

분류 모형

1. 이진 분류와 로직스틱 회귀 모형 소개

1. 이진 분류와 로지스틱 회귀 모형 소개

이진 분류 (Binary Classification)

이진 분류는 데이터를 두 개의 클래스 중 하나로 분류하는 문제를 다룬다. 예를 들어, 이메일이 스팸인지 아닌지, 특정 질병이 있는지 없는지 등을 예측하는 문제이다. 이진 분류를 수행하기 위해 다양한 알고리즘이 사용될 수 있다. 대표적인 알고리즘으로는 **결정 트리**, **서포트 벡터 머신**, **나이브 베이즈**, **k-최근접 이웃** 등이 있다.

로지스틱 회귀 (Logistic Regression)

로지스틱 회귀는 통계학적 모델로, 이진 분류 문제를 해결하기 위해 사용되는 알고리즘 중 하나이다. 로지스틱 회귀는 독립 변수들의 선형 결합을 통해 종속 변수의 확률을 모델링한다. 이 확률을 **로지스틱 함수(시그모이드 함수)**를 통해 0과 1 사이의 값으로 변환하여, 특정 클래스에 속할 확률을 예측한다.

1. 이진 분류와 로지스틱 회귀 모형 소개

주요 차이점

1. 모델의 본질:

1. **이진 분류**: 일반적인 분류 문제로, 다양한 알고리즘을 사용하여 데이터를 두 클래스 중 하나로 분류한다.
2. **로지스틱 회귀**: 이진 분류 문제를 해결하기 위한 특정 알고리즘으로, 독립 변수들의 선형 결합을 바탕으로 종속 변수의 확률을 예측한다.

2. 사용되는 알고리즘:

1. **이진 분류**: 여러 분류 알고리즘이 사용될 수 있다 (결정 트리, 서포트 벡터 머신, k-최근접 이웃 등).
2. **로지스틱 회귀**: 이진 분류 문제를 해결하기 위한 회귀 기반 알고리즘이다.

1. 이진 분류와 로지스틱 회귀 모형 소개

3. 결과 해석:

1. **이진 분류:** 데이터가 어느 클래스에 속하는지를 결정.
2. **로지스틱 회귀:** 특정 클래스에 속할 확률을 예측하며, 이 확률을 기준으로 최종 클래스를 결정

4. 적용 가능성:

1. **이진 분류:** 다양한 형태의 데이터와 문제에 적용 가능
2. **로지스틱 회귀:** 주로 연속형 독립 변수와 이산형 종속 변수를 다루는 문제에 적합.

2. 로직스틱 회귀 모형

2. 로지스틱 회귀 모형

1. 로지스틱 회귀 개요

로지스틱 회귀(Logistic Regression) 모형은 종속 변수가 이진형(binary)인 경우에 사용되는 통계적 모형이다. 이 모형은 특정 사건이 발생할 확률을 예측하는 데 사용되며, 분류 문제에서 중요한 역할을 한다.

2. 로지스틱 회귀 기본 개념

로지스틱 회귀 모형은 선형 회귀와 달리, 종속 변수가 0과 1 사이의 값을 가지도록 한다. 이를 위해 로지스틱 함수(시그모이드 함수)를 사용하여 출력값을 확률로 변환한다.

- **종속 변수:** 이진형 값(0 또는 1)을 가지는 변수이다. 예를 들어, 환자의 질병 유무(있음=1, 없음=0) 또는 이메일의 스팸 여부(스팸=1, 스팸 아님=0) 등이 있다.
- **독립 변수:** 종속 변수에 영향을 미치는 연속형 또는 범주형 변수들이다.

2. 로지스틱 회귀 모형

3. 로지스틱 함수

로지스틱 함수는 다음과 같은 형태를 가진다:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

여기서 z 는 독립 변수의 선형 결합이다:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

- β_0 : 절편(intercept)
- $\beta_1, \beta_2, \dots, \beta_n$: 독립 변수 x_1, x_2, \dots, x_n 에 대한 회귀 계수

로지스틱 함수는 입력값 z 를 0과 1 사이의 확률 값으로 변환한다.

오일러의 수(자연상수 e)

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \\ = 2.71828182845904523536 \dots$$

2. 로지스틱 회귀 모형

4. 최대우도 추정법

로지스틱 회귀 모형의 회귀 계수 β 는 최대우도 추정법(Maximum Likelihood Estimation, MLE)을 사용하여 추정된다. 최대우도 추정법은 주어진 데이터에서 관찰된 결과가 실제로 일어날 확률을 최대화하는 파라미터 값을 찾는 방법이다.

우도 함수(Likelihood Function)는 다음과 같이 정의된다:

$$L(\beta) = \prod_{i=1}^N P(y_i|x_i; \beta)$$

로그 우도 함수(Log-Likelihood Function)는 우도 함수의 로그를 취한 형태로, 계산의 편의를 위해 사용된다:

$$\log L(\beta) = \sum_{i=1}^N [y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i))]$$

최대화된 로그 우도 함수를 통해 최적의 회귀 계수 β 를 추정한다.

5. 활용 분야

로지스틱 회귀 모형은 다양한 분야에서 활용된다:

- **의료 분야:** 질병 예측, 환자 생존율 예측 등
- **마케팅 분야:** 고객 이탈 예측, 구매 가능성 예측 등
- **금융 분야:** 대출 상환 여부 예측, 신용 카드 사기 탐지 등
- **사회과학 분야:** 투표 행동 분석, 범죄 발생 예측 등

.

6. 다중 분류 확장

로지스틱 회귀는 기본적으로 이진 분류에 사용되지만, 다중 클래스 분류 문제에서도 사용할 수 있다. 대표적인 방법으로는 다중 클래스 로지스틱 회귀(multi-class logistic regression)와 소프트맥스 회귀(softmax regression)가 있다.

- **일대다(One-vs-Rest, OvR):** 각 클래스에 대해 이진 로지스틱 회귀 모델을 학습하고, 가장 높은 확률을 가진 클래스를 선택한다.
- **소프트맥스 회귀:** 모든 클래스에 대해 동시에 확률을 계산하고, 확률의 합이 1이 되도록 정규화하여 가장 높은 확률을 가진 클래스를 선택한다.

8. 장점과 단점

로지스틱 회귀 모형은 다음과 같은 장점과 단점을 가진다:

• 장점:

- 이해하기 쉽고 해석이 용이하다.
- 계산 비용이 낮아 대규모 데이터에서도 효율적으로 동작한다.
- 과적합을 방지하기 위해 정규화 기법(L1, L2 정규화 등)을 사용할 수 있다.

• 단점:

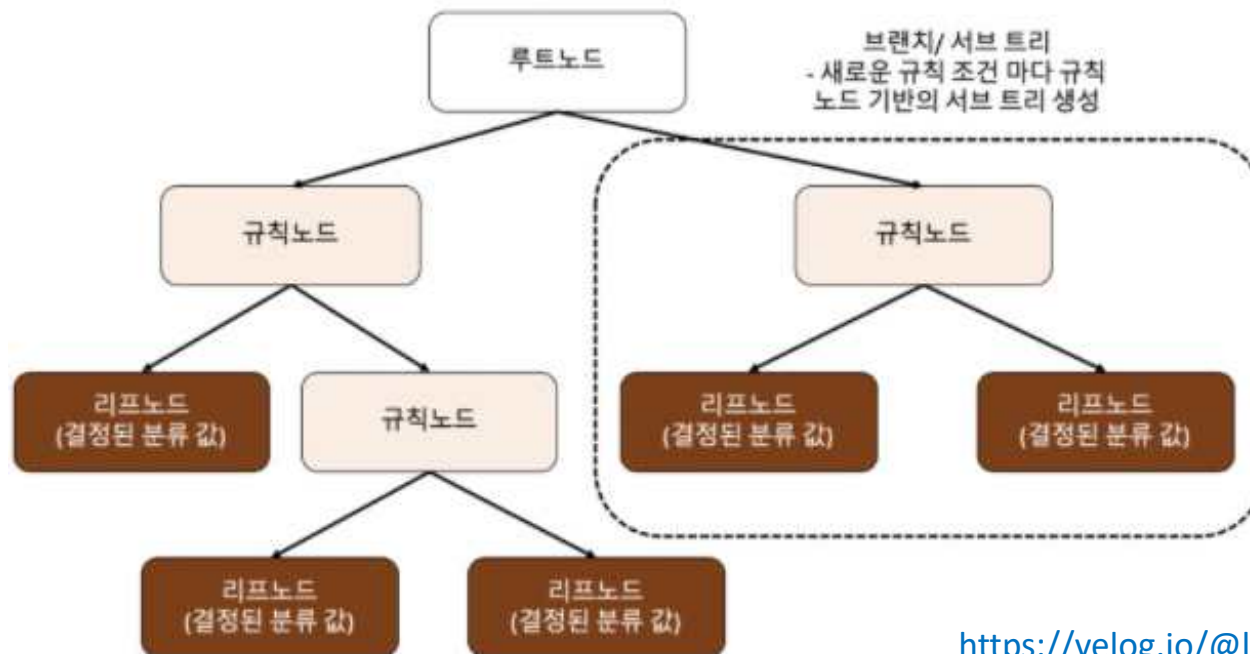
- 비선형 관계를 잘 설명하지 못한다.
- 독립 변수들 간의 다중 공선성(multicollinearity)에 민감하다.
- 고차원 데이터에서 성능이 저하될 수 있다.

3. 의사 결정 트리(Decision Tree)

3. 의사결정 트리(Decision Tree)

Decision Tree결정트리 모델은 ?

데이터에 있는 규칙을 학습을 통해 자동으로 찾아내 트리 (Tree)기반의 분류 규칙을 만드는 것으로 이 모양이 나무를 닮아 Tree모델이다.



루트노드(root) : 트리가 시작된 곳(뿌리)
규칙(중간)노드 : 규칙조건이 되는 곳
리프(터미널)노드 : 결정된 클래스 값

<https://velog.io/@ljs7463/%EA%B2%B0%EC%A0%95%ED%8A%B8%EB%A6%ACDecision-Tree%EB%AA%A8%EB%8D%B8Model>

3. 의사결정 트리(Decision Tree)

작동 원리

- 1.루트 노드 생성:** 트리의 최상단에 루트 노드를 생성하고, 전체 데이터를 이 노드에 배치한다.
- 2.특성 선택:** 데이터를 가장 잘 분할할 수 있는 특성을 선택한다. 일반적으로 엔트로피(entropy)나 지니 지수(gini index) 등의 기준을 사용한다.
- 3.노드 분할:** 선택된 특성을 기준으로 데이터를 분할하여 새로운 하위 노드를 생성한다.
- 4.반복 분할:** 하위 노드에 대해 위의 과정을 반복하여 데이터를 계속 분할한다.
- 5.종료 조건:** 모든 데이터를 분류하거나 예측할 수 있을 때까지, 또는 더 이상 분할할 의미가 없을 때까지 분할을 반복한다.

3. 의사결정 트리(Decision Tree)

엔트로피와 정보 이득

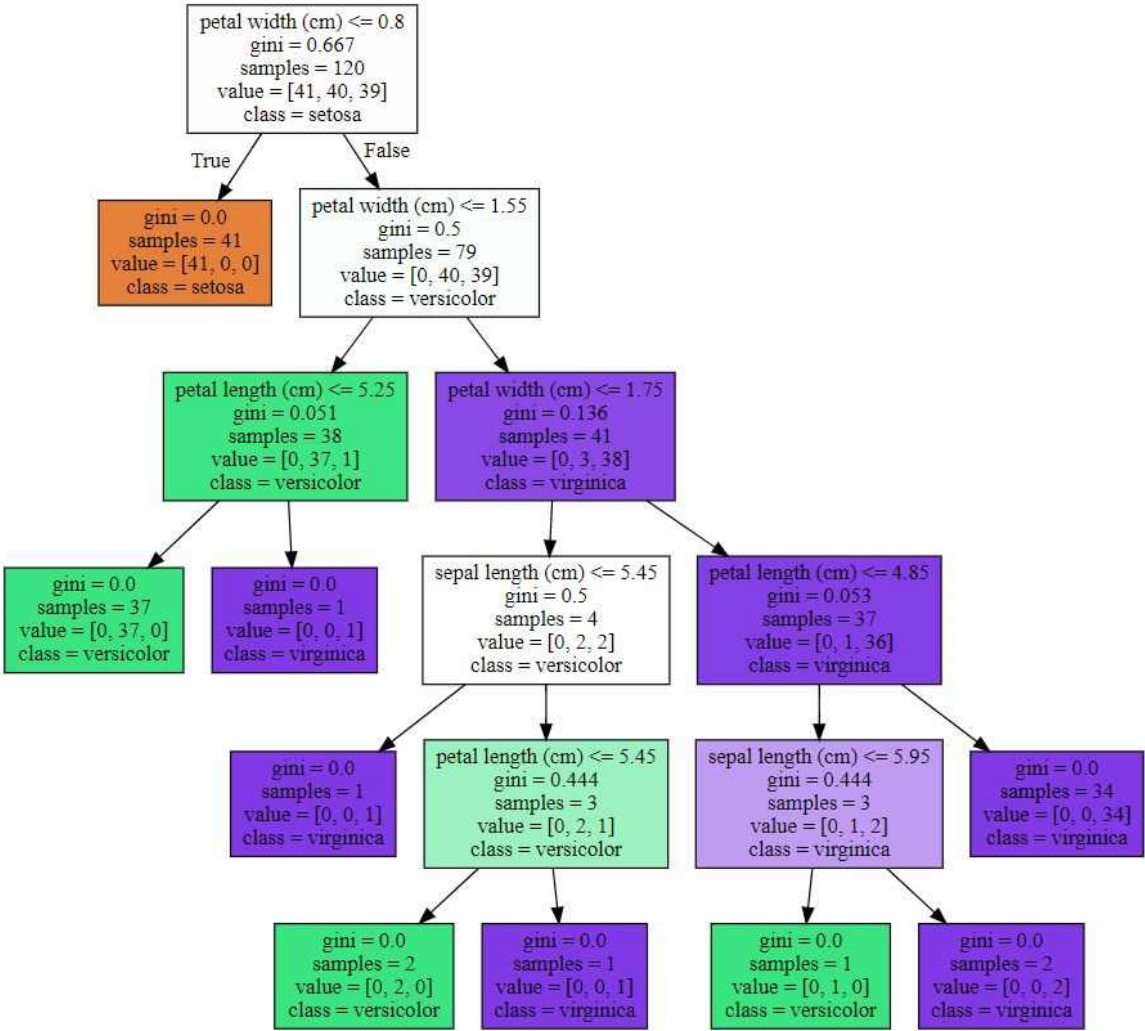
- **엔트로피(Entropy)**: 데이터의 불확실성을 나타내는 척도이다. 엔트로피가 낮을수록 데이터의 불확실성이 적다.
- **정보 이득(Information Gain)**: 특정 특성을 기준으로 데이터를 분할했을 때, 엔트로피가 얼마나 감소하는지를 나타낸다. 정보 이득이 큰 특성을 선택하여 데이터를 분할한다.

지니 지수

- **지니 지수(Gini Index)**: 불순도를 측정하는 또 다른 방법이다. 값이 작을수록 데이터가 순수하게 분할되었음을 의미한다.
- **지니 감소(Gini Decrease)**: 지니 지수를 기준으로 특성을 선택하여 분할할 때의 순수도 증가를 측정한다.

3. 의사결정 트리(Decision Tree)

결정트리 예



가지치기

- **가지치기(Pruning)**: 과적합을 방지하기 위해 트리의 일부 가지를 제거하는 과정이다.
- **사전 가지치기(Pre-pruning)**: 트리를 생성하는 동안 미리 가지치기를 수행하여 트리의 깊이를 제한한다.
- **사후 가지치기(Post-pruning)**: 트리를 완성한 후, 성능을 평가하여 불필요한 가지를 제거한다.

3. 의사결정 트리(Decision Tree)

장점

- 직관적이고 이해하기 쉬운 모델을 제공한다.
- 다양한 데이터 타입을 처리할 수 있다.
- 데이터 전처리가 거의 필요 없다.

단점

- 과적합에 취약할 수 있다.
- 작은 변동에도 트리 구조가 크게 변할 수 있다.
- 연속형 변수를 다루기 어려울 수 있다.

활용 예시

- **분류 문제:** 이메일 스팸 필터링, 질병 진단, 고객 이탈 예측 등
- **회귀 문제:** 주택 가격 예측, 주식 가격 예측 등

결정트리 알고리즘은 다양한 데이터 분석 및 예측 문제에 적용할 수 있으며, 이해하기 쉬운 모델을 제공하여 많은 분야에서 널리 사용된다.

.

3. 의사결정 트리(Decision Tree)

분류 모델 평가

로지스틱 회귀 모형의 성능을 평가하기 위해 다양한 지표가 사용된다

- **정확도(Accuracy):** 전체 예측 중에서 맞게 예측한 비율이다.
- **정밀도(Precision):** 양성으로 예측한 것 중에서 실제로 양성인 비율이다.
- **재현율(Recall):** 실제 양성인 것 중에서 양성으로 예측한 비율이다.
- **F1 점수:** 정밀도와 재현율의 조화 평균으로, 두 지표 간의 균형을 평가한다.
- **ROC 곡선과 AUC:** 수신자 조작 특성 곡선(ROC Curve)과 곡선 아래 면적(AUC)으로, 모델의 분류 능력을 평가한다.

3. 의사결정 트리(Decision Tree)

혼동 행렬

		예측 클래스 (Predicted Class)	
		Negative(0)	Positive(1)
실제 클래스 (Actual Class)	Negative(0)	TN (True Negative)	FP (False Positive)
	Positive(1)	FN (False Negative)	TP (True Positive)

$$\text{정밀도} = \text{TP} / (\text{FP} + \text{TP})$$

$$\text{재현율} = \text{TP} / (\text{FN} + \text{TP})$$

$$\text{F1 스코어} = 2 / ((1/\text{재현율}) + (1/\text{정밀도}))$$

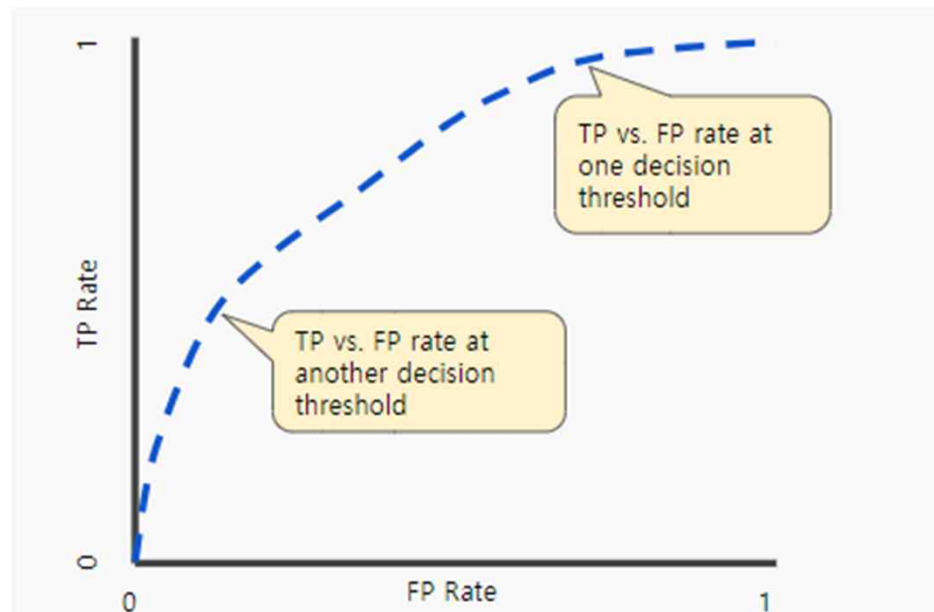
3. 의사결정 트리(Decision Tree)

ROC AUC(Area Under Curve) : ROC 곡선의 면적, 1에 가까울수록 좋음, 1이 최대

ROC(Receiver Operation Curve) :

수신자 판단 곡선

- TNR(True Negative Rate, 특이성) : $TN / (FP + TN)$
- x축을 FPR(False Positive Rate) : $FP / (FP + TN) = 1 - TNR = 1 - \text{특이성}$
- y축을 TPR(True Positive Rate, 재현율, 민감도) : $TP / (FN + TP)$



4. 이진 분류 모델 구현 실습


4. 이진 분류 모델 구현 실습

Breast Cancer.csv 데이터 파일 엑셀로 확인


#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE		
1	mean radi	mean text	mean peri	mean are	mean smc	mean con	mean con	mean con	mean con	mean sym	mean frac	radius err	texture en	perimeter	area error	smoothne	compactn	concavity	concave	symmetry	fractal	din	worst radi	worst text	worst peri	worst are	worst smc	worst con	worst con	worst con	worst sym	worst frac	target
2	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.0064	0.04904	0.05373	0.01587	0.03003	0.00619	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601	0.1189	0		
3	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.00523	0.01308	0.0186	0.0134	0.01389	0.00353	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275	0.08902	0		
4	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.00457	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613	0.08758	0		
5	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.00921	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.173	0		
6	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.00512	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364	0.07678	0		
7	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.00508	15.47	23.75	103.4	741.6	0.1791	0.5249	0.5355	0.1741	0.3985	0.1244	0		
8	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.00431	0.01382	0.02254	0.01039	0.01369	0.00218	22.88	27.66	153.2	1606	0.1442	0.2576	0.3784	0.1932	0.3063	0.08368	0		
9	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.00881	0.03029	0.02488	0.01448	0.01486	0.00541	17.06	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196	0.1151	0		
10	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.00573	0.03502	0.03553	0.01226	0.02143	0.00375	15.49	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378	0.1072	0		
11	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.00715	0.07217	0.07743	0.01432	0.01789	0.01008	15.09	40.68	97.65	711.4	0.1853	1.058	1.105	0.221	0.4366	0.2075	0		
12	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.00403	0.00927	0.01101	0.00759	0.0146	0.00304	19.19	33.88	123.8	1150	0.1181	0.1551	0.1459	0.09975	0.2948	0.08452	0		
13	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.00577	0.04061	0.02791	0.01282	0.02008	0.00414	20.42	27.28	136.5	1299	0.1396	0.5609	0.3965	0.181	0.3792	0.1048	0		
14	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.00314	0.08297	0.0889	0.0409	0.04484	0.01284	20.96	29.94	151.7	1332	0.1037	0.3903	0.3639	0.1767	0.3176	0.1023	0		
15	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.00977	0.03126	0.05051	0.01992	0.02981	0.003	16.84	27.66	112	876.5	0.1131	0.1924	0.2322	0.1119	0.2809	0.06287	0		
16	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.00643	0.05936	0.05501	0.01628	0.01961	0.00809	15.03	32.01	108.8	697.7	0.1651	0.7725	0.6943	0.2208	0.3596	0.1431	0		
17	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.00561	0.0424	0.04741	0.0109	0.01857	0.00547	17.46	37.13	124.1	943.2	0.1678	0.6577	0.7026	0.1712	0.4218	0.1341	0		
18	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.00572	0.01162	0.01998	0.01109	0.0141	0.00209	19.07	30.88	123.4	1138	0.1464	0.1871	0.2914	0.1609	0.3029	0.08216	0		
19	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.00703	0.02501	0.03188	0.01297	0.01689	0.00414	20.96	31.48	136.8	1315	0.1789	0.4233	0.4784	0.2073	0.3706	0.1142	0		
20	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.00649	0.01893	0.03391	0.01521	0.01356	0.002	27.32	30.88	186.8	2398	0.1512	0.315	0.5372	0.2388	0.2768	0.07615	0		
21	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.00846	0.0146	0.02387	0.01315	0.0198	0.0023	15.11	19.26	99.7	711.2	0.144	0.1773	0.239	0.1288	0.2977	0.07259	1		
22	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.0041	0.01898	0.01698	0.00649	0.01678	0.00243	14.5	20.49	96.09	630.5	0.1312	0.2776	0.189	0.07283	0.3184	0.08183	1		
23	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.00961	0.01432	0.01985	0.01421	0.02027	0.00297	10.23	15.66	65.13	314.9	0.1324	0.1148	0.08867	0.06227	0.245	0.07773	1		
24	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.00679	0.05328	0.06446	0.02252	0.03672	0.00439	18.07	19.08	125.1	980.9	0.139	0.5954	0.6305	0.2393	0.4667	0.09946	0		
25	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.00473	0.01259	0.01715	0.01038	0.01083	0.00199	29.17	35.59	188	2615	0.1401	0.26	0.3155	0.2009	0.2822	0.07526	0		
26	16.65	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633	0.8068	0.9017	5.455	102.6	0.00605	0.01882	0.02741	0.0113	0.01468	0.0028	26.46	31.56	177	2215	0.1805	0.3578	0.4695	0.2095	0.3613	0.09564	0		
27	17.14	16.4	116	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413	1.046	0.976	7.276	111.4	0.00803	0.03799	0.03732	0.02397	0.02308	0.00744	22.25	21.4	152.4	1461	0.1545	0.3949	0.3853	0.255	0.4066	0.1059	0		
28	14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924	0.2545	0.9832	2.11	21.05	0.00445	0.03055	0.02681	0.01352	0.01454	0.00371	17.62	33.21	122.4	896.9	0.1525	0.6643	0.5539	0.2701	0.4264	0.1275	0		
29	18.61	20.25	122.1	1094	0.0944	0.1066	0.149	0.07731	0.1697	0.05699	0.8529	1.849	5.632	93.54	0.01075	0.02722	0.05081	0.01911	0.02293	0.00422	21.31	27.26	139.9	1403	0.1338	0.2117	0.3446	0.149	0.2341	0.07421	0		
30	15.3	25.27	102.4	732.4	0.1082	0.1697	0.1683	0.08751	0.1926	0.0654	0.439	1.012	3.498	43.5	0.00523	0.03057	0.03576	0.01083	0.01768	0.00297	20.27	36.71	149.3	1269	0.1641	0.611	0.6335	0.2024	0.4027	0.09876	0		
31	17.57	15.05	115	955.1	0.09847	0.1157	0.09875	0.07953	0.1739	0.06149	0.6003	0.8225	4.655	61.1	0.00563	0.03033	0.03407	0.01354	0.01925	0.00374	20.01	19.52	134.9	1227	0.1255	0.2812	0.2489	0.1456	0.2756	0.07919	0		
32	18.63	25.11	124.8	1088	0.1064	0.1887	0.2319	0.1244	0.2183	0.06197	0.8307	1.466	5.574	105	0.00625	0.03374	0.05196	0.01158	0.02007	0.00456	23.15	34.01	160.5	1670	0.1491	0.4257	0.6133	0.1848	0.3444	0.09782	0		
33	11.84	18.7	77.93	440.6	0.1109	0.1516	0.1218	0.05182	0.2301	0.07799	0.4825	1.03	3.475	41	0.00555	0.03414	0.04205	0.01044	0.02273	0.00567	16.82	28.12	119.4	888.7	0.1637	0.5775	0.6956	0.1546	0.4761	0.1402	0		
34	17.02	23.98	112.8	899.3	0.1197	0.1496	0.2417	0.1203	0.2248	0.06382	0.6009	1.398	3.999	67.78	0.00827	0.03082	0.05042	0.01112	0.02102	0.00385	20.88	32.09	136.1	1344	0.1634	0.3559	0.5588	0.1847	0.353	0.08482	0		
35	19.27	26.47	127.9	1162	0.09401	0.1719	0.1657	0.07593	0.1853	0.06261	0.5558	0.6062	3.528	68.17	0.00502	0.03318	0.03497	0.00964	0.01543	0.0039	24.15	30.9	161.4	1813	0.1509	0.659	0.6091	0.1785	0.3672	0.1123	0		
36	16.13	17.88	107	807.2	0.104	0.1559	0.1354	0.07752	0.1998	0.06515	0.334	0.6857	2.183	35.03	0.00419	0.02868	0.02664	0.00907	0.01703	0.00382	20.21	27.26	132.7	1261									

4. 이진 분류 모델 구현 실습

 Breast Cancer.csv
스프레드시트

 데이터 분석해줘(한글로).



 데이터의 구조는 다음과 같습니다:

- 총 31개의 열(column)과 여러 행(row)으로 구성되어 있습니다.
- 주요 열(column)들은 암의 다양한 특성을 나타내며, 마지막 열 "target"은 종양이 양성(0)인지 악성(1)인지를 나타냅니다.

각 열의 의미는 다음과 같습니다:

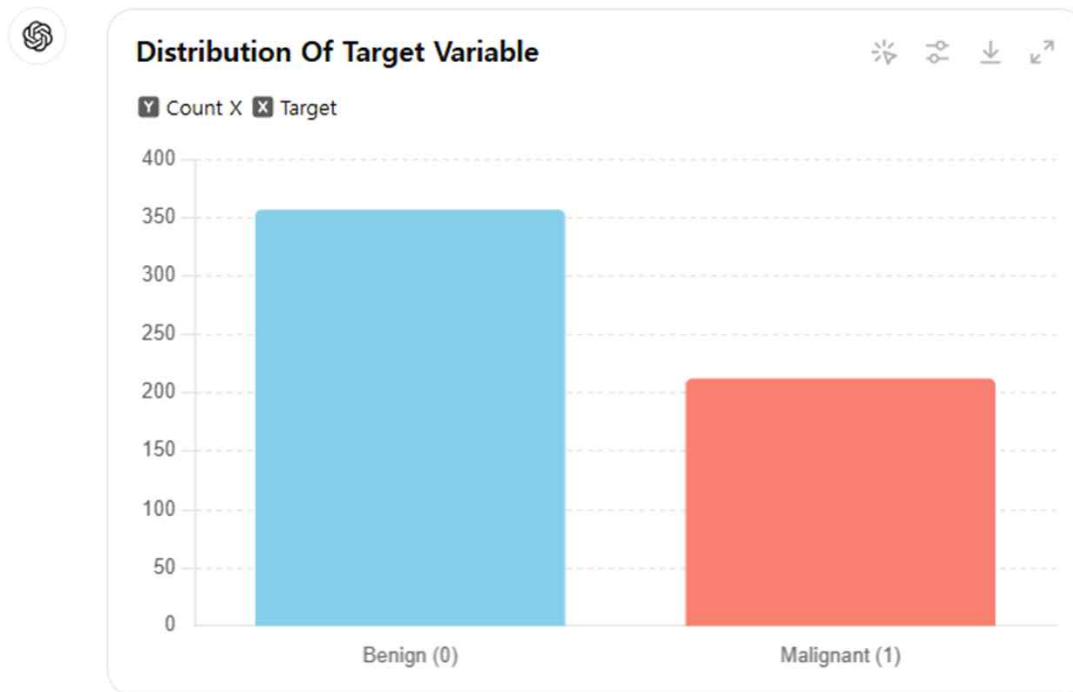
- ``mean radius``: 평균 반경
- ``mean texture``: 평균 텍스처
- ``mean perimeter``: 평균 둘레
- ``mean area``: 평균 면적
- ``mean smoothness``: 평균 평활도
- ``mean compactness``: 평균 조밀도
- ``mean concavity``: 평균 오목도
- ``mean concave points``: 평균 오목한 점의 수

4. 이진 분류 모델 구현 실습

- ``mean symmetry``: 평균 내성성
- ``mean fractal dimension``: 평균 프랙탈 차원
- ``worst radius``: 최악의 반경
- ``worst texture``: 최악의 텍스처
- ``worst perimeter``: 최악의 둘레
- ``worst area``: 최악의 면적
- ``worst smoothness``: 최악의 평활도
- ``worst compactness``: 최악의 조밀도
- ``worst concavity``: 최악의 오목도
- ``worst concave points``: 최악의 오목한 점의 수
- ``worst symmetry``: 최악의 대칭성
- ``worst fractal dimension``: 최악의 프랙탈 차원
- ``target``: 종양의 종류 (0: 양성, 1: 악성)

4. 이진 분류 모델 구현 실습

target의 분포를 그래프로 그려줘.



위 그래프는 `target` 변수의 분포를 보여줍니다.

- `0`은 양성 종양 (Benign)을 나타내며, `1`은 악성 종양 (Malignant)을 나타냅니다.
- 그래프에서 볼 수 있듯이, 양성 종양의 수가 악성 종양의 수보다 더 많습니다.

수치로 알려줘.

`target` 변수의 분포는 다음과 같습니다:

- 악성 종양 (Malignant, 1): 357개
- 양성 종양 (Benign, 0): 212개

4. 이진 분류 모델 구현 실습

결정 트리를 사용한 이진분류 모델 만들어주고 평가지표를 알려줘.



결정 트리를 사용한 이진 분류 모델의 평가지표는 다음과 같습니다:

- **정확도(Accuracy):** 0.9415
- **혼동 행렬(Confusion Matrix):**

lua

코드 복사

```
[[ 60,  3],  
 [  7, 101]]
```

- 양성(Benign) 종양 (0): 60개의 참 양성(True Positive), 3개의 거짓 양성(False Positive)
- 악성(Malignant) 종양 (1): 101개의 참 악성(True Negative), 7개의 거짓 음성(False Negative)

4. 이진 분류 모델 구현 실습

- 분류 보고서(Classification Report):

markdown

코드 복사

	precision	recall	f1-score	support
Benign	0.90	0.95	0.92	63
Malignant	0.97	0.94	0.95	108
accuracy			0.94	171
macro avg	0.93	0.94	0.94	171

4. 이진 분류 모델 구현 실습

markdown

코드 복사

- ****Benign****:
 - 정밀도 (Precision): 0.90
 - 재현율 (Recall): 0.95
 - F1-점수 (F1-score): 0.92
- ****Malignant****:
 - 정밀도 (Precision): 0.97
 - 재현율 (Recall): 0.94
 - F1-점수 (F1-score): 0.95

이 모델은 정확도 94.15%로 좋은 성능을 보이고 있습니다. 추가 분석이나 다른 질문이 있으시면 알려주세요.

4. 이진 분류 모델 구현 실습

이 모델의 ROC_AUC 값을 알려줘.



이 모델의 ROC AUC 값은 0.9438입니다.

ROC AUC (Receiver Operating Characteristic Area Under Curve)는 분류 모델의 성능을 평가하는 지표로, 값이 1에 가까울수록 모델의 분류 성능이 뛰어남을 의미합니다. 이 값은 모델이 양성과 음성을 얼마나 잘 구분하는지를 나타내며, 0.5는 랜덤 분류와 동일한 성능을, 1은 완벽한 분류를 의미합니다.

따라서 이 모델은 매우 좋은 성능을 보이고 있습니다. 추가로 궁금한 점이 있거나 다른 분석이 필요하시면 말씀해 주세요. [>-]



4. 이진 분류 모델 구현 실습

[참고] macro avg

각 클래스의 성능 지표(precision, recall, f1-score)를 단순 평균한 값이다. 이 지표는 클래스 간의 불균형이 있을 때 각 클래스의 성능을 균형 있게 평가할 수 있도록 도와준다.

	precision	recall	f1-score	support
Benign	0.90	0.95	0.92	63
Malignant	0.97	0.94	0.95	108
accuracy			0.94	171
macro avg	0.93	0.94	0.94	171
weighted avg	0.94	0.94	0.94	171

여기서 `macro avg`는 다음과 같이 계산됩니다:

- Precision: $(0.90 + 0.97) / 2 = 0.935$
- Recall: $(0.95 + 0.94) / 2 = 0.945$
- F1-score: $(0.92 + 0.95) / 2 = 0.935$

4. 이진 분류 모델 구현 실습

[참고] weighted avg

각 클래스의 성능 지표(precision, recall, f1-score)를 클래스의 샘플 수로 가중 평균한 값. 이는 클래스 간 불균형이 있을 때 전체 데이터에 대한 모델의 성능을 더 정확하게 반영할 수 있도록 도와준다.

	precision	recall	f1-score	support
Benign	0.90	0.95	0.92	63
Malignant	0.97	0.94	0.95	108
accuracy			0.94	171
macro avg	0.93	0.94	0.94	171
weighted avg	0.94	0.94	0.94	171

여기서 `weighted avg`는 다음과 같이 계산됩니다:

- Precision: $(0.90 * 63 + 0.97 * 108) / (63 + 108) = 0.94$
- Recall: $(0.95 * 63 + 0.94 * 108) / (63 + 108) = 0.94$
- F1-score: $(0.92 * 63 + 0.95 * 108) / (63 + 108) = 0.94$

5. 다중 분류 모델 구현 실습

5. 다중 분류 모델 구현 실습

iris.csv 엑셀로 열어 확인하기

	A	B	C	D	E
1	sepal length	sepal width	petal length	petal width	label
2	5.1	3.5	1.4	0.2	0
3	4.9	3	1.4	0.2	0
4	4.7	3.2	1.3	0.2	0
5	4.6	3.1	1.5	0.2	0
6	5	3.6	1.4	0.2	0
7	5.4	3.9	1.7	0.4	0
8	4.6	3.4	1.4	0.3	0
9	5	3.4	1.5	0.2	0
10	4.4	2.9	1.4	0.2	0
11	4.9	3.1	1.5	0.1	0
12	5.4	3.7	1.5	0.2	0
13	4.8	3.4	1.6	0.2	0
14	4.8	3	1.4	0.1	0
15	4.3	3	1.1	0.1	0
16	5.8	4	1.2	0.2	0
17	5.7	4.4	1.5	0.4	0
18	5.4	3.9	1.3	0.4	0
19	5.1	3.5	1.4	0.3	0
20	5.7	3.8	1.7	0.3	0

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2
150 rows x 5 columns					

5. 다중 분류 모델 구현 실습



📄 데이터셋 설명해줘(한글로)



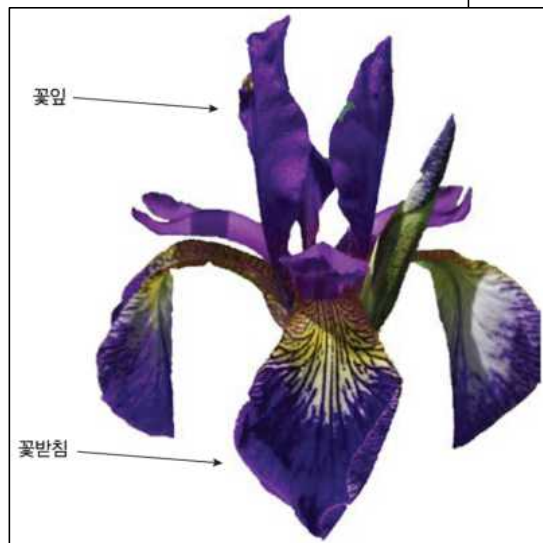
이 데이터셋은 유명한 "아이리스(Iris)" 데이터셋입니다. 각 행은 하나의 꽃 샘플을 나타내며, 각 열은 해당 샘플의 특성을 나타냅니다. 데이터셋의 열은 다음과 같습니다:

1. ``sepal length (cm)`` - 꽃받침의 길이 (센티미터)
2. ``sepal width (cm)`` - 꽃받침의 너비 (센티미터)
3. ``petal length (cm)`` - 꽃잎의 길이 (센티미터)
4. ``petal width (cm)`` - 꽃잎의 너비 (센티미터)
5. ``label`` - 꽃의 종류를 나타내는 라벨 (0, 1, 2로 구분됨)

각 라벨은 다음과 같은 꽃 종류를 의미합니다:

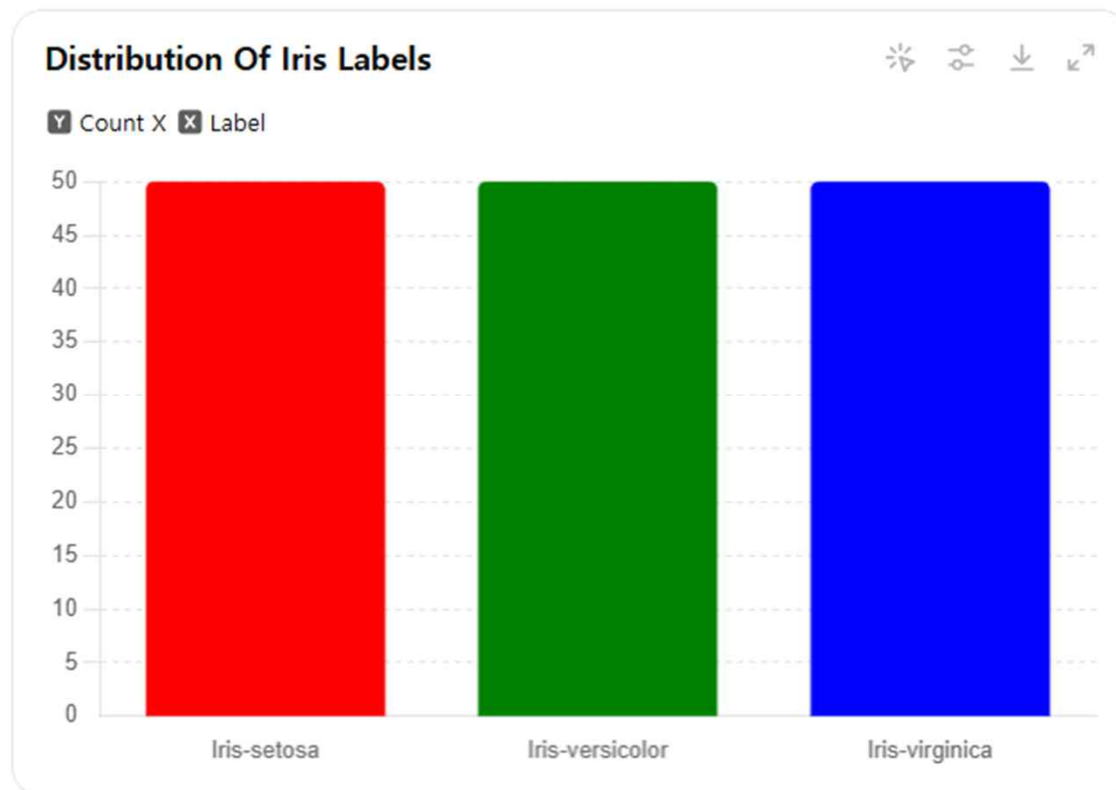
- 0: Iris-setosa
- 1: Iris-versicolor
- 2: Iris-virginica

이 데이터셋은 머신러닝 및 데이터 분석 실습에 자주 사용되며, 꽃의 특성들을 바탕으로 꽃의 종류를 예측하는 모델을 만드는 데 유용합니다. [↗]



5. 다중 분류 모델 구현 실습

label의 분포를 수치로 알려주고 그래프로 출력해줘.



라벨의 분포는 다음과 같습니다:

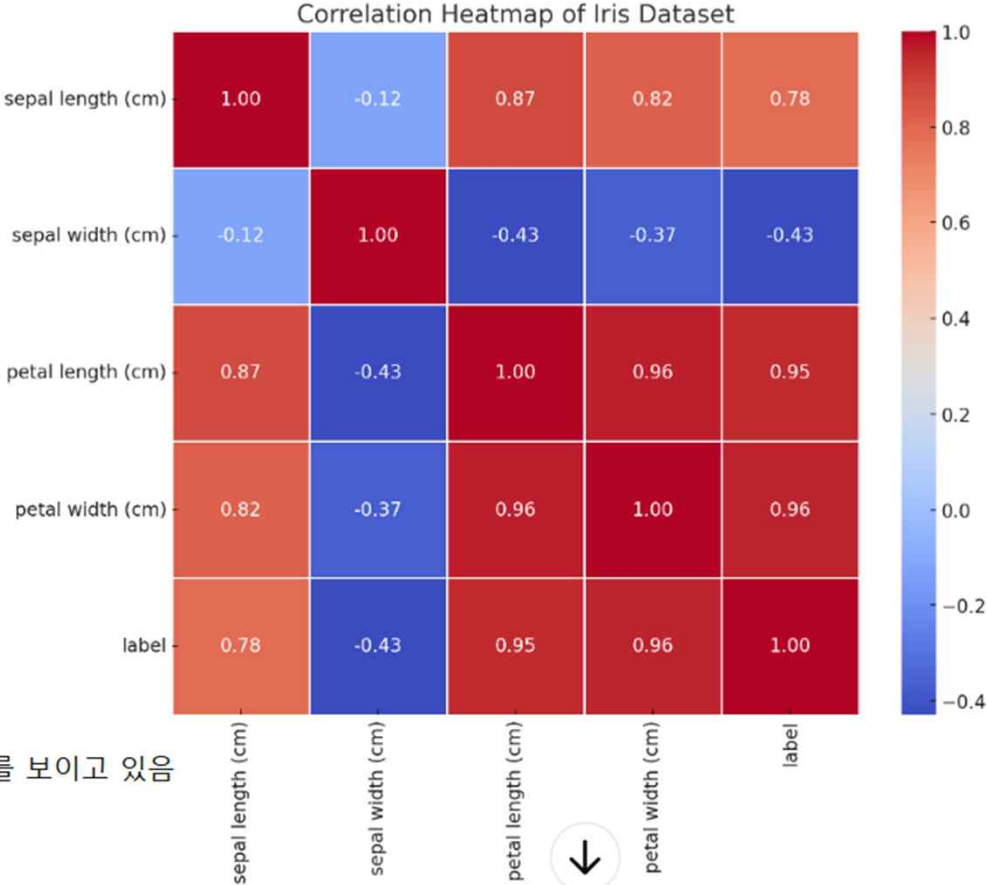
- Iris-setosa: 50개
- Iris-versicolor: 50개
- Iris-virginica: 50개

5. 다중 분류 모델 구현 실습

상관 관계를
히트맵으로
그려줘(값도
함께 표시).



Correlation Heatmap Of Iris Dataset



`petal length (cm)`와 `petal width (cm)`는 강한 양의 상관 관계를 보이고 있음

5. 다중 분류 모델 구현 실습

다중 분류 모델을 만들어주고 평가지표도 알려줘.

Confusion Matrix					↓ ↗
		Iris-setosa	Iris-versicolor	Iris-virginica	
1	Iris-setosa	10	0	0	
2	Iris-versicolor	0	9	0	
3	Iris-virginica	0	0	11	

정확도 (Accuracy)

- 1.0 (100%)

5. 다중 분류 모델 구현 실습

평가지표를 테이블 형태로 출력해줘.



Classification Report					↓ ↗
	class	precision	recall	f1-score	
1	Iris-setosa	1.0	1.0	1.0	
2	Iris-versicolor	1.0	1.0	1.0	
3	Iris-virginica	1.0	1.0	1.0	
4	accuracy				
5	macro avg	1.0	1.0	1.0	
6	weighted avg	1.0	1.0	1.0	

감사합니다