LAPORAN TUGAS Praktikum Machine Learning: Naive Bayes

Nama: Surya Dwi Satria

NIM: 434231048

Kelas: C7

BAB I – PENDAHULUAN

1.1 Latar Belakang

Naive Bayes merupakan algoritma klasifikasi yang menggunakan pendekatan probabilistik berdasarkan teorema Bayes. Algoritma ini sering digunakan karena kesederhanaannya, efisiensi tinggi, dan hasil yang cukup akurat untuk dataset dengan atribut kategorikal. Pada tugas ini, model Naive Bayes diterapkan untuk memprediksi Exam_Score siswa berdasarkan berbagai faktor seperti Hours_Studied, Attendance, Motivation_Level, dan sebagainya. Hasil perhitungan dilakukan secara manual di Excel, otomatis menggunakan Python, dan diimplementasikan melalui UI Streamlit untuk membandingkan hasilnya

1.2 Tujuan

1. Memahami proses perhitungan manual Naive Bayes.
2. Membandingkan hasil manual dengan hasil dari program Python dan Streamlit.
3. Membuat implementasi UI sederhana berbasis model Naive Bayes.

BAB II – DATASET DAN PERSIAPAN DATA

2.1 Sumber Dataset

Dataset yang digunakan berasal dari file: dataset_final.xlsx. Dataset ini berisi 20 atribut input dan 1 atribut target (Exam_Score).

2.2 Struktur Dataset

Dataset memiliki 21 kolom, dengan 20 kolom fitur dan 1 kolom target. Semua kolom bersifat kategorikal setelah preprocessing.

2.3 Preprocessing

Data telah dibersihkan dari nilai kosong, dikategorikan menjadi nilai diskrit, dan disimpan dalam format Excel untuk memudahkan perhitungan manual.

```
Tipe data tiap kolom:
 Hours_Studied                object
Attendance                    object
Parental_Involvement          object
Access_to_Resources           object
Extracurricular_Activities    object
Sleep_Hours                   object
Previous_Scores               object
Motivation_Level              object
Internet_Access               object
Tutoring_Sessions             object
Family_Income                 object
Teacher_Quality               object
School_Type                   object
Peer_Influence                object
Physical_Activity             object
Learning_Disabilities         object
Parental_Education_Level      object
Distance_from_Home            object
Gender                        object
Exam_Score                    object
dtype: object
```

BAB III – PERHITUNGAN MANUAL NAIVE BAYES

3.1 Menentukan Data Uji (x_test)

Satu baris data dipilih sebagai data uji (x_test) dengan kombinasi nilai fitur tertentu.

Feature	Nilai_x_test
Hours_Studied	High
Attendance	Met
Parental_Involvement	Medium
Access_to_Resources	High
Extracurricular_Activities	Yes
Sleep_Hours	Medium
Previous_Scores	High
Motivation_Level	Medium
Internet_Access	Yes
Tutoring_Sessions	Low
Family_Income	Medium
Teacher_Quality	High
School_Type	Public
Peer_Influence	Positive
Physical_Activity	Medium
Learning_Disabilities	No
Parental_Education_Level	College
Distance_from_Home	Near
Gender	Male

3.2 Perhitungan Frekuensi Tiap Fitur

Langkah pertama adalah menghitung berapa kali setiap nilai fitur muncul pada tiap kelas target (High, Medium, Low) menggunakan fungsi COUNTIFS di Excel.

Rumus Excel:
=COUNTIFS(Dataset!A:A, B2, Dataset!N:N, "High")

	A	B	C	D	E
1	Fitur	Nilai	High	Medium	Total
2	Hours_Studied	High	1834	726	2560
3		Medium	1266	1605	2871
4		Low	59	276	335
5	Attendance	Met	2862	936	3798
6		NotMet	513	1688	2201
7	Parental_Involvement	Low	548	661	1209
8		Medium	1745	1298	3043
9		High	954	756	1710
10	Access_to_Resources	High	1171	765	1936
11		Medium	1683	1344	3027
12		Low	548	569	1117
13	Extracurricular_Activities	No	1282	1130	2412
14		Yes	2093	1494	3587
15	Sleep_Hours	High	1595	1247	2842
16		Medium	1051	825	1876
17		Extreme	566	434	1000
18		Low	163	118	281
19	Previous_Scores	High	2431	1593	4024
20		Medium	944	1031	1975
21	Motivation_Level	Low	906	872	1778
22		Medium	1753	1287	3040
23		High	716	465	1181
24	Internet_Access	Yes	3162	2388	5550
25		No	213	236	449
26	Tutoring_Sessions	Low	2604	2258	4862
27		Medium	682	339	1021
28		High	83	25	108
29		Extreme	6	2	8
30	Family_Income	Low	1264	1164	2428
31		Medium	1399	1026	2425
32		High	712	434	1146
33	Teacher_Quality	Low	314	295	609
34		Medium	1987	1647	3634
35		High	1074	682	1756
36	School_Type	Public	2339	1843	4182
37		Private	1036	781	1817
38	Peer_Influence	Positive	1439	940	2379
39		Negative	608	645	1253
40		Neutral	1328	1039	2367
41	Physical_Activity	Low	1064	854	1918
42		Medium	2098	1615	3713
43		High	213	155	368
44	Learning_Disabilities	No	3085	2281	5366
45		Yes	290	343	633
46	Parental_Education_Leve	HighSchool	1579	1409	2988
47		College	1031	782	1813
48		Postgraduate	765	433	1198
49	Distance_from_Home	Near	2102	1467	3569
50		Moderate	982	836	1818
51		Far	291	321	612
52	Gender	Male	1975	1485	3460
53		Female	1400	1139	2539

3.3 Perhitungan Probabilitas (Laplace Smoothing)

Dari frekuensi diperoleh probabilitas bersyarat menggunakan Laplace Smoothing untuk menghindari pembagian dengan nol.

Rumus Excel:
=(Frekuensi_Kelas!C2 + 1) / (SUM(Frekuensi_Kelas!C:C) + 3)

	A	B	C	D
1	Fitur	Nilai	P(High)	P(Medium)
2	Hours_Studied	High	0.028756797	0.014543781
3		Medium	0.019855511	0.032128353
4		Low	0.000940277	0.005541441
5	Attendance	Met	0.044866872	0.018744874
6		NotMet	0.008055038	0.033788785
7	Parental_Involvement	Low	0.008603532	0.013243443
8		Medium	0.027362054	0.025986757
9		High	0.014966072	0.015143937
10	Access_to_Resources	High	0.018366739	0.015323984
11		Medium	0.026390434	0.026906996
12		Low	0.008603532	0.011402965
13	Extracurricular_Activities	No	0.020106251	0.022625883
14		Yes	0.032815659	0.029907776
15	Sleep_Hours	High	0.025011362	0.024966491
16		Medium	0.016486186	0.016524296
17		Extreme	0.008885615	0.008702263
18		Low	0.00257009	0.002380619
19	Previous_Scores	High	0.038112551	0.031888291
20		Medium	0.014809359	0.020645368
21	Motivation_Level	Low	0.01421385	0.017464541
22		Medium	0.027487424	0.025766699
23		High	0.011236307	0.009322424
24	Internet_Access	Yes	0.049568256	0.047792426
25		No	0.003353654	0.004741233
26	Tutoring_Sessions	Low	0.040823682	0.04519175
27		Medium	0.010703484	0.006801768
28		High	0.001316387	0.000520135
29		Extreme	0.000109699	6.00156E-05
30	Family_Income	Low	0.019824168	0.02330606
31		Medium	0.021939791	0.020545342
32		High	0.011173622	0.008702263
33	Teacher_Quality	Low	0.004936453	0.00592154
34		Medium	0.031154503	0.032968572
35		High	0.016846625	0.013663553
36	School_Type	Public	0.036670793	0.036889591
37		Private	0.016251117	0.015644067
38	Peer_Influence	Positive	0.022566642	0.018824894
39		Negative	0.009543809	0.01292336
40		Neutral	0.02082713	0.020805409
41	Physical_Activity	Low	0.016689912	0.017104447
42		Medium	0.032894015	0.032328405
43		High	0.003353654	0.003120811
44	Learning_Disabilities	No	0.048361568	0.045651869
45		Yes	0.004560342	0.006881789
46	Parental_Education_Leve	HighSchool	0.024760621	0.028207334
47		College	0.01617276	0.015664073
48		Postgraduate	0.0120042	0.008682257
49	Distance_from_Home	Near	0.0329567	0.029367636
50		Moderate	0.015404867	0.016744354
51		Far	0.004576014	0.006441675
52	Gender	Male	0.030966448	0.029727729
53		Female	0.021955462	0.02280593

3.4 Menghitung Nilai Posterior dan Menentukan Prediksi Akhir

Menghitung probabilitas gabungan setiap kelas dengan mengalikan seluruh probabilitas fitur terhadap kelas dan kelas dengan nilai posterior tertinggi menjadi hasil prediksi akhir.

Rumus Excel:
=PRODUCT(C2:C20)

	A	B	C	D
1	**Feature**	**Nilai_x_test**	P(High)	P(Medium)
2	Hours_Studied	High	0.014543781	0
3	Attendance	Met	0.018744874	0
4	Parental_Involvement	Medium	0.032128353	0
5	Access_to_Resources	High	0.014543781	0
6	Extracurricular_Activities	Yes	0.029907776	0
7	Sleep_Hours	Medium	0.032128353	0
8	Previous_Scores	High	0.014543781	0
9	Motivation_Level	Medium	0.032128353	0
10	Internet_Access	Yes	0.029907776	0
11	Tutoring_Sessions	Low	0.005541441	0
12	Family_Income	Medium	0.032128353	0
13	Teacher_Quality	High	0.014543781	0
14	School_Type	Public	0.036889591	0
15	Peer_Influence	Positive	0.018824894	0
16	Physical_Activity	Medium	0.032128353	0
17	Learning_Disabilities	No	0.022625883	0
18	Parental_Education_Level	College	0.015664073	0
19	Distance_from_Home	Near	0.029367636	0
20	Gender	Male	0.029727729	0
21				
22		Posterior (P(Class|X))	3.05775E-32	0
23				
24		Prediksi Akhir	High	

BAB IV – IMPLEMENTASI PROGRAM

4.1 Implementasi di Python

File program: finalhasil.ipynb.

Python digunakan untuk melatih model Naive Bayes, menyimpan model ke file .pkl, dan melakukan prediksi otomatis untuk data baru.

```python
from sklearn.preprocessing import OrdinalEncoder
from sklearn.naive_bayes import CategoricalNB
from sklearn.pipeline import Pipeline
import numpy as np
import joblib

# ========================
# 1️⃣ Siapkan Encoder Aman
# ========================
class SafeOrdinalEncoder(OrdinalEncoder):
    """OrdinalEncoder yang otomatis ubah -1 menjadi max(category)+1"""
    def transform(self, X):
        X_enc = super().transform(X)
        # ubah -1 (kategori baru) jadi max+1 dari kolom masing-masing
        for i in range(X_enc.shape[1]):
            mask = X_enc[:, i] < 0
            if np.any(mask):
                max_val = np.nanmax(X_enc[:, i][~mask]) if np.any(~mask) else 0
                X_enc[:, i][mask] = max_val + 1
        return X_enc

encoder = SafeOrdinalEncoder(handle_unknown='use_encoded_value', unknown_value=-1)

# ========================
# 2️⃣ Buat Pipeline
# ========================
model = Pipeline([
    ('encoder', encoder),
    ('nb', CategoricalNB())
])

# ========================
# 3️⃣ Latih Model
# ========================
model.fit(X_train, y_train)

# ========================
# 4️⃣ Simpan
# ========================
joblib.dump(model, "naive_bayes_balanced_model.pkl")
print("✅ Model aman disimpan dan siap digunakan di Streamlit.")
```

4.2 Implementasi di Streamlit

File program: app.py.

Aplikasi Streamlit memudahkan pengguna memilih kategori setiap fitur dan menampilkan hasil prediksi Exam_Score secara interaktif.

BAB V – HASIL DAN PEMBAHASAN

Hasil perbandingan antara perhitungan manual, Python menunjukkan kesamaan hasil prediksi.

	A	B
1	Metode	Prediksi
2	Manual Excel	High
3	Program Python	High

BAB VI – KESIMPULAN

1. Algoritma Naive Bayes dapat diimplementasikan dengan baik baik secara manual maupun otomatis.
2. Proses perhitungan di Excel memberikan pemahaman mendalam tentang probabilitas dan smoothing.
3. Hasil dari Excel, Python, dan Streamlit konsisten dan akurat.
4. Implementasi UI Streamlit mempermudah pengguna awam melakukan prediksi.

LAMPIRAN

- StudentPerformanceFactors.xlsx (Dataset Original)
- Naive_Bayes_Manual_Excel.xlsx
- finalhasil.ipynb (Mengolah Dataset Original)
- dataset_final.xlsx (Data Yang Digunakan Untuk Prediksi Mode)
- naive_bayes_balanced_model.pkl (Model Categori Naïve Bayes)
- app.py (Streamlit)