**Stock Price Report**

**Overall**

Saham dapat diartikan sebagai tanda penyertaan modal seseorang atau pihak (badan usaha) pada suatu perusahaan atau perseroan [1]. Indonesia sendiri memiliki pasar sahamnya sendiri yang dinamakan Bursa Efek Jakarta yang digabungkan dengan Bursa Efek Surabaya. Pada projek ini, mencoba memprediksi harga pasar saham menggunakan model machine learning serta mengetahui proses pengambilan data dari database. Dataset yang akan digunakan berasal dari saham apple selama 10 tahun yang didapatkan dari website Kaggle [2]. Software yang digunakan antara lain Pentaho, PostgreSQL (Pg Admin 4), JupyterNotebook, Microsoft Excel dengan model machine learning yang akan digunakan antara lain Convolutional Neural Network.

**Business Understanding**

Pada pernyataan yang telah dijelaskan, sehingga masalah yang diangkat adalah

1. Bagaimana proses pengambilan data dari database?
2. Bagaimana performa model machine learning yang digunakan?

Tujuan dari masalah yang diangkat adalah

1. Mengetahui proses pengambilan data dari database
2. Mengetahui performa model machine learning yang digunakan

Solusi yang dapat dilakukan

1. Menggunakan korelasi antar fitur pada dataset menggunakan library sklearn dan heatmap seaborn
2. Mengevaluasi hasil model menggunakan MSE dan RMSE

**Data Understanding**

Dataset yang digunakan memiliki 2518 baris dan 6 fitur, antara lain.

* Date: Tanggal harga saham dibuka
* Close: Harga terakhir/penutup saham pada hari tersebut
* Volume: Jumlah saham yang diperdagangkan pada hari tersebut.
* Open: Pembukaan harga saham pada hari tersebut
* High: Harga tertinggi sebuah saham pada hari tersebut
* Low: Harga terendah sebuah saham pada hari tersebut.

Dari 6 fitur, fitur High akan digunakan sebagai **label** dikarenakan untuk mengetahui harga tertinggi sebuah saham pada hari tersebut. Pada tahapan ini akan dilakukan proses ETL dari database PostgreSQL dan Pentaho. Banyangkan dataset yang dimiliki berada dalam sebuah database perusahaan dan ingin dilakukan pengambilan data untuk dilakukan proses lebih lanjut, disitulah ETL digunakan. ETL adalah proses pengumpulan data dari berbagai macam sumber dengan tujuan untuk memperoleh data yang berkualitas [3]. Berikut proses ETL yang dilakukan.

1. **Pembuatan Database**

CREATE DATABASE "Stock\_market"

WITH

OWNER = postgres

ENCODING = 'UTF8'

CONNECTION LIMIT = -1;

1. **Pembuatan Tabel dan Fitur**

CREATE TABLE IF NOT EXISTS public.stock\_hist

(

date date,

close character(10),

volume integer,

open character(10),

high character(10),

low character(10)

);

ALTER TABLE public.stock\_hist

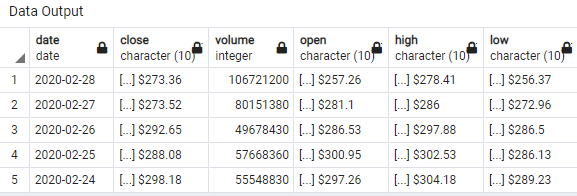
OWNER to postgres;

1. **Import Dataset dan Hasil Import**



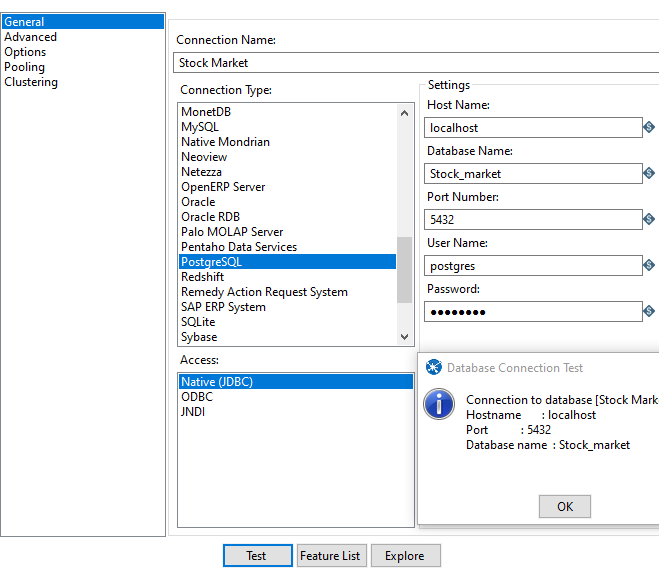
**Gambar 1 Import data ke PostgreSQL**

SELECT \* FROM stock\_hist LIMIT 5



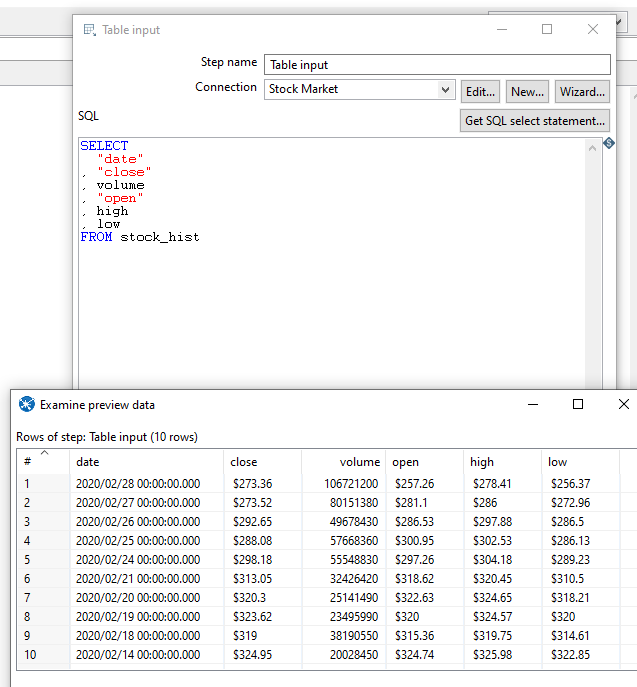
1. **Pengambilan Data dari Database Menggunakan Pentaho**

Pertama yang perlu dilakukan adalah melakukan connection ke PostgreSQL. Untuk melakukan connection ke PostgreSQL dari Pentaho diperlukan sebuah file bernama **postgresql-42.5.0.jar**. File tersebut tergantung dari versi Java yang diinstal dan pada projek ini versi Java yang terinstall adalah JDK 8. File yang telah di download akan diletakkan pada folder lib pada Pentaho. Setelah diletakkan, maka akan dilakukan connection ke PostgreSQL.



**Gambar 2 Connection Pentaho ke PostgreSQL**

Setelah melakukan koneksi, tahapan selanjutnya adalah pengambilan data. Pengambilan data dilakukan pada fitur design dengan menggunakan fungsi **Table Input**. Pada Table Input, pilih koneksi database yang ingin diperoleh datanya dan klik **Get SQL select statement.**



**Gambar 3. Tabel Input**

Diagram

Description automatically generated

**Gambar 4. Rangkaian ETL**

Rangkaian diatas merupakan rangkaian ETL sederhana yang hanya sekedar pengambilan data dari sebuah database. Hal ini dikarenakan proses processing data akan dilakukan pada bahasa program Python. Namun akan dijelaskan singkat Table Output dan Microsoft Excel Output.

Table Output berfungsi sebagai pembentukan tabel baru pada database terkait. Dengan mengatur parameter seperti pada gambar dibawah ini.

Graphical user interface, text, application

Description automatically generated

**Gambar 5. Parameter Table Output**

Step name berguna untuk menamai step pada pentaho, connection sebagai database terkait, Target schema sebagai schema table dimana schema table yang digunakan pada projek ini adalah public, dan Target table sebagai nama table yang akan dibuat pada database Stock Market.

Graphical user interface, text

Description automatically generated

**Gambar 6 Hasil Table Output**

Microsoft Excel Output digunakan sebagai file output dalam format xls. Dengan mensetting lokasi hasil file dan kolom yang digunakan.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, application, table

Description automatically generated

**Gambar 7 Parameter Microsoft Excel Output**



**Gambar 8 Hasil Microsoft Excel Output**

**Data Preparation**

Tahapan yang dilakukan pada data preparation antara lain Exploratory Data Analisis dan Preprocessing

Text

Description automatically generated  
**Gambar 9 Tampilan 5 dataset pertama**

**Exploratory Data Analisis (EDA)**

Dataset tidak memiliki missing value pada tiap fitur

Table

Description automatically generated

**Gambar 10 Missing value**

Dataset memiliki 1 tipe datetime, 4 tipe object, dan 1 tipe integer. Jika dilihat 4 tipe object seharusnya float, untuk merubah tipe object ke tipe float akan dilakukan penghapusan symbol ‘$’ pada fitur-fitur tersebut.

Text

Description automatically generated

**Gambar 11 Tipe fitur**

Table

Description automatically generated

**Gambar 12 Hasil removes symbol**

Text

Description automatically generated

**Gambar 13 Tipe data baru pada fitur**

Tipe data yang telah dirubah akan dilakukan pengecekan outlier pada dataset

Chart

Description automatically generated

Dapat dilihat bahwa tiap-tiap fitur memiliki outlier masing dan jika dihitung didapatkan sebagai berikut

Table

Description automatically generated with medium confidence

Namun pada projek ini, tidak akan dilukan pembersihan outlier dan dilakukan split dataset menjadi train dan test



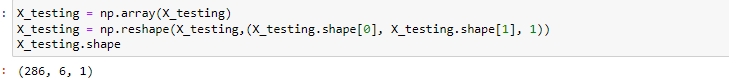
Setelah dilakukan pemisahan, diperlukan normalisasi. Hal ini dilakukan karena untuk menyamakan distribusi pada tiap data train dan testing dengan testing hanya dilakukan transform tanpa fit.

Chart

Description automatically generated with medium confidence

Text

Description automatically generated



Graphical user interface, text, application

Description automatically generated

Jika dilakukan normalisasi terlebih dahulu, setelah itu split dataset akan menyebabkan persamaan semua distribusi dan hal tersebut dapat menyebabkan bias pada distribusi data. Berikut visualisasi split dataset

Chart, scatter chart

Description automatically generated

Setelah data prepation selesai dilakukan, tahapan selanjutnya adalah modeling.

**Modeling & Evaluation**

Model yang digunakan pada projek ini adalah LSTM dengan optimasi Adam, loss MAE dan metrics RMSE

Table

Description automatically generated



Menggunakan epochs = 5, didapatkan loss sebagai berikut

Chart, line chart

Description automatically generated

Berikut hasil predict dari sisi Train dan sisi Testing.

Chart, scatter chart

Description automatically generated



Jika dilihat pada Train, prediksi yang didapatkan mendekati dengan nilai aslinya. Namun pada data testing menjauhi nilai aslinya, untuk mengetahui penyebabnya mari melihat nilai RMSE data testing



Ternyata nilai RMSE yang didapatkan cukup tingggi, hal ini dapat ditingkatkan dengan menambah layer pada model dan mengubah nilai parameter optimizer dan jumlah epochs.

[1] “Saham .:: SIKAPI ::.” https://sikapiuangmu.ojk.go.id/FrontEnd/CMS/Category/64 (accessed Aug. 25, 2022).

[2] Nasdaq, “Apple (AAPL) Historical Stock Data | Kaggle.” https://www.kaggle.com/datasets/tarunpaparaju/apple-aapl-historical-stock-data (accessed Aug. 25, 2022).

[3] R. Dharayani, K. A. Laksitowening, and A. P. Yanuarfiani, “Implementasi ETL ( Extract , Transform , Load ) Pangkalan Data Perguruan Tinggi dengan Menggunakan State-Space Problem,” *e-Proceeding Eng.*, vol. Vol.2, No., no. 2355–9365, pp. 1159–1165, 2015.