بسم الله الرحمن الرحيم

گزارش پروژه Airline Satisfaction Prediction

استاد: میرزارضایی
دانشجو: سید محمدرضا هاشمی
درس: یادگیری ماشین
دی ماه 1404

مسئله پیش‌بینی رضایت مسافران خطوط هوایی

---

1. مقدمه (Introduction)

هدف این پروژه، **پیش‌بینی سطح رضایت مسافران خطوط هوایی** بر اساس ویژگی‌های جمعیتی، نوع سفر، کلاس پرواز و میزان رضایت از خدمات مختلف هواپیمایی است.
دیتاست شامل دو فایل `train.csv` و `test.csv` بوده و متغیر هدف `Satisfaction` دارای دو کلاس است:

- `satisfied`
- `neutral or dissatisfied`

این مسئله یک **مسئله طبقه‌بندی دودویی (Binary Classification)** با داده‌های عددی و دسته‌ای است.

---

2. شرح داده‌ها (Dataset Description)

تعداد نمونه‌ها:

Features	Samples	Dataset
22	103,904	Train
22	25,976	Test

متغیر هدف (Target Distribution)

```
neutral or dissatisfied: ٪56.67
satisfied: ٪43.33
```

⚠ داده‌ها **نامتوازن (imbalanced)** هستند، بنابراین معیار Accuracy به تنهایی مناسب نیست.

---

3. Task 1 – تحلیل اکتشافی داده‌ها (EDA)

3.1 مقادیر گمشده

تنها ستون دارای مقدار گمشده:

Missing	Feature
310	Arrival Delay in Minutes

نسبت داده‌های گمشده کم است و با **میانه** (**Median**) جایگزین شد.

3.2 تحلیل متغیر هدف

- توزیع کلاس‌ها نامتوازن است
- استفاده از معیارهایی مانند **F1-score** ضروری است

3.3 تحلیل ویژگی‌ها

- ویژگی‌های عددی (سن، تأخیر، مسافت پرواز) دارای **outlier** هستند
- نمودارهای Boxplot نشان دادند که تأخیر پرواز تأثیر مستقیم بر نارضایتی دارد
- همبستگی‌ها نشان دادند که:
  - `Inflight entertainment`
  - `Online boarding`
  - `Seat comfort`
- ارتباط مثبتی با رضایت دارند

---

4. Task 2 – پیش‌پردازش داده‌ها (Preprocessing)

مراحل انجام‌شده:

4.1 حذف ستون‌های غیرضروری

`Unnamed: 0, id`

4.2 کدگذاری متغیرهای دسته‌ای

- Binary Encoding:
  - Gender
  - Customer Type
  - Type of Travel
- Ordinal Encoding:
  - Eco > Eco Plus > Business

4.3 مدیریت داده‌های گمشده

- استفاده از **Median imputation** فقط بر اساس داده‌های train

4.4 نرمال‌سازی داده‌ها

- استفاده از `StandardScaler`
- ضروری برای مدل‌هایی مانند KNN

---

5. Task 3 – انتخاب ویژگی (Feature Selection)

5.1 Forward Selection

- مدل پایه: Logistic Regression
- معیار: F1-score
- تعداد ویژگی‌های منتخب: 10

F1 = 0.8482 ± 0.0035

5.2 Backward Elimination

- انتخاب 10 ویژگی

F1 = 0.8491 ± 0.0033

5.3 ویژگی‌های مشترک (8 ویژگی مهم)

```
Customer Type
Type of Travel
Class
Inflight wifi service
Online boarding
On-board service
Leg room service
Checkin service
```

PCA 5.4

- حفظ 95٪ واریانس
- 17 مؤلفه اصلی

F1 = 0.8397

- PCA نسبت به انتخاب ویژگی عملکرد ضعیف‌تری داشت (کاهش interpretability)

---

6. Task 4 & 5 – طبقه‌بندی با KNN و یافتن K بهینه

بهترین نتایج KNN

Test F1	Best K	Feature Set
0.9226	11	Common(8)
0.9233	9	Backward(10)

- مجموعه ویژگی backward عملکرد بهتری داشت.

---

7. Task 6 – درخت تصمیم (ID3 و C4.5)

بهترین مدل در این بخش:

ID3 با ویژگی‌های مشترک

F1	Features	Model
0.9315	Common	ID3
0.9161	Common	C4.5

- هرس (pruning) باعث کاهش بیش‌برازش شد اما دقت نهایی کاهش یافت.

---

8. Task 7 – روش‌های Ensemble

نتایج کلیدی

Test F1	Features	Model
0.9362	Backward	Random Forest
0.9346	Backward	Bagging DT
0.9257	Backward	Gradient Boosting
0.8927	Backward	AdaBoost

- **Random Forest** بهترین عملکرد را داشت

---

9. Task 8 – معیار ارزیابی مناسب

با توجه به:

- نامتوازن بودن داده‌ها
- اهمیت تشخیص درست مشتریان ناراضی

بهترین معیار:

☑ **F1-score**

دلایل:

- ترکیب Precision و Recall
- جلوگیری از گمراهی ناشی از Accuracy بالا
- مناسب برای کاربردهای تجاری و رضایت مشتری

---

10. Task 9 – بهبود عملکرد مدل (Performance Enhancement)

اقدامات انجام‌شده:

1. **تنظیم ابرپارامترها (GridSearchCV)**
   استفاده از `class_weight='balanced'`
   تنظیم آستانه تصمیم (Threshold tuning)

بهترین مدل نهایی:

- `RandomForest`
- `n_estimators=200`
- `min_samples_leaf=5`
- `class_weight=balanced`

نتیجه نهایی:

Value	Metric
0.9517	Accuracy
0.9446	F1-score

- این بهترین نتیجه کل پروژه است.

---

11. جمع‌بندی نهایی (Conclusion)

- مهم‌ترین ویژگی‌ها مربوط به **کیفیت خدمات داخل پرواز** هستند
- مدل‌های Ensemble به‌طور واضح بهتر از مدل‌های پایه عمل کردند
- انتخاب ویژگی + تنظیم آستانه → بهبود قابل توجه عملکرد
- **Random Forest تنظیم‌شده بهترین مدل نهایی پروژه است**