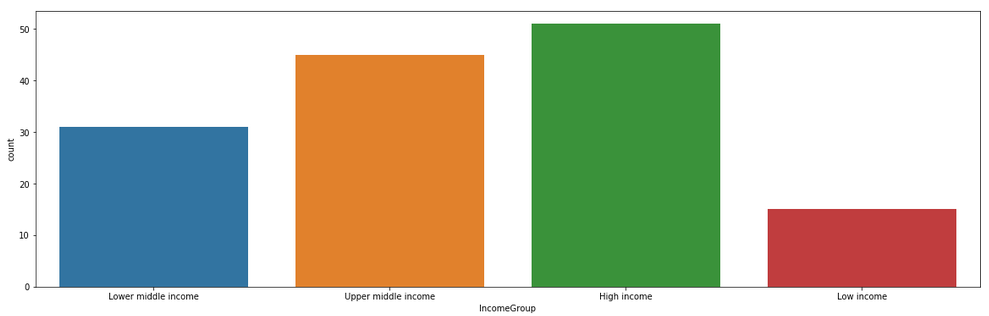
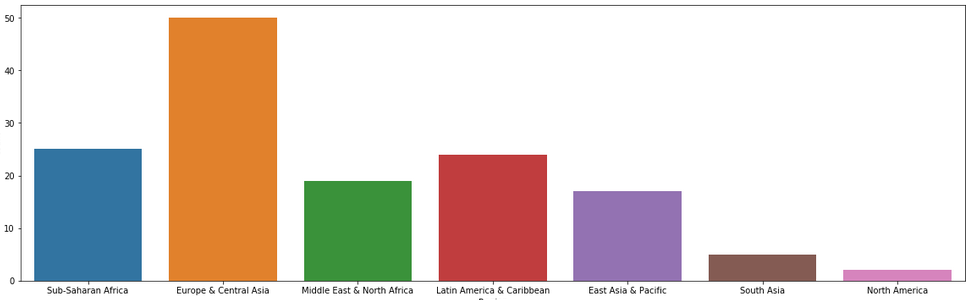
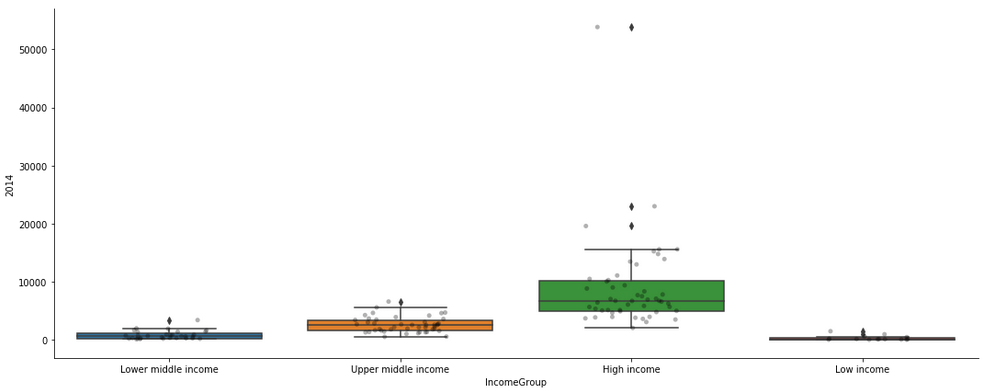
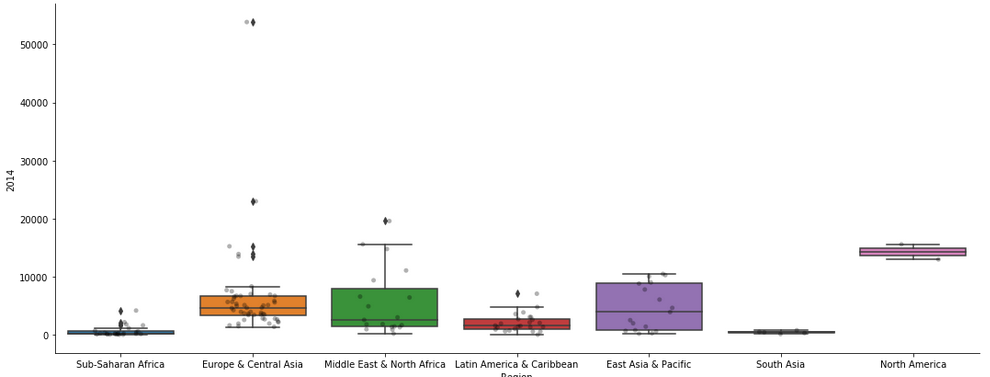
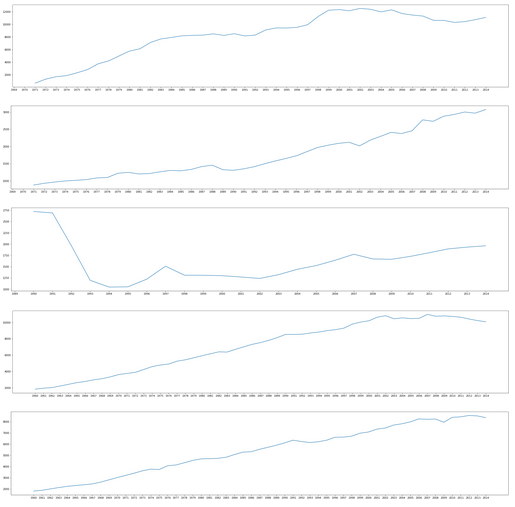
# Analisis

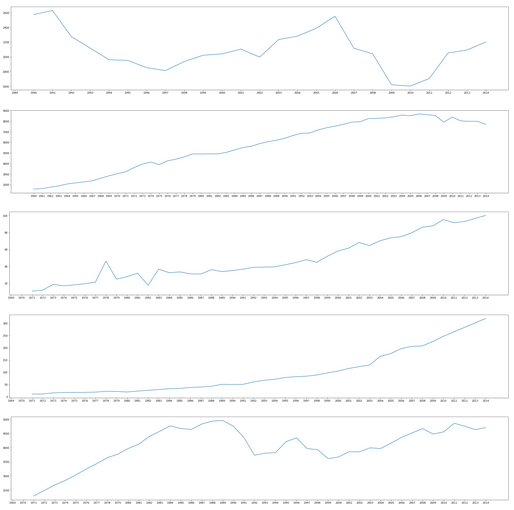
## Eksplorasi Data

Di dalam dataset ada 4 jenis fitur, yaitu Country Code, Region, Income Group, dan penggunaan listrik per tahun. Tetapi karena sudah dipastikan bahwa Country Code tidak akan digunakan pada analisis, maka kolom itu kami tidak masukkan ke analisis. Mari kita eksplorasi fitur yang lain.

Persebaran dari data kategorikal terlihat merata. Lalu kita lihat hubungan masing-masing data kategorikal dengan data numerik.



Untuk fitur IncomeGroup, cukup terlihat pengaruhnya terhadap penggunaan listrik masing-masing negara. Untuk fitur Region, pengaruh terhadap penggunaan listrik masing-masing negara tidak terlalu terlihat, maka kami tidak akan mengikitsertakan fitur Region. Selanjutnya mari kita lihat grafik untuk data penggunaan listrik berbagai negara per tahun.



Plot tersebut menjelaskan penggunaan listrik dari 10 Negara pertama. Semakin tahun, penggunaan listrik cenderung semakin tinggi.

## Praproses Data

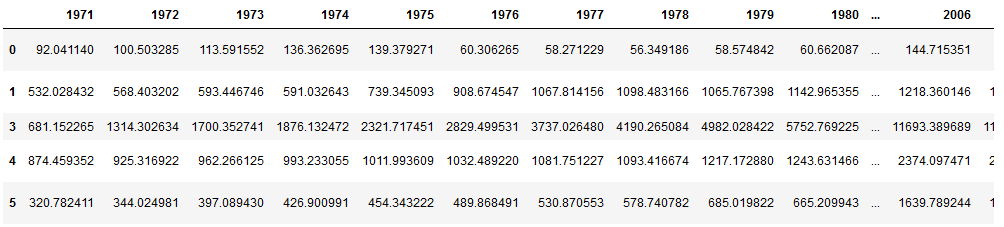
Sebelum melakukan permodelan, kami melihat dulu apakah value di masing-masing fitur kosong atau tidak.

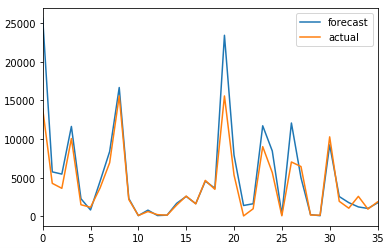
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Region | 42 | 1973 | 35 | 1988 | 33 | 2003 | 2 |
| IncomeGroup | 42 | 1974 | 35 | 1989 | 33 | 2004 | 2 |
| 1960 | 152 | 1975 | 35 | 1990 | 10 | 2005 | 1 |
| 1961 | 152 | 1976 | 35 | 1991 | 9 | 2006 | 1 |
| 1962 | 152 | 1977 | 35 | 1992 | 9 | 2007 | 1 |
| 1963 | 152 | 1978 | 35 | 1993 | 9 | 2008 | 1 |
| 1964 | 152 | 1979 | 35 | 1994 | 9 | 2009 | 1 |
| 1965 | 151 | 1980 | 35 | 1995 | 7 | 2010 | 1 |
| 1966 | 151 | 1981 | 34 | 1996 | 7 | 2011 | 1 |
| 1967 | 151 | 1982 | 34 | 1997 | 7 | 2012 | 1 |
| 1968 | 151 | 1983 | 34 | 1998 | 7 | 2013 | 1 |
| 1969 | 151 | 1984 | 34 | 1999 | 7 | 2014 | 1 |
| 1970 | 151 | 1985 | 33 | 2000 | 2 |  |  |
| 1971 | 35 | 1986 | 33 | 2001 | 2 |  |  |
| 1972 | 35 | 1987 | 33 | 2002 | 2 |  |  |

Terlihat untuk periode 1960 sampai 1970 lebih dari 90% value kosong. Karena berbahaya untuk analisis, maka tidak diikutsertakan dalam analisis. Untuk data numerik yang lain, kami melakukan penanganan menggunakan metode *Simple Imputer* dengan parameter *mean.* Untuk fitur kategorikal, dilakukan drop baris untuk baris yang tidak ada value-nya karena sulit untuk di *handle.*

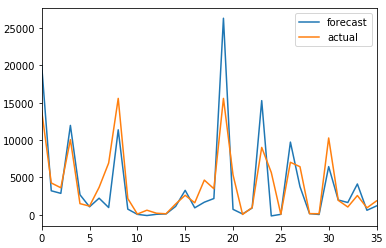
Lalu untuk bisa mengikutsertakan fitur kategorikal, tentu saja kita harus merubah isi dari ketegori ke format yang dimengerti oleh model. Maka kami menggunakan algoritma One-Hot Encoder untuk merubah fitur kategorikal agar dimengerti oleh model.

## Modelling dengan metode Long-Short Term Neural Network

Setelah melakukan eksplorasi dan praproses data, dilakukan pengubahan permasalahan *univariate time series* menjadi permasalahan *supervised learning* dengan membagi data menjadi input (X) dan output (y). Data yang menjadi output adalah data penggunaan listrik di tahun 2014 dan yang menjadi input adalah data penggunaan listrik tahun 1971-2013 serta data IncomeGroup.

Setelah itu data dibagi menjadi data ​train dan data ​validation dengan perbandingan 75:25. Model yang digunakan adalah *Long Short Term Memory (LSTM) Neural Network.* Dimana terdapat beberapa variasi LSTM diantaranya *Vanilla LSTM, Stacked LSTM,* dan *Bidirectional LSTM* yang akan di test satu persatu untuk mencari yang terbaik*.* ​*Vanilla LSTM* ​yang digunakan memiliki parameter basis yang terdiri dari *layer* LSTM dengan jumlah ​*neuron* sebanyak 50 dan ​*layer dense* sebagai layer penghubung. Fungi ​*loss* yang akan di minimalisasikan oleh model adalah ​*mean squared error*​. Model dioptimalisasikan menggunakan ​*Adam Optimizer.* ​Pada proses ​*training,* ​model dilatih menggunakan data ​*train* yang sudah ditransformasikan dengan jumlah ​*epoch* sebanyak 100 kali. Lalu dilakukan proses testing pada model yang sudah dilatih, dan mendapatkan ​*RMSE* sebesar 2728,26.

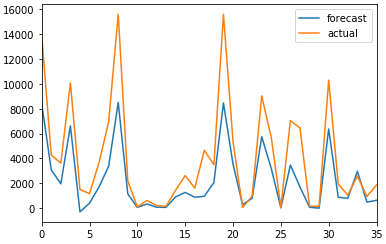
|  |  |  |
| --- | --- | --- |
|  | actual | forecast |
| count | 36 | 36 |
| mean | 4085.024939 | 5236.027344 |
| std | 4420.549958 | 6313.709961 |
| min | 51.195008 | 80.305069 |
| 25% | 965.006324 | 1166.806213 |
| 50% | 2383.460472 | 2435.357422 |
| 75% | 5865.441455 | 8048.509644 |
| max | 15590.59991 | 25653.1582 |

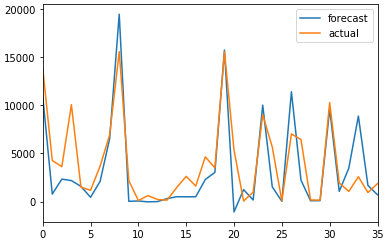
Selanjutnya dilakukan testing dengan variasi *Stacked LSTM.* Dimana hampir sama dengan *Vanilla LSTM,* hanya saja dilakukan dua kali pemasukan algoritma ke model. *Stacked LSTM* memiliki hasil RMSE sebesar 3139,80.

|  |  |  |
| --- | --- | --- |
|  | actual | forecast |
| count | 36 | 36 |
| mean | 4085.024939 | 3878.784 |
| std | 4420.549958 | 6068.86 |
| min | 51.195008 | -143.687 |
| 25% | 965.006324 | 498.7148 |
| 50% | 2383.460472 | 1421.389 |
| 75% | 5865.441455 | 3405.02 |
| max | 15590.59991 | 26319.91 |

Selanjutnya dilakukan testing dengan variasi *Bidirectional LSTM.* Dimana hampir sama dengan *Vanilla LSTM,* hanya saja dilakukan pemasukan algoritma *Bidirectional* ke model. *Stacked LSTM* memiliki hasil RMSE sebesar 2696.59.

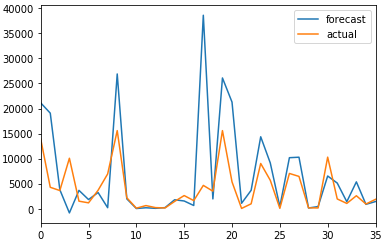
|  |  |  |
| --- | --- | --- |
|  | actual | forecast |
| count | 36 | 36 |
| mean | 4085.024939 | 2247.457 |
| std | 4420.549958 | 2599.647 |
| min | 51.195008 | -301.107 |
| 25% | 965.006324 | 388.5134 |
| 50% | 2383.460472 | 1053.304 |
| 75% | 5865.441455 | 3238.677 |
| max | 15590.59991 | 8502.434 |



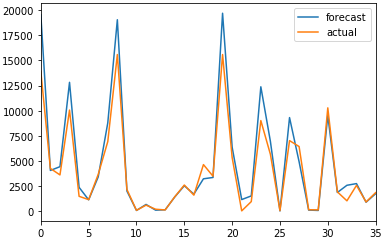
Karena *Bidirectional LSTM* menunjukkan RMSE paling kecil, maka variasi ini dipilih. Selanjutnya dilakukan tuning parameter di algoritma dengan mengatur nilai dari *epoch* agar prediksi semakin akurat. Untuk testing tuning parameter yang pertama dilakukan dengan nilai *epoch* sebesar 200. Dan didapat RMSE sebesar 2729,13.

|  |  |  |
| --- | --- | --- |
|  | actual | forecast |
| count | 36 | 36 |
| mean | 4085.024939 | 2247.457 |
| std | 4420.549958 | 4928.026 |
| min | 51.195008 | -1077.6 |
| 25% | 965.006324 | 268.0146 |
| 50% | 2383.460472 | 1384.35 |
| 75% | 5865.441455 | 3140.878 |
| max | 15590.59991 | 19482.03 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | actual | forecast | |
| count | 36 | | 36 |
| mean | 4085.024939 | | 6770.211 |
| std | 4420.549958 | | 9535.901 |
| min | 51.195008 | | -844.556 |
| 25% | 965.006324 | | 583.6573 |
| 50% | 2383.460472 | | 1959.441 |
| 75% | 5865.441455 | | 9392.81 |
| max | 15590.59991 | | 38626.73 |

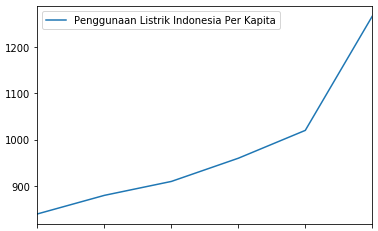
Karena hasil *forecast* dirasa belum normal karena nilai minimum bernilai negative maka dilakukan testing dilakukan dengan tuning parameter *epoch* sebesar 400. Dan didapatkan RMSE sebesar 7814.63.

Karena hasil *forecast* dirasa masih belum normal karena nilai minimum bernilai negative maka dilakukan testing dilakukan dengan tuning parameter *epoch* sebesar 600. Dan didapatkan RMSE sebesar 1678,22.



|  |  |  |
| --- | --- | --- |
|  | actual | forecast |
| count | 36 | 36 |
| mean | 4085.024939 | 4812.001 |
| std | 4420.549958 | 5619.286 |
| min | 51.195008 | 47.39497 |
| 25% | 965.006324 | 1168.449 |
| 50% | 2383.460472 | 2571.867 |
| 75% | 5865.441455 | 6525.833 |
| max | 15590.59991 | 19695.58 |

Karena model dirasa sudah memberikan hasil yang baik dan logis. Dipakai model hasil tuning parameter dengan nilai *epoch* 600.

Lalu selanjutnya dilakukan prediksi untuk penggunaan listrik Indonesia pada tahun 2019. Data Indonesia diambil dari BPS dan mendapat data penggunaan listrik per kapita untuk tahun 2013 sampai 2018. Untuk data yang lain diambil dari rata-rata penggunaan listrik dengan IncomeGroup “Lower middle income” dan “Upper middle income” 40 tahun sebelumnya karena Indonesia termasuk negara dengan kategori “Middle Income”. Lalu didapat penggunaan listrik per kapita untuk tahun 2019 sebesar 1265,919 kWh per kapita.

2014

2015

2016

2017

2018

2019

Hasil ini menunjukkan bahwa penggunaan listrik perkapita di Indonesia pada tahun 2019 akan naik.

## Simpulan

* Model *LSTM* dengan variasi *Bidirectional LSTM* dipilih menjadi model akhir karena memiliki performa yang lebih baik dari variasi yang lain
* Tahun 2019 akan ada kenaikan penggunaan listrik per kapita sebesar 24% ke angka 1265,919 kWh berdasarkan prediksi dari model akhir.

## Saran

* Dataset yang mengandung pemasukan wilayah, populasi wilayah, GDP, pertumbuhan populasi, pertumbuhan pendapatan, dan kecepatan pembangunan dirasa akan memberi hasil yang lebih baik untuk prediksi.
* Jika ada data asli dari Indonesia untuk 45 tahun terakhir akan memberikan prediksi yang lebih akurat dan sesuai dengan kondisi di Indonesia.