

2024 Probabilistic Model Class

Using Genetic Algorithms to Optimize Penguin Data Prediction



순천향대학교 미래융합기술학과

Senseable AI Lab

석사과정 김병훈

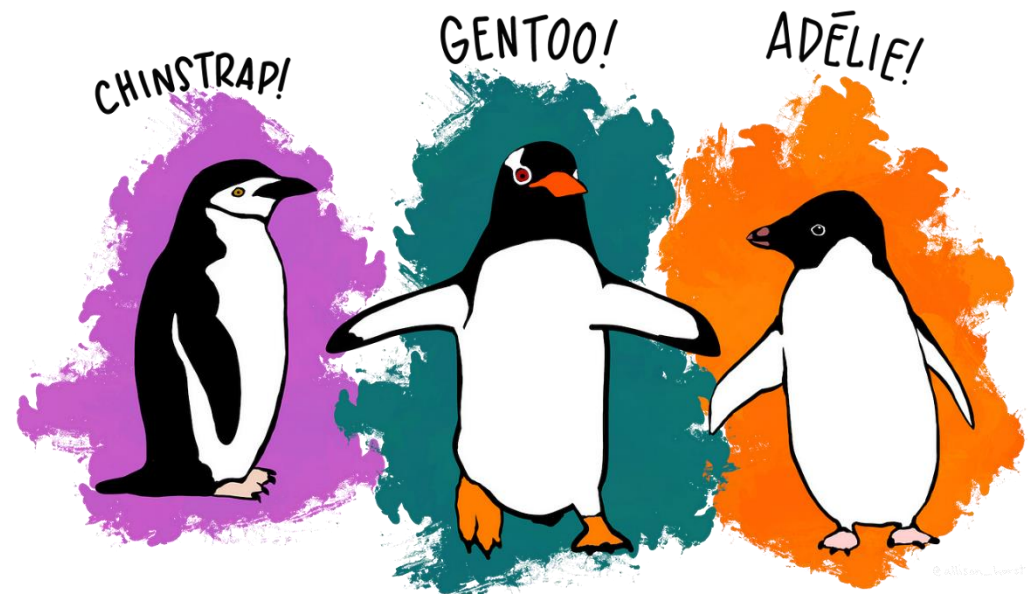
Introduction

Penguin Datasets

통계학 및 데이터 과학 교육용으로 널리 사용되는 데이터셋 중 하나

[변수 목록]

- species: 펭귄의 종(Adelie, Gentoo, Chinstrap)
- island: 펭귄이 발견된 섬(Biscoe, Dream, Torgersen)
- bill_length_mm: 부리의 길이(밀리미터)
- bill_depth_mm: 부리의 깊이(밀리미터)
- flipper_length_mm: 날개의 길이(밀리미터)
- body_mass_g: 몸무게(그램)
- sex: 펭귄의 성별



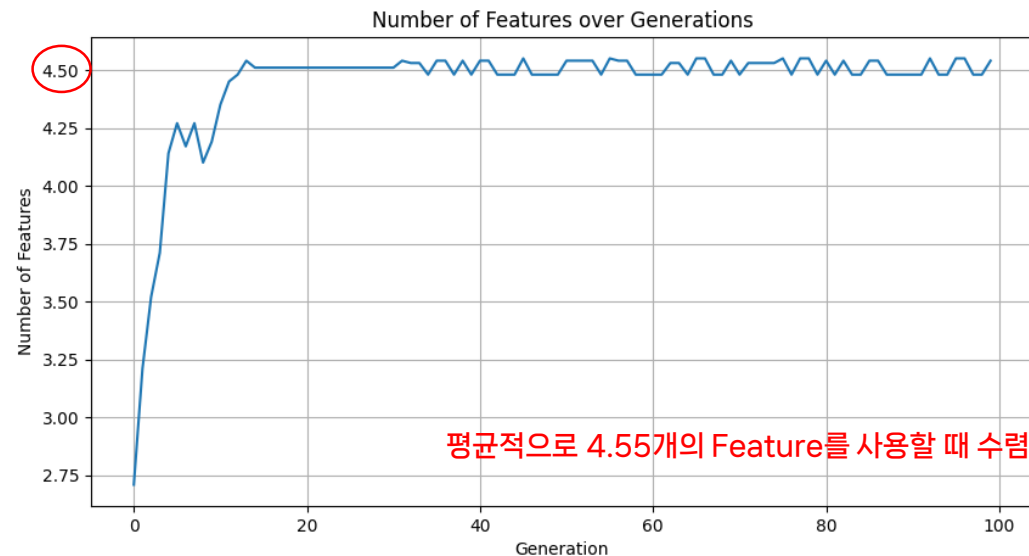
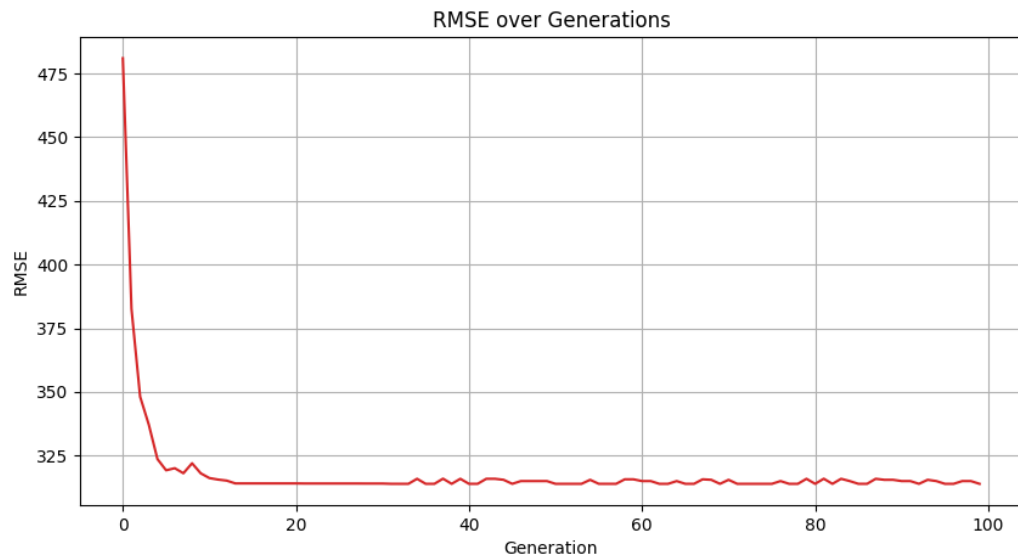
Example 1(몸무게 예측 + 변수 최소 선택)

문제 1. 몸무게 예측 + 변수 최소 선택

- RMSE를 최소화 하면서 독립변수 개수를 최소화 하는 모델 구축
- 총 Feature 개수(One-Hot Encoding 적용했을 때): 11개



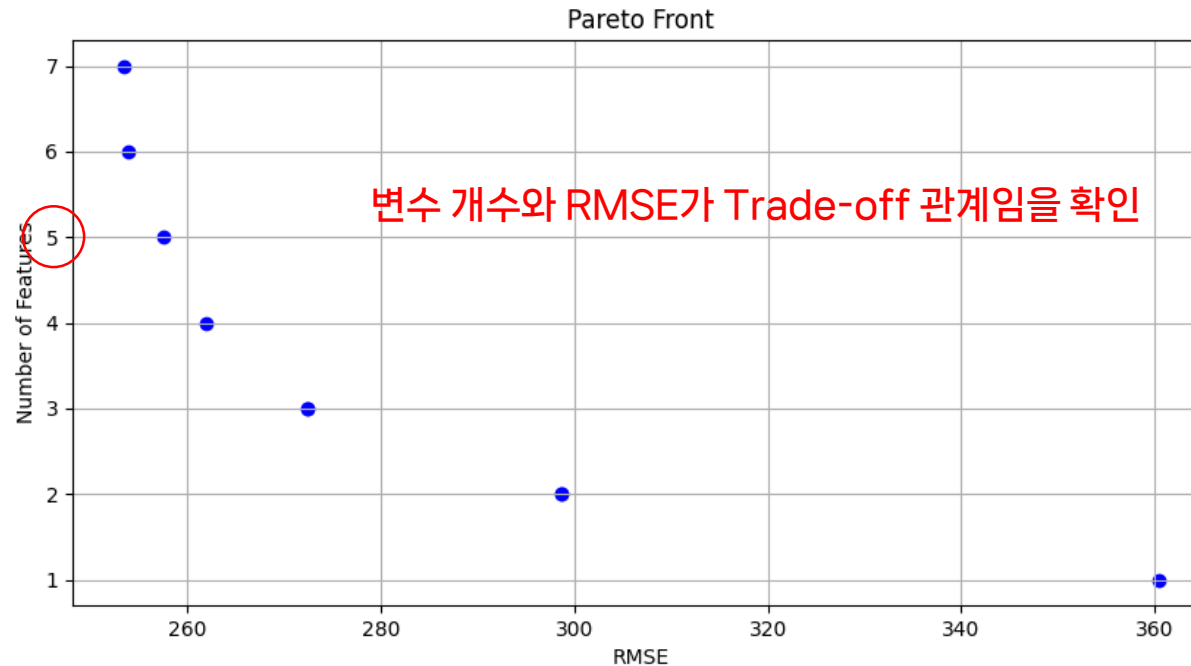
Example 1(몸무게 예측 + 변수 선택)



n_gen	n_eval	n_nds	eps	indicator
1	72	4	-	-
2	172	5	0.2008514994	ideal
3	272	7	0.0615265518	ideal
4	372	6	0.2000000000	nadir
5	472	8	0.0040820828	ideal
6	572	9	0.000000E+00	f
7	672	9	0.000000E+00	f
8	772	9	0.000000E+00	f
9	872	9	0.000000E+00	f
10	972	9	0.000000E+00	f
96	9572	9	5.859510E-17	f
97	9672	9	5.859510E-17	f
98	9772	9	5.859510E-17	f
99	9872	9	5.859510E-17	f
100	9972	9	5.859510E-17	f

Number of Non-Dominated Solutions(파레토 프론트)

Example 1(몸무게 예측 + 변수 선택)



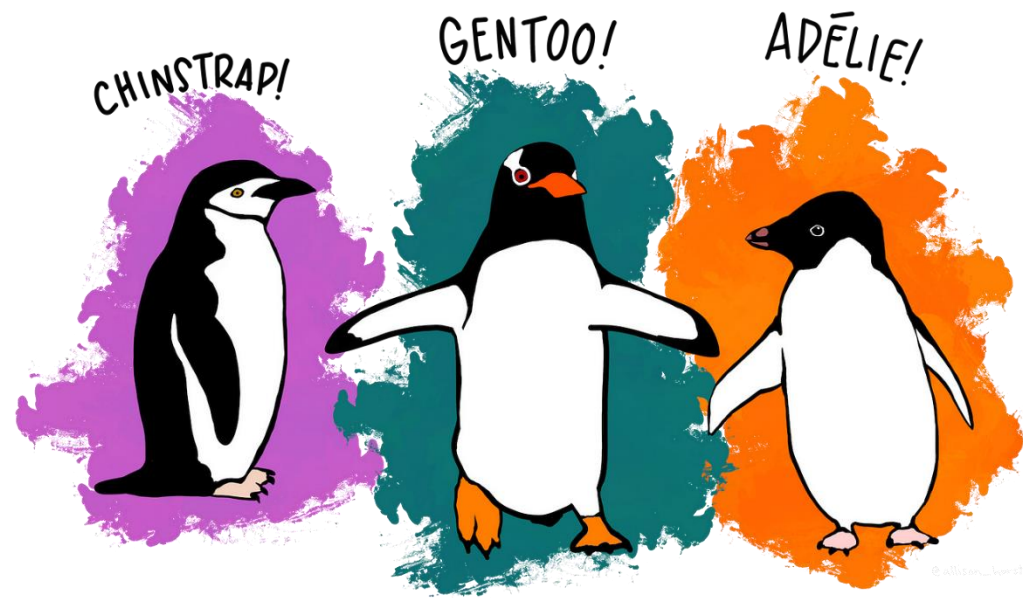
솔루션	선택된 feature	Test RMSE	Feature 개수
1	flipper_length_mm	360.398	1
2,5	flipper_length_mm, species_Gentoo, sex_FEMALE(MALE)	272.407	3
3,8	species_Gentoo, sex_FEMALE(MALE)	298.722	2
4	bill_length_mm, bill_depth_mm, flipper_length_mm, species_Adelie, species_Gentoo, sex_MALE	253.9	6
6	bill_depth_mm, flipper_length_mm, species_Gentoo, sex_FEMALE	261.871	4
7	bill_length_mm, bill_depth_mm, flipper_length_mm, species_Adelie, species_Chinstrap, island_Biscoe, sex_FEMALE	253.464	7
9	bill_length_mm, flipper_length_mm, species_Chinstrap, island_Biscoe, sex_MALE	257.497	5

→ 솔루션 9 선택

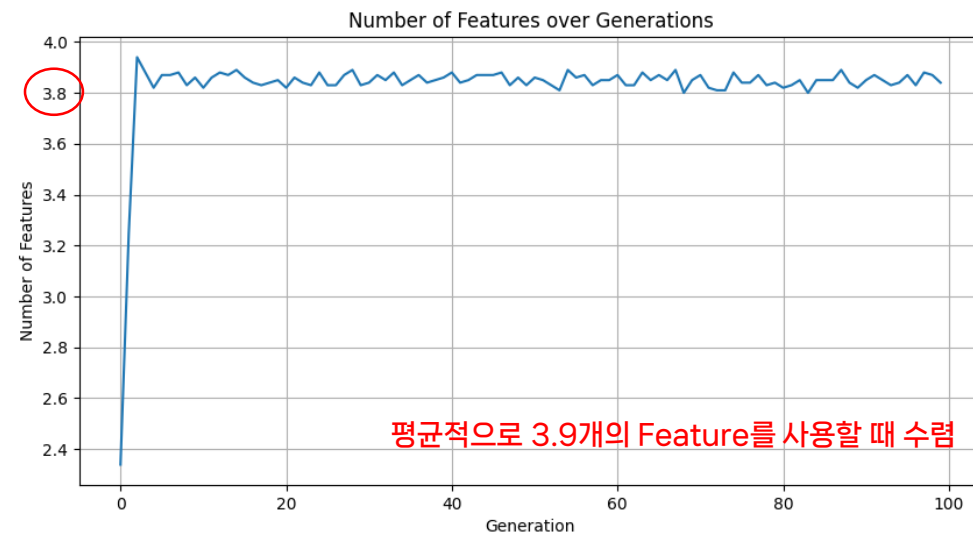
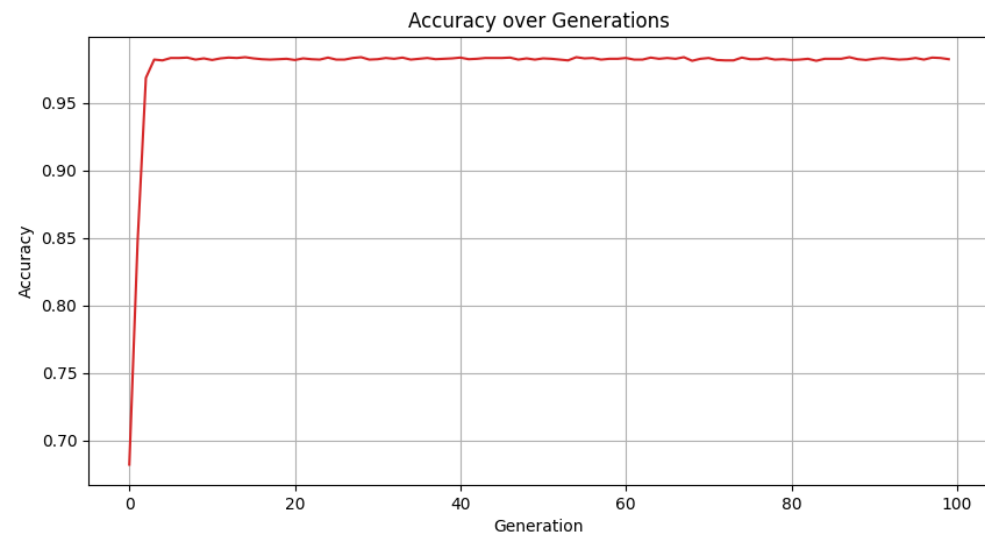
Example 2(펭귄 종 예측 + 변수 최소 선택)

문제 2. 펭귄 종 예측 + 변수 최소 선택

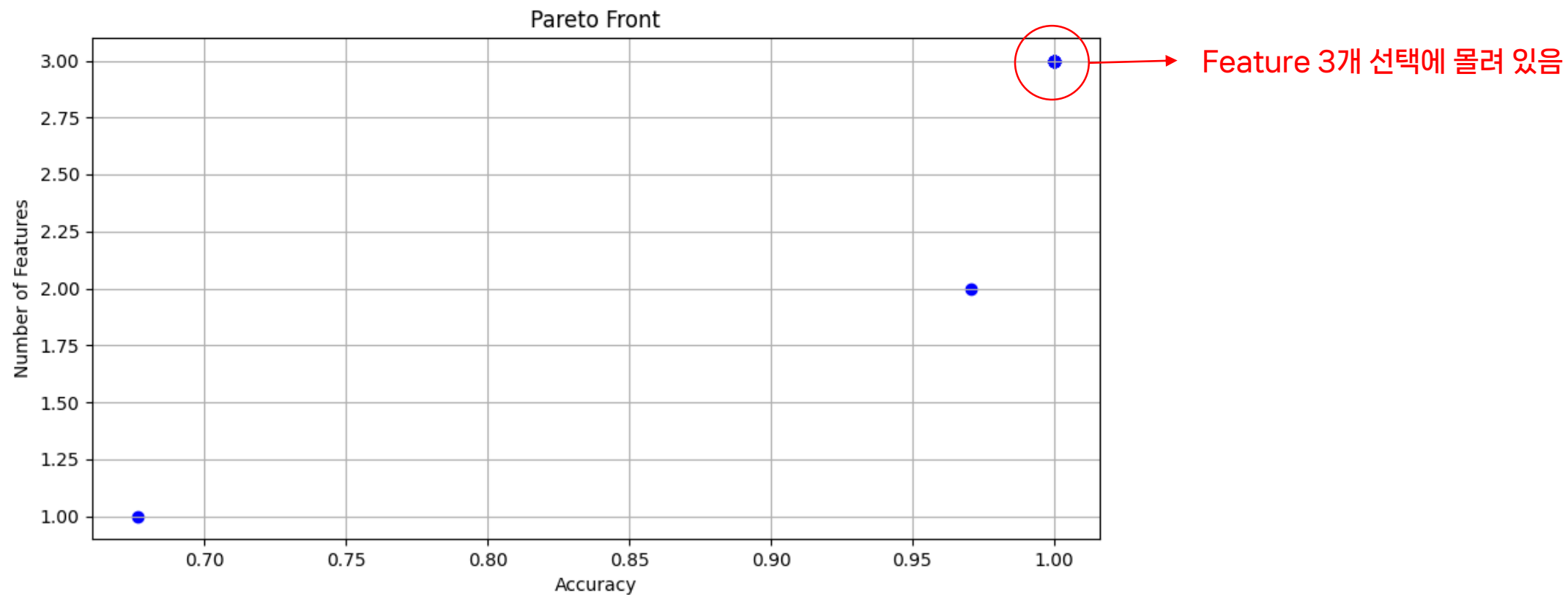
- 정확도를 최대화 하면서 독립변수 개수를 최소화 하는 모델 구축
- 총 Feature 개수(One-Hot Encoding 적용했을 때): 9개



Example 2(펭귄 종 예측 + 변수 선택)



Example 2(펭귄 종 예측 + 변수 선택)



Example 2(펭귄 종 예측 + 변수 선택)

솔루션	선택된 feature	Test Accuracy	Feature 개수
1	bill_length_mm, flipper_length_mm, island_Dream	1	3
2	bill_depth_mm	0.676471	1
3	bill_length_mm, bill_depth_mm, sex_MALE	1	3
4	bill_length_mm, bill_depth_mm, island_Biscoe	1	3
5	bill_length_mm, bill_depth_mm, island_Dream	1	3
6	bill_length_mm, flipper_length_mm, sex_MALE	1	3
7	bill_length_mm, flipper_length_mm	0.970588	2
8	bill_length_mm, island_Biscoe, sex_FEMALE	1	3
9	bill_length_mm, flipper_length_mm, sex_FEMALE	1	3
10	bill_length_mm, bill_depth_mm, sex_FEMALE	1	3
11	bill_length_mm, bill_depth_mm, body_mass_g	1	3
12	bill_length_mm, flipper_length_mm, island_Biscoe	1	3
13	bill_length_mm, island_Dream, sex_FEMALE	1	3
14	bill_length_mm, bill_depth_mm, flipper_length_mm	1	3
15	bill_length_mm, island_Dream, sex_MALE	1	3
16	bill_length_mm, island_Biscoe, sex_MALE	1	3

정확도 1인 것들 중, **bill_length_mm** 는 14번 포함
bill_depth_mm, flipper_length_mm은 8번씩 포함으로 두번째로 많음

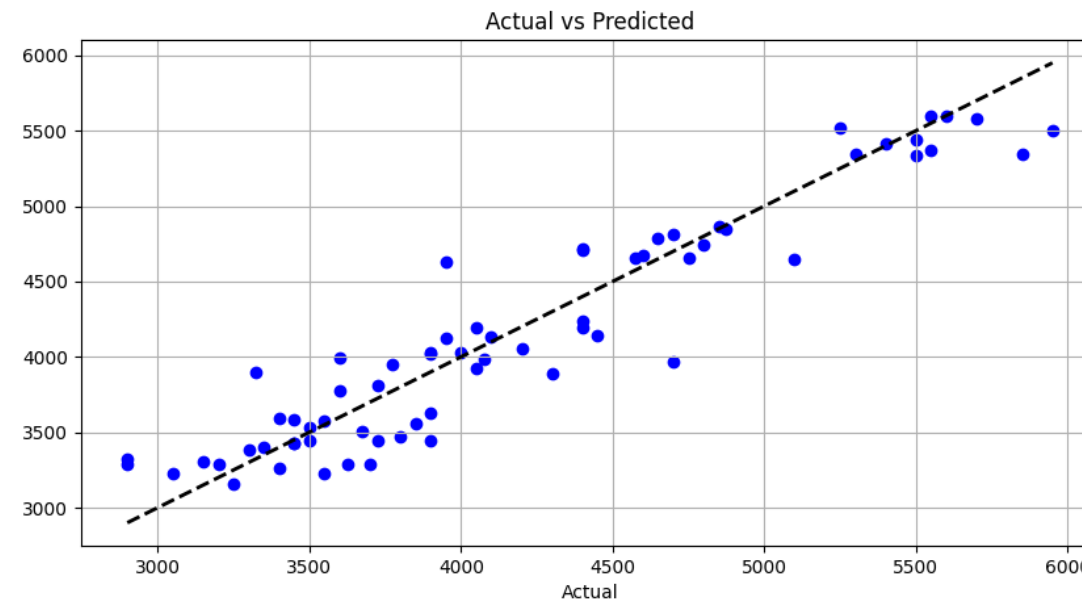
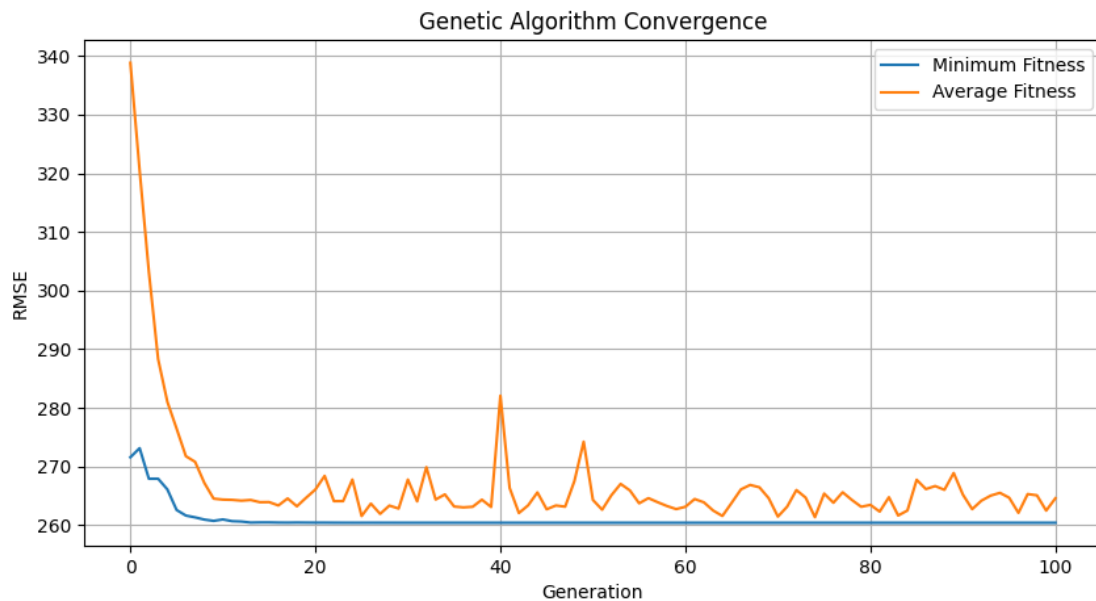
Example 3(몸무게 예측)

문제 1. 몸무게 예측

- 목표: RMSE를 최소화 시키는 **앙상블 모형** 만들기(모델 가중치 최적화)
- 사용 모델
 - Linear Regression
 - DecisionTree Regression
 - RandomForest Regression
 - Support Vector Regression
 - K-NN Regression



Example 3(몸무게 예측)



Best Weights: [1.02582795 -0.06217455 0.08237867 0.02853572 -0.0745678]

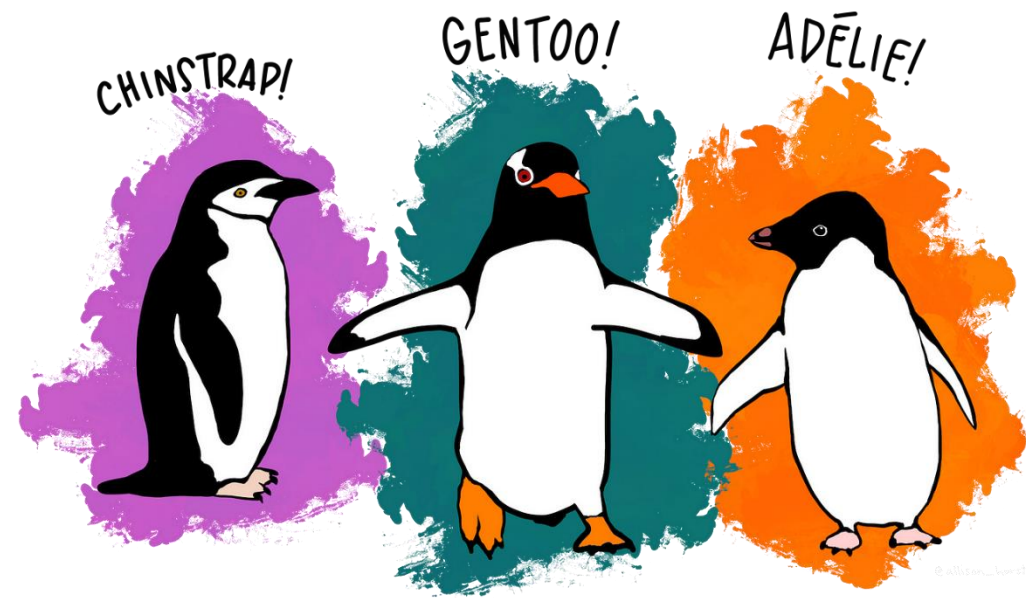
Linear, DT, RF, SVM, K-NN순

Final RMSE: 260.4304628170861

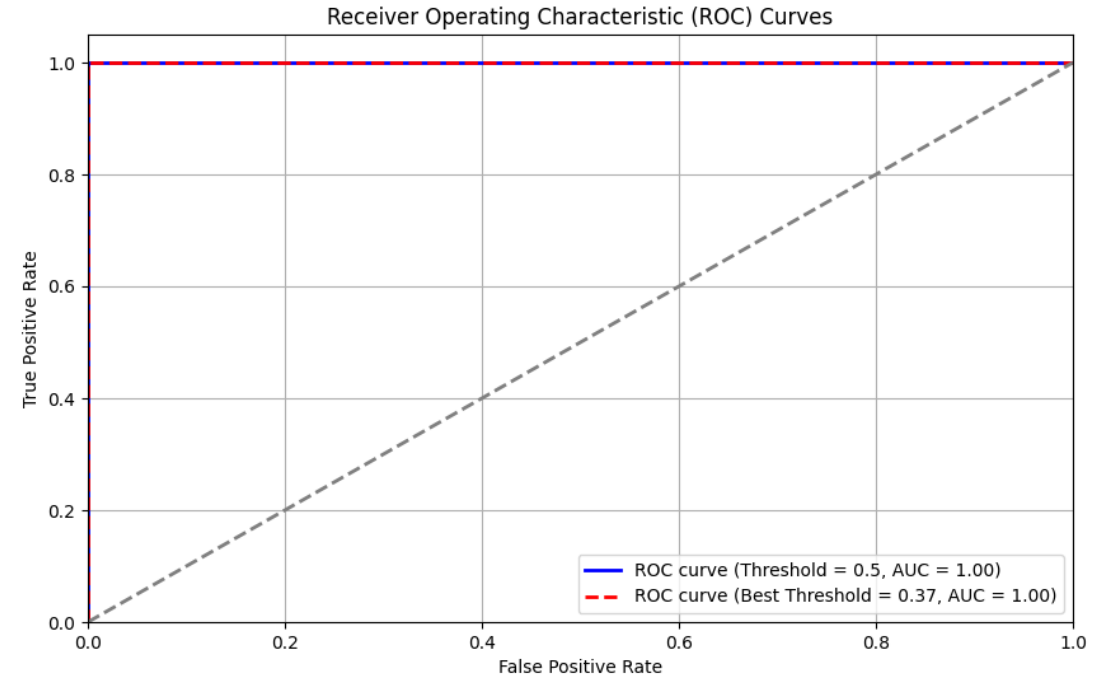
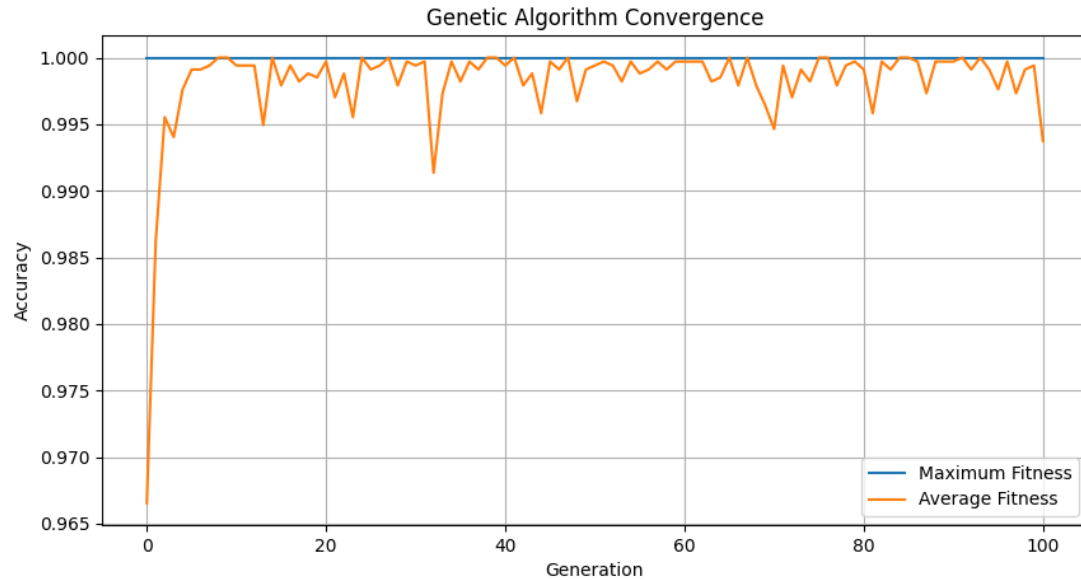
Example 4(펭귄 종 예측)

문제 2. 펭귄 종 예측

- 목표: 확률 값(0-1) 바탕으로 **최적의 임계값** 탐색
- 사용 모델: Logistic Regression



Example 4(펭귄 종 예측)



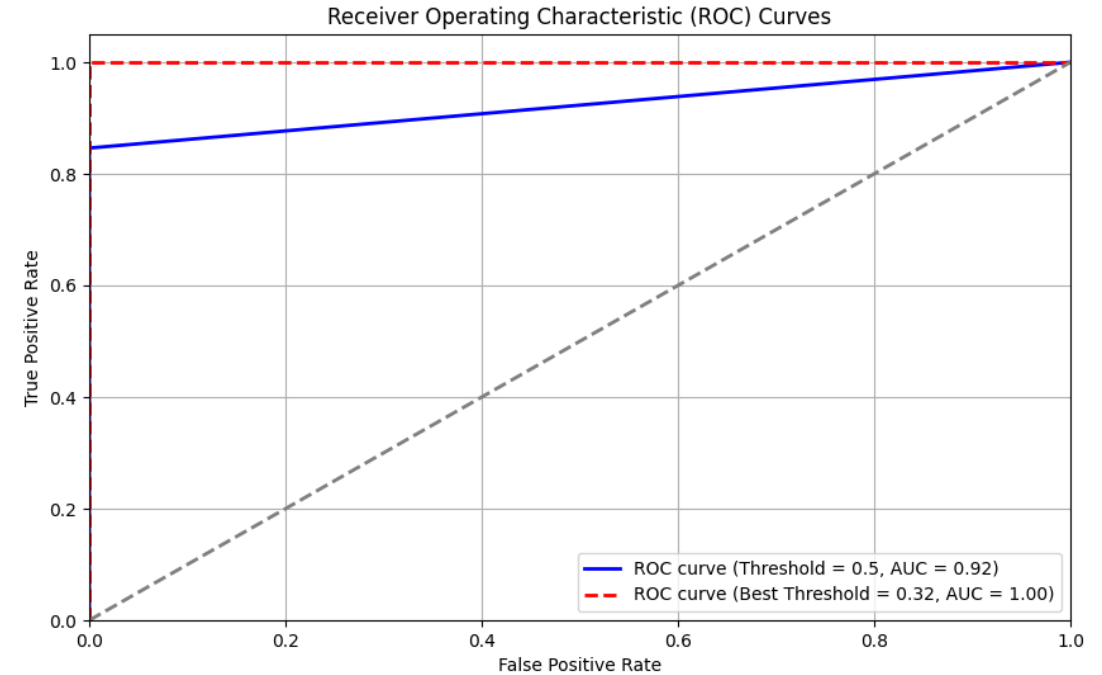
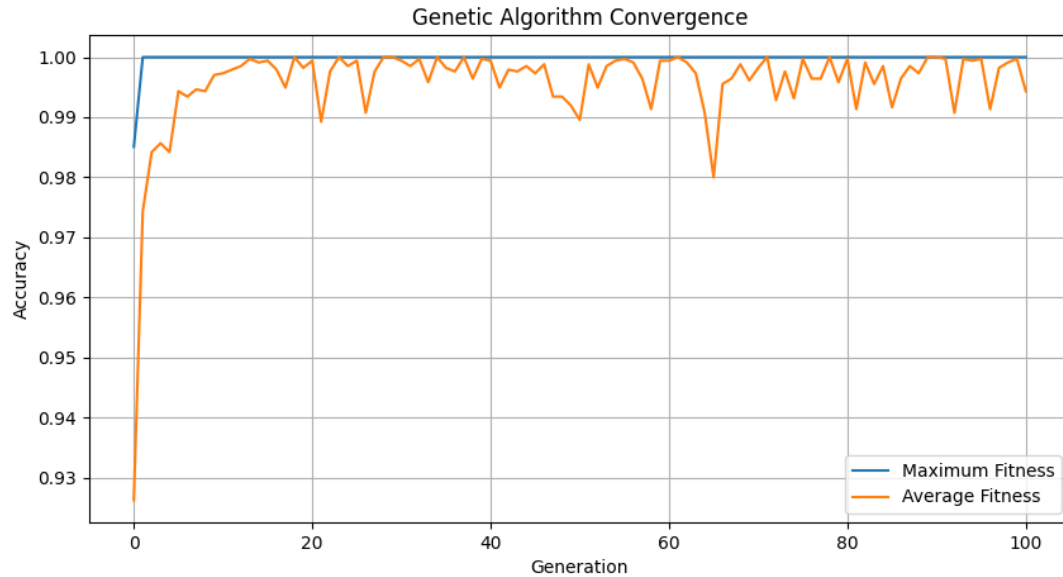
[Adelie]

Best Threshold: 0.37

Best Threshold Accuracy: 1.0

Default(0.5) Accuracy: 1.0

Example 4(펭귄 종 예측)



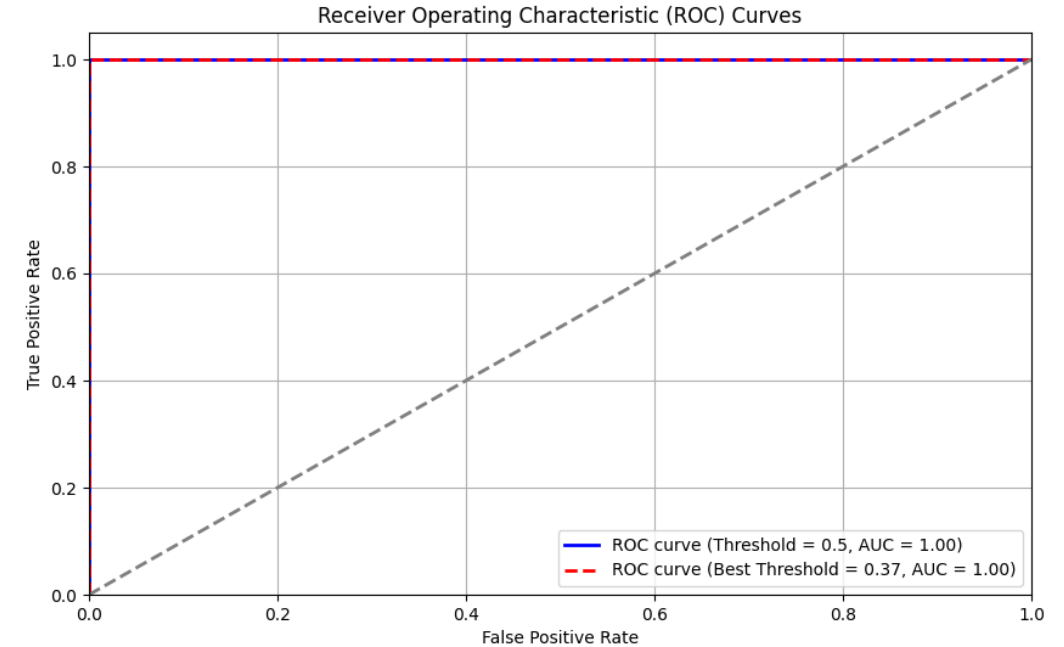
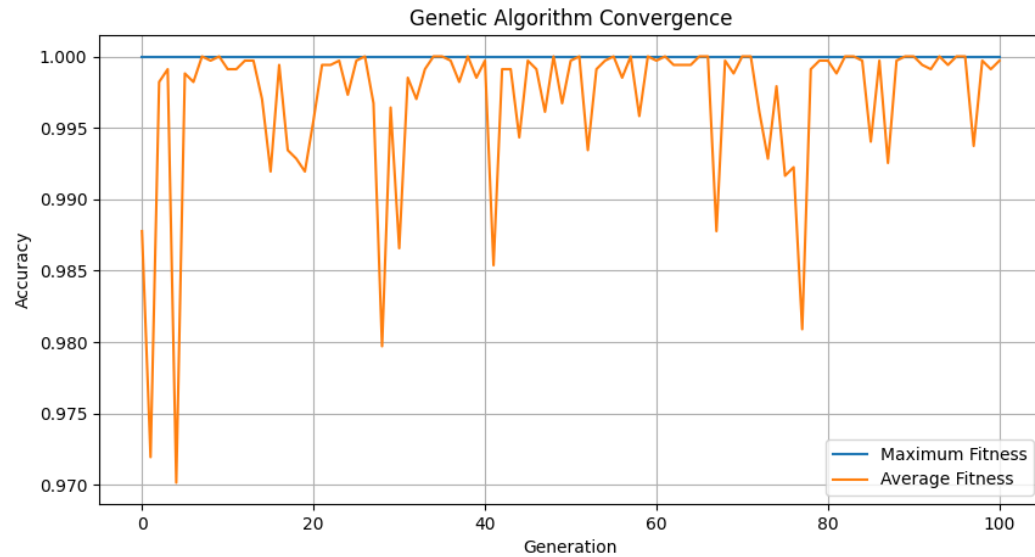
[Chinstrap]

Best Threshold: 0.3241

Best Threshold Accuracy: 1.0

Default(0.5) Accuracy: 0.9701492537313433

Example 4(펭귄 종 예측)



[Gentoo]

Best Threshold: 0.37

Best Threshold Accuracy: 1.0

Default(0.5) Accuracy: 1.0

번외(TPOT 라이브러리)

TPOT 라이브러리

- 예측 모델링 작업을 위한 AutoML 라이브러리
- genetic programming으로 머신러닝 파이프라인을 최적화
- 여러 머신러닝 모델을 기반으로 feature selection과 hyperparameter 튜닝을 자동으로 수행

```
# Average CV score on the training set was: -89667.24995761643
exported_pipeline = make_pipeline(
    StackingEstimator(estimator=RandomForestRegressor(bootstrap=True, max_features=1.0, min_samples_leaf=13, min_samples_split=13, n_estimators=100)),
    KNeighborsRegressor(n_neighbors=13, p=1, weights="uniform")
)

exported_pipeline.fit(training_features, training_target)
results = exported_pipeline.predict(testing_features)
```

변수 모두 사용 / 스택킹 방법으로 RF, KNN 사용할 때 가장 최적의 값이 나옴

RMSE: 249.063422719208

Thanks