JALA Tech Data Scientist Take Home Test

Faisal Nur Khairudin

1. Hasil evaluasi kelengkapan data:
   a. Tidak terdapat *feature description* untuk dataset **feed_tray** dan **mortalities**
   b. Terdapat *missing values* yang lebih dari 10% pada beberapa kolom di beberapa dataset
   c. Terdapat beberapa kolom tanggal yang masih memiliki tipe data "object"
   d. Terdapat *negative value* pada beberapa kolom yang tidak masuk akal
   e. Terdapat duplikat data berdasarkan semua kolom terutama pada *feeds dataset* yang memiliki jumlah paling banyak
2. Hasil perhitungan SR dan ADG udang di setiap siklus budidaya (10 baris pertama):
   a. Survival Rate (%)

	cycle_id	updated_at	size	created_at	weight	id	harvested_at	status	selling_price	jumlah_ekor	pond_id	total_seed	survival_rate(%)
0	15013.0	2022-07-18 08:42:59	399.86	2022-07-18 08:42:59	0.01	11960.0	2022-04-17	Failed	NaN	3.9986	31254	798822	0.000501
1	11848.0	2021-10-25 23:29:53	300.00	2021-10-24 08:38:53	46.00	7520.0	2021-10-22	Full	1.610000e+06	13800.0000	21231	105000	13.142857
2	11848.0	2021-10-25 23:29:35	300.00	2021-10-25 23:29:35	10.00	7546.0	2021-10-21	Partial	3.500000e+05	3000.0000	21231	105000	2.857143
3	4666.0	2020-10-08 02:07:18	110.00	2020-10-08 02:07:18	270.50	3009.0	2020-08-31	Failed	NaN	29755.0000	15345	65550	45.392830
4	4666.0	2020-10-08 02:07:37	190.00	2020-10-08 02:07:37	270.50	3010.0	2020-08-31	Failed	NaN	51395.0000	15345	65550	78.405797
5	18747.0	2022-10-20 01:06:50	72.30	2022-10-20 01:06:50	5322.28	13876.0	2022-10-19	Full	5.322280e+03	384800.8440	36101	492936	78.063043
6	27095.0	2024-02-19 02:10:48	41.50	2023-12-27 12:00:59	7843.00	25322.0	2023-12-12	Full	4.925404e+08	325484.5000	45264	450000	72.329889
7	27095.0	2024-02-19 02:08:11	67.50	2024-02-12 05:06:57	1244.50	26728.0	2023-11-08	Partial	7.653675e+07	84003.7500	45264	450000	18.667500
8	27095.0	2024-02-19 02:08:08	176.00	2024-02-12 05:06:57	5.00	26729.0	2023-11-08	Partial	1.350000e+05	880.0000	45264	450000	0.195556
9	27095.0	2024-02-19 02:08:07	67.50	2024-02-12 05:06:58	2.00	26735.0	2023-11-08	Partial	9.225000e+04	135.0000	45264	450000	0.030000

   b. Average Growth Rate/Average Daily Gain (ADG)

	cycle_id	sampled_at	average_weight	ABWt-ABW0	ht-h0	ADG
0	3458.0	2020-04-10	4.37	4.37	34.0	0.128529
1	3458.0	2020-04-17	5.23	0.86	7.0	0.122857
2	3458.0	2020-04-24	6.69	1.46	7.0	0.208571
3	3458.0	2020-05-01	8.21	1.52	7.0	0.217143
4	3458.0	2020-05-08	9.68	1.47	7.0	0.210000
5	3458.0	2020-05-15	11.59	1.91	7.0	0.272857
6	3458.0	2020-05-22	13.82	2.23	7.0	0.318571
7	3458.0	2020-05-29	15.62	1.80	7.0	0.257143
8	3458.0	2020-06-05	18.15	2.53	7.0	0.361429
9	3458.0	2020-06-12	20.49	2.34	7.0	0.334286

3. Hasil prediksi model:
   a. Survival Rate (in percentage) forecast
      - Hasil metrik evaluasi model training:

	Model	MAE	r2_score
0	RandomForestRegressor	**4.330758**	**0.826690**
2	XGBRegressor	5.076364	0.803925
4	CatBoostRegressor	6.585183	0.706695
3	LGBMRegressor	7.218221	0.649573
1	AdaBoostRegressor	17.688743	-0.297745

- Hasil *cross validation*:

Model	MAE	r2	Adj-r2	std-MAE
CatBoostRegressor	**9.379899**	**0.351475**	0.350068	0.327404
RandomForestRegressor	9.513719	0.284244	0.282692	0.333111
LGBMRegressor	9.524140	0.342994	0.341569	0.344540
XGBRegressor	9.720559	0.265480	0.263887	0.235580
AdaBoostRegressor	16.853372	-0.262010	-0.264747	2.408002

4. Kesimpulan dan penjelasan terkait fitur atau variabel apa yang penting dalam membuat prediksi
5. Rekomendasi mengenai bagaimana budidaya udang perlu dilakukan untuk mencapat hasil yang optimal:
    - Diharapkan para petambak udang mulai mencatat segala data dengan sistem digital, sehingga semua data dapat tercatat. Kemudian, dari data-data yang telah dikumpulkan, sistem bisa melakukan Analisa sesuai dengan kebutuhan para petambak. Harapannya hasil panen udang bisa lebih optimal.
6. Link ke repository (github) ke project:
    - Link Github: https://github.com/fnkhairudin/Assignment_JALA
    - MLflow tracking: https://dagshub.com/fnkhairudin/Assignment_JALA.mlflow