can be used as new variables and these new variables can also affect bus demand. For example we know that bus demand will rise when holidays are coming.

3.2.2. Data Visualization

Data Heatmap provided will be analyzed more and data visualization will be created in form of heatmap and visualization of data distributions to achieve clear visualization. Data visualization is done to get clear image so further analysis and action on the data can be carried out.

B) Heatmap Visualization

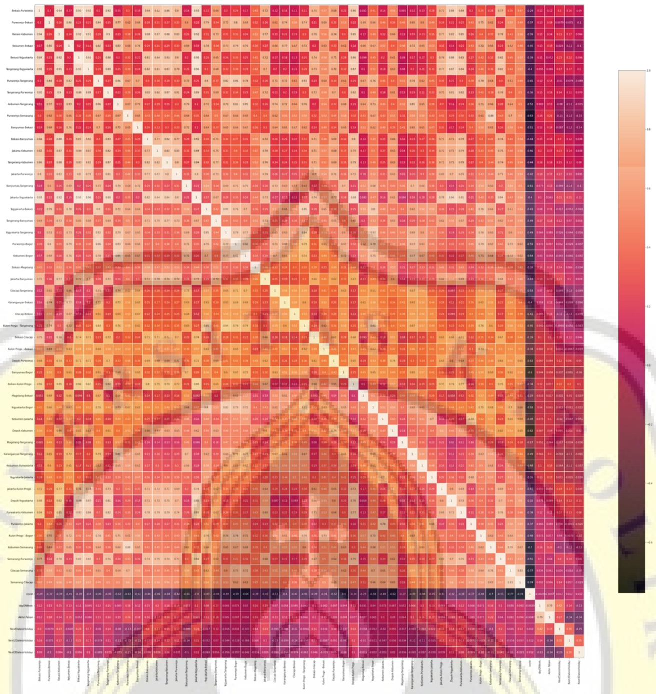

Figure 3.2: Heatmap

With help of seaborn library to create heatmap to achieve clear visualization about how each variable correlates with each other and how strong the correlation is, so it can be ensured that the data can be predicted. Heatmap is a visualization method that shows how each data variable correlate with each other. Data variables can have correlation with other data either positively correlate or negatively correlate with each other or even the data will not have correlation at all. Heatmap visualization is important in time series task since the time series task will have many variables and each variable must have correlation with one another, if not then the time series data could not be predicted further.

C) *Histogram Visualization*

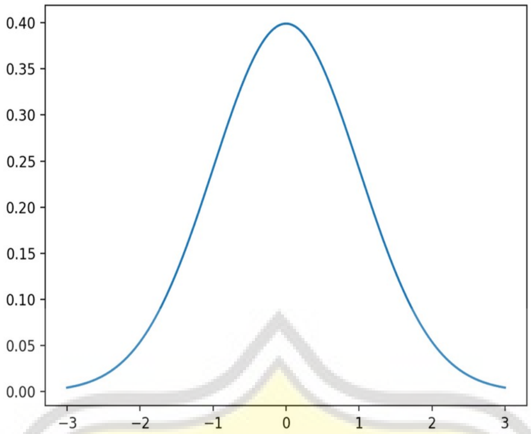

Figure 3.3: Gaussian distribution

In addition to the heatmap, a histogram illustration will also be carried out. Histogram here is used to show how the data distribution is, so if the data distribution is not normal then further step could be done besides that from histogram get a clear visualization of the Gaussian distribution of the data in order to get a clear picture of what actions should be taken to make the data more have more gaussian like distribution.

3.2.3. Split Data

The data that was previously analyzed will be entered and split in this stage. The data will usualy divided into two parts, the dependent variable and the independent variable. Dependent variable is a data variable that depend on the independent variable, and indepenent variable is data that dosent depand with other data. with the dependent variable being the bus route that will be predicted and the independent variable is the categorical data. In order to classify the data per record in time-series, and then the data will be divided into two parts, 80% training set and 20% test set.

3.2.4. Feature Scaling

The data that have been splited then must be scaled first. Its unkown for what feature scaling is the best for machine learining and deep learning models. But feature scaling is important to performed for the data so deep learning or machine learning models could perform better job. There are many method to do feature scaling, normalizaation, standarization, power transformer an so on. Normalization is a method to normalize the data and scaled into bettween 0 an 1 value. Figure below is showing how the normalization is calculated.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Figure 3.4: Normalization

Standarization is a method to transform the data with its standard deviation. Figure below is the formula of standarization method, it seen that the x value is substracted with the mean times x then divided by the standard deviation. After standarization performed the data mean should be near 1.

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

Figure 3.5: standarization formula

3.3. it must be noted that feature scaling is must be done after the data is splited because if feature scaling is done before then the splited data could not be inverted back into its normal form.

3.4. Creating Models

The models that are frequently used are Convolution Neural Networks, Long Short Term Memory and GRU models, each of them have their own uniqueness [1]. This research will use LSTM-Autoencoders-Bi-LSTM and Bi-LSTM models to forecast the bus route demand then Autoencoder-Bi-LSTM and Bi-LSTM Models performance will be compared to find autoencoders architecture effect on time series forecasting problem. 60% of data will become the training set for the deep learning to learn, and the models will validate the gained weight with the help of validation sets then the last to calculate the performance will be explained in next part.

3.4.1. LSTM Autoencoder-Bi-LSTM Hybrid Models

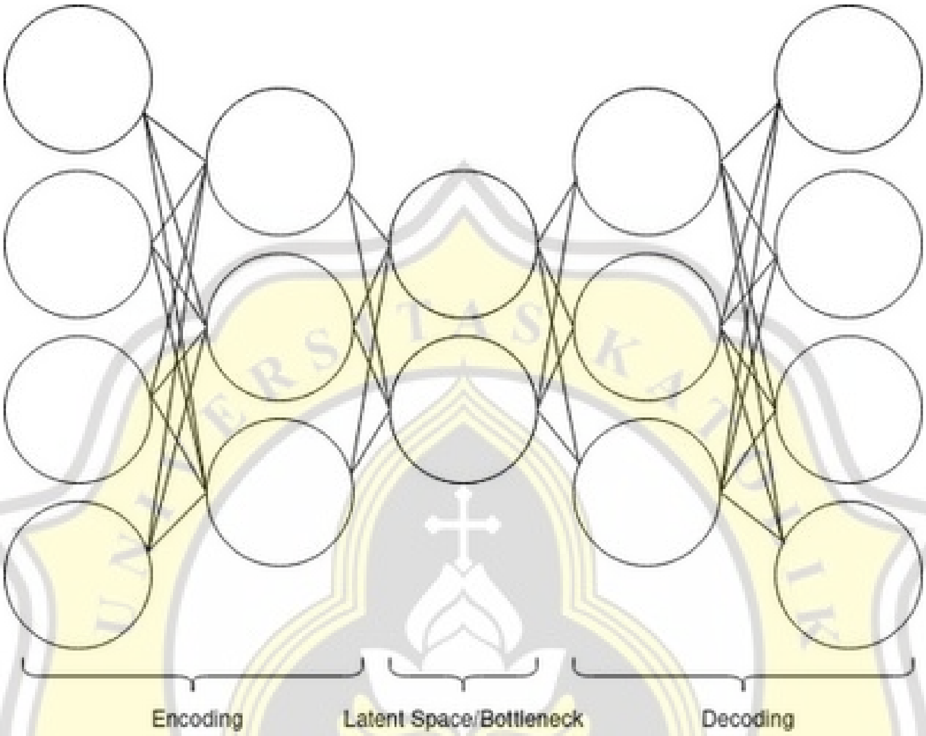

Figure 3.6: Autoencoder architecture

LSTM or Long Short Term Memory is an improvisation of RNN that can overcome the vanishing gradient problem. The vanishing gradient problem can be overcome by the LSTM because LSTM have an input gate, forget gate and output gate that LSTM layer have. The ability to extract feature of unlabeled data is one of the reasons why autoencoder was used to solve the unsupervised learning problem. The autoencoders will have 2 stages the first stage is the encoding stage, in encoding step the large data variable sequence will be compressed into bottleneck/latent space then after encoding the data then will be decompressed again and try to decode the latent space variable back to its normal dimension when trying to reconstruct the data variable back to its original form to hopefully models got new information that will help the models to learn more from the data. The output obtained from the autoencoders will be entered back into the Bi-LSTM model. Bi-LSTM or

Bidirectional Long Short Term Memory is one type of artificial neural network layer that based on LSTM but learns the data Bidirectionally (forward and backward) at the same time. Bi-LSTM in here is used to calculate and provide results to dense layer for the models output in the form of a regression which is a prediction of the number of bus requests for a given route each sequence.

3.4.2. Bi-LSTM Model

As mentioned before Bi-LSTM is an artificial neural network layer that based on LSTM but learns the data Bidirectionally. This model will use Bi-LSTM models to create a comparison to find the effect of autoencoders in time series problem. For this models the data will be fed directly into the models and Bi-LSTM will directly calculate and provide the predictions per sequences.

3.5. Models Evaluation and Comparison

The performance of the system will be evaluated in this part. Since the forecasting problem would have high error rate the metrics used will be RMSE because RMSE will root the given error first so RMSE will give more weight to larger error and for the training loss function is MSE. RMSE is calculating the root-mean-square-error of the predicted and the real value. RMSE metrics are calculated like below image.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Figure 3.7: RMSE and MSE formula

2 models will be made namly the LSTM Autoencoder-Bi-LSTM Hybrid Models and Bi-LSTM models, so that comparisons can be made by looking at the RMSE results obtained. After training finished the models loss and metrics history during training epoch

will be saved for each model's then compared to find the effect of autoencoders in time series forecasting task.