پیش‌بینی و ارزیابی مدل‌ها

امیر مهدی اعرابی | ۹۹۵۲۲۲۷۵

در این مرحله، داده‌های پاک‌سازی‌شده و آماده به‌کارگیری هستند. هدف این است که عملکرد مدل‌های مختلف یادگیری ماشین و یادگیری عمیق برای پیش‌بینی نتیجه دیابت ارزیابی شود و بهترین مدل انتخاب گردد.

---

2. آماده‌سازی داده‌ها

بارگذاری داده‌ها

داده‌ها از فایل `cleaned_diabetes_data.csv` بارگذاری شدند. ستون‌های ورودی به‌عنوان ویژگی‌ها (X) و ستون `Outcome` به‌عنوان متغیر هدف (y) تعریف شدند. داده‌ها به نسبت 70-30 برای مجموعه‌های آموزش و آزمایش تقسیم شدند.

پیش‌پردازش

- داده‌های غیرعددی (مثلاً دسته‌بندی‌های BMI یا سن) به مقادیر عددی با استفاده از **Label Encoding** تبدیل شدند.
- داده‌های فاقد مقدار (`NaN`) حذف شدند.
- مقیاس‌گذاری ویژگی‌ها (Scaling) برای مدل‌هایی مانند **Logistic Regression** و **SVM** انجام شد.

---

3. آموزش و ارزیابی مدل‌ها

مدل‌های ارزیابی‌شده

1. **Logistic Regression**
2. **Random Forest**
3. **(Support Vector Machine (SVM**
4. **(K-Nearest Neighbors (KNN**
5. **شبکه عصبی (مدل یادگیری عمیق)**

نتایج مدل‌ها

الف) مدل‌های کلاسیک

برای هر مدل، معیارهای زیر ارزیابی شدند:

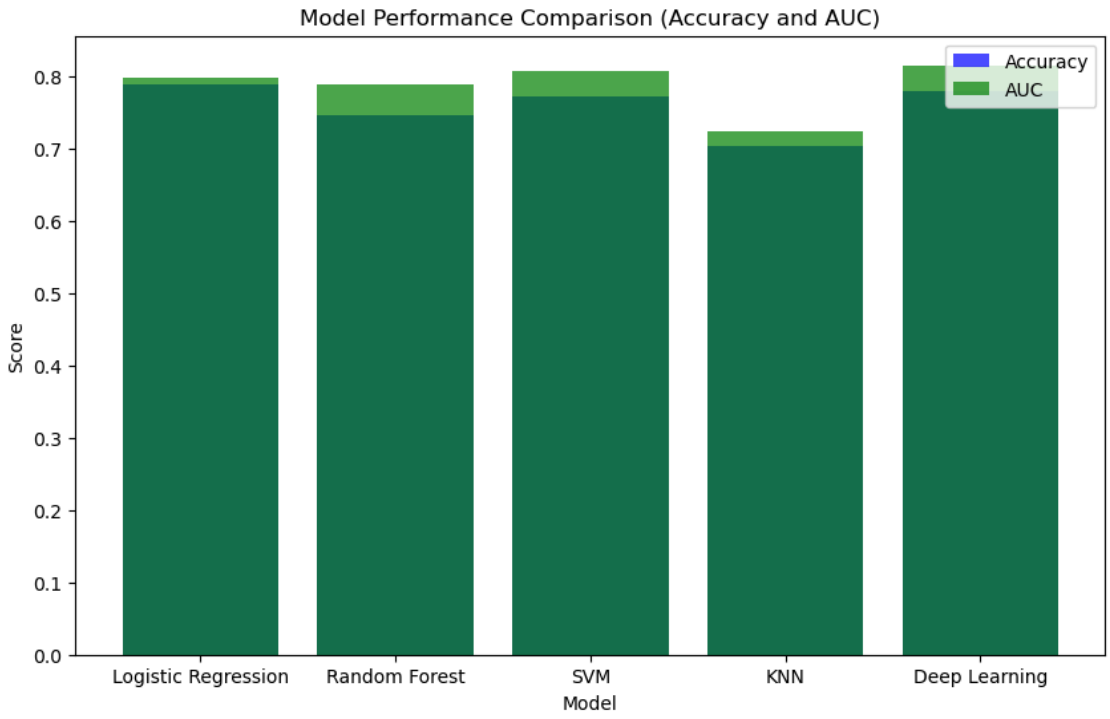
- **Accuracy:** دقت پیش‌بینی
- **AUC:** مساحت زیر منحنی ROC
- **Classification Report:** متریک‌های دقت، یادآوری و F1-Score
- **Confusion Matrix:** ماتریس اغتشاش برای ارزیابی جزئی عملکرد مدل.

نتایج اولیه:

مدل	Accuracy	AUC
Logistic Regression	0.79	0.80
Random Forest	0.75	0.79
SVM	0.77	0.81
KNN	0.70	0.72

ب) مدل یادگیری عمیق

- **ساختار مدل:**
  - **لایه ورودی:** 64 نرون با تابع فعال‌سازی ReLU.
  - **لایه میانی:** 32 نرون با تابع ReLU.
  - **لایه خروجی:** 1 نرون با تابع سیگموئید.
- **دقت مدل:**
  - Accuracy: 0.78
  - AUC: 0.81



---

4. بهینه‌سازی مدل‌ها (Hyperparameter Tuning)

با استفاده از **GridSearchCV**، مدل‌ها با مقادیر مختلف پارامترها بهینه‌سازی شدند. برخی از تغییرات قابل توجه:

- **Logistic Regression:** بهبود با تنظیم پارامترهای `C` و `solver`.
- **Random Forest:** افزایش تعداد درخت‌ها (n_estimators) و تنظیم عمق حداکثری (max_depth).
- **SVM:** آزمایش هسته‌های مختلف (`linear`, `rbf`) و مقدار `gamma`.

نتایج بهترین پارامترها برای هر مدل:

مدل	بهترین پارامترها
Logistic Regression	`'C=10, solver='liblinear`
Random Forest	`n_estimators=100, max_depth=None`
SVM	`'C=1, kernel='linear', gamma='scale`

---

5. ارزیابی مدل‌های ترکیبی (Ensemble Methods)

روش‌های ترکیبی بررسی‌شده:

1. **(Random Forest (Ensemble**
2. **Gradient Boosting**
3. **AdaBoost**

نتایج:

مدل	Accuracy	AUC
(Random Forest (Ensemble	0.75	0.79
Gradient Boosting	0.76	0.79
AdaBoost	0.73	0.76

---

6. انتخاب مدل نهایی

مدل **SVM** به‌عنوان بهترین مدل انتخاب شد:

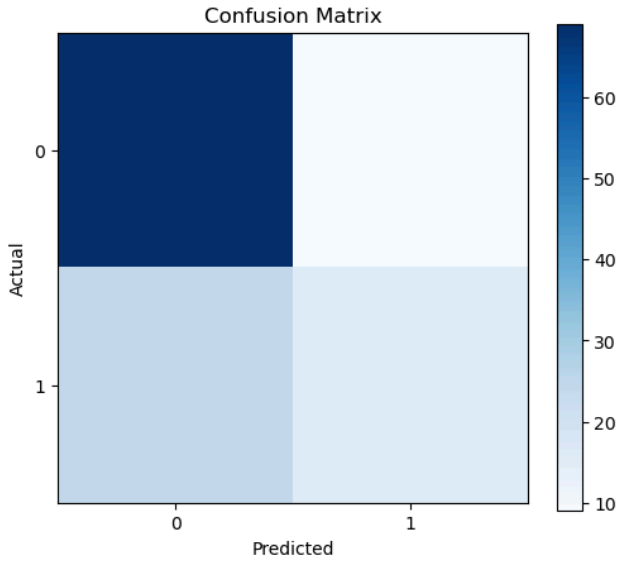
- Accuracy: 0.77
- AUC: 0.80

---

7. تجزیه و تحلیل عملکرد مدل

SVM با بهترین پارامترها:

- **ماتریس اغتشاش:**
نشان‌دهنده تعادل نسبتاً خوب میان پیش‌بینی‌های مثبت و منفی.



- **منحنی ROC:**
نشان‌دهنده تعادل میان نرخ مثبت‌های واقعی (TPR) و نرخ مثبت‌های کاذب (FPR)