アジア人のデータで学習した糖尿病予測モデルは、他人種にも同等の性能を発揮できるのかを検証した

はじめに

元看護師でデータサイエンスを勉強している。自分の知識が活かせる分野でデータ分析をしてみようと思いヘルスケアデータをkaggleで探した。

目次

---

1. 本記事の概要

いくつかの人種が含まれるヘルスケアデータを用いて、アジア人のデータで学習した機械学習モデルが他人種にも適用できるかを検証した。

- どんな人に読んで欲しいか

  データサイエンスを始めた方、看護師などの医療職、医療にデータサイエンスを使うことに興味がある方

- この記事に書くこと、わかること

  ヘルスケアデータセットから分析できること

  糖尿病リスク予測モデルをアジア人に対して構築し、それが他人種に適用できるか

  糖尿病患者を予測するツールとして次にどう展開できるか

- この記事で扱わないこと

  妊娠糖尿病、1型糖尿病に関すること

  糖尿病患者を予測するツールの開発

【実行環境】

本分析は Google Colaboratory 上で実施した。使用言語は Python 3.10 系である。

主要ライブラリとして pandas、numpy、scikit-learn、matplotlib、seaborn を使用した。学習および評価は CPU 環境で実行し、GPU は使用していない。再現性確保のため、乱数シードを固定した。

【使用したデータセット】

https://www.kaggle.com/datasets/alamshihab075/health-and-lifestyle-data-for-diabetes-prediction

---

2. 作成したプログラム

臨床で日本人の糖尿病患者の看護をしていたため、このデータセットで多人種間での糖尿病に関する分析をしようと思った。

妊娠糖尿病および1型糖尿病は、発生の機序が異なり、糖尿病の大部分は2型糖尿病である※1ため前処理の段階でデータセットから除外した。その後、全体データを数値カラムとカテゴリカルカラムでそれぞれ可視化した。いくつか疑問点があったためそれについて調査した。（表1）

疑問点	調べたこと	わかったこと
insulin_level=2と数値が低いデータが14000人ほどいて数が極端に多い ※単位はμU/mLだがこの表では省略	健康な人のインスリン分泌のタイミングと正常値 (健常人は空腹時2-10、食後30m～1hで30-60、食後2時間で10-30)	健康な人が10000人弱いて、その人たちのデータが占めているので数が多いと思われる
BMIの平均が25.6と日本の基準からすると適正数値を超える人が多く、肥満ぎみの人を多く集めているのかという点	海外のBMIの基準 (欧米やWHOの基準では「標準: 18.5～24.9」「過体重：25～29.9」「肥満：30以上」)	日本でも普通体重は18.5以上25未満であり、欧米・WHO基準に照らし合わせてみてもデータセットが肥満に偏っているとは言えない

表1. 可視化した時の疑問点

初期検証およびモデル設計を迅速に行うため、全体データ（約97,000件）から目的変数の分布を保持した層化サンプリングにより10,000件を抽出した。最終的な評価では、より大規模なデータを用いた再学習を行うことを想定していた。説明変数となるカラムを確認したところ、diabetes_risk_scoreは計算方法が不明であり、他の説明変数から派生している可能性があるため、本分析ではデータリークを防ぐ目的で説明変数から除外した。

サンプリングしデータを10,000に減らした後も、diagnosed_diabetesの1（診断済み）、0（未診断）の割合が変わっていないことを確認した。診断済み：未診断は全体データと同様の60：40であった。BMI、HbA1c、insulin_levelなど生理学的データでもっとも糖尿病の診断に直結する項目を可視化した。

人種はAsian、White、Hispanic、Black、Otherという5種類であった。Otherという人種に関しては混血や少数民族、不明や未回答などを含むと予測し医学的・疫学的な解釈が困難であるため、本分析の人種間比較の対象からは除外した。サンプル10,000件の内容（民族）と数は以下の図1の通りである。

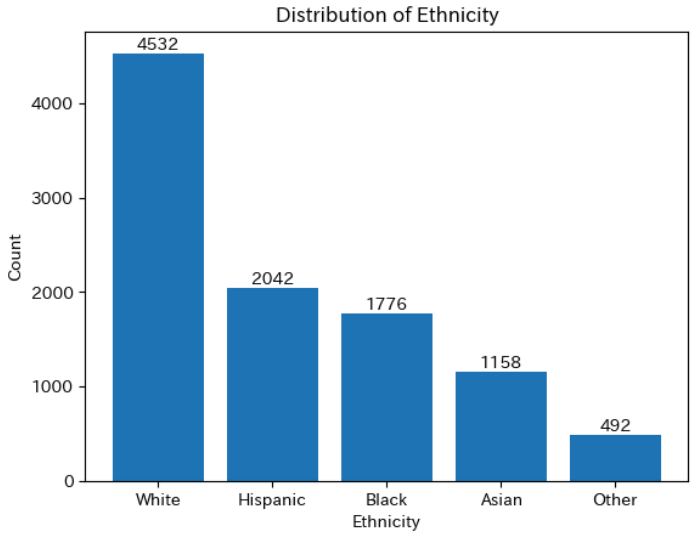

図1. サンプリング10,000件の人種分布

まずはAsianのデータを使ってモデルに学習させる際に、特徴量31カラムのうち何を選択するか検討した。income_level、educational_level、employment_statusは環境素因として糖尿病発症に関わってくるものであり、人種の身体的・生理学的な差というより社会経済差が現れるのではないかと考えたため除外した。またhypertension_history、cardiovascular_historyという既往歴から直接糖尿病を発症することはないため除外した。

その他、生活習慣に関する項目、脂質に関する採血結果や血圧などの生理学的項目などがあったが、ベースラインモデルを作るにあたって最低限の項目として以下の10項目を特徴量として採用した。糖尿病の病態と直接関連する生理学的指標および生活習慣指標に限定した。

英語タイトル (Original)	日本語訳
Age	年齢
gender	性別

ethnicity	民族
education_level	教育水準
income_level	所得水準
employment_status	雇用形態
smoking_status	喫煙状況
alcohol_consumption_per_week	週あたりのアルコール摂取量
physical_activity_minutes_per_week	週あたりの運動時間（分）
diet_score	食事スコア
sleep_hours_per_day	1日あたりの睡眠時間
screen_time_hours_per_day	1日あたりのスクリーンタイム（テレビ・スマホ等）
family_history_diabetes	糖尿病の家族歴
hypertension_history	高血圧の既往歴
cardiovascular_history	心血管疾患の既往歴
bmi	BMI（体格指数）
waist_to_hip_ratio	ウエスト・ヒップ比
systolic_bp	収縮期血圧（最高血圧）
diastolic_bp	拡張期血圧（最低血圧）
heart_rate	心拍数
cholesterol_total	総コレステロール
hdl_cholesterol	HDLコレステロール（善玉）
ldl_cholesterol	LDLコレステロール（悪玉）
triglycerides	中性脂肪
glucose_fasting	空腹時血糖値
glucose_postprandial	食後血糖値
insulin_level	インスリン濃度
hba1c	ヘモグロビンA1c
diabetes_risk_score	糖尿病リスクスコア
diabetes_stage	糖尿病のステージ（タイプ等）
diagnosed_diabetes	糖尿病の診断有無（0=なし、1=あり）

図2. 全特徴量

Age	**bmi**

gender	**waist_to_hip_ratio**
physical_activity_minutes_per_week	**glucose_fasting**
diet_score	**insulin_level**
sleep_hours_per_day	**hba1c**

図3. ベースラインモデルに使用した特徴量10個

Asianのデータのみを用いて、数値特徴量は標準化、カテゴリカル特徴量はOne-Hot Encodingをおこなった上でロジスティック回帰モデルを構築した（これをアジア人モデルと呼ぶ）。今回は、解釈性および再現性が高いためロジスティック回帰モデルを採用した。ロジスティック回帰は医療分野において広く用いられており、各特徴量が予測に与える影響を定量的に解釈できる点から、公平性の検証に適していると判断した。

```
# 学習に使うカラム（説明変数）を指定
# ひとまずベースラインモデルなので最低限の項目とする
feature_cols = [
   "Age",
   "gender",
   "bmi",
   "waist_to_hip_ratio",
   "glucose_fasting",
   "hba1c",
   "insulin_level",
   "diet_score",
   "physical_activity_minutes_per_week",
   "sleep_hours_per_day"
]
# asianを分割
df_asian = df_eval[df_eval["ethnicity"] == "Asian"]

X = df_asian[feature_cols]
y = df_asian["diagnosed_diabetes"]

X_train, X_test, y_train, y_test = train_test_split(
   X, y,
   test_size=0.2,
   stratify=y,
   random_state=42
)
```

```
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression

# 数値の説明変数とカテゴリカル変数をそれぞれ分ける
numeric_features = [
    "Age",
    "bmi",
    "waist_to_hip_ratio",
    "glucose_fasting",
    "hba1c",
    "insulin_level",
    "diet_score",
    "physical_activity_minutes_per_week",
    "sleep_hours_per_day"
]

categorical_features = [
    "gender"
]
```

```
# 前処理（数値の特徴量に正規化、カテゴリカル変数にOneHotEncoding）＋モデル

preprocess = ColumnTransformer(
    transformers=[
        ("num", StandardScaler(), numeric_features),
        ("cat", OneHotEncoder(drop="first"), categorical_features)
    ]
)
model = Pipeline(
    steps=[
        ("preprocess", preprocess),
        ("classifier", LogisticRegression(
            max_iter=1000,
            class_weight="balanced",
            random_state=42
        ))
    ]
)
```

```
# Pipelineを使うと、このfitだけで前処理のfit、transform、モデル学習のfitを順番に行う
# asianだけのモデルが構築できた
model.fit(X_train, y_train)
```

Asianのテストデータに対して予測をおこなった精度が以下である。

```
# asianモデルの精度を評価する
from sklearn.metrics import classification_report, roc_auc_score

y_pred = model.predict(X_test)
y_prob = model.predict_proba(X_test)[:, 1]

print(classification_report(y_test, y_pred))
print("ROC-AUC:", roc_auc_score(y_test, y_prob))
```

	precision	recall	f1-score	support
0	0.8	0.88	0.84	92
1	0.92	0.86	0.89	140
accuracy			0.87	232
ROC-AUC:	0.913431677			

表2. サンプリングデータのアジア人テストデータに対する予測精度

糖尿病ありクラスに対して、再現率 0.86、適合率 0.92 と高い値を示しており、見逃しを抑えつつ過剰な誤検出も少ない、バランスの取れたモデルであることが確認された。ROC-AUC は 0.913 を示し、糖尿病あり・なしを高い精度で識別できていることが確認された。

次に、このアジア人モデルを他人種のデータに用いて予測をおこなった。White、Black、Hispanicに分けて実施し、予測結果は以下である。

```
# asian以外の人種をdf_non_asiaとする
df_non_asian = df_eval[df_eval["ethnicity"] != "Asian"]
# 人種ごとに分ける
ethnicities = ["White", "Black", "Hispanic"]
# 人種ごとに分けてモデルで予測し、それぞれの精度を確かめる
from sklearn.metrics import classification_report, roc_auc_score
```

```
for eth in ethnicities:
   print(f"\n===== {eth} =====")

   df_eth = df_non_asian[df_non_asian["ethnicity"] == eth]

   X_eth = df_eth[feature_cols]
   y_eth = df_eth["diagnosed_diabetes"]

   y_pred = model.predict(X_eth)
   y_prob = model.predict_proba(X_eth)[:, 1]

   print(classification_report(y_eth, y_pred))
   print("ROC-AUC:", roc_auc_score(y_eth, y_prob))
```

White

	precision	recall	f1-score	support
0	0.82	0.91	0.87	1767
1	0.94	0.88	0.91	2765
accuracy			0.89	4532
ROC-AUC:	0.9334092684			

Black

	precision	recall	f1-score	support
0	0.82	0.91	0.87	702
1	0.94	0.87	0.9	1074
accuracy			0.89	1776
ROC-AUC:	0.927666099			

Hispanic

	precision	recall	f1-score	support
0	0.85	0.91	0.88	885
1	0.93	0.87	0.9	1157
accuracy			0.89	2042

ROC-AUC:	0.9359936325			

表3.サンプリングデータのアジア人以外のデータに対する予測精度

本分析では、アジア人のデータにおいて他人種と比較して再現率および ROC-AUC が僅かに低い傾向が見られた。この結果は、アジア人における糖尿病の病態的特徴（欧米人よりBMIが低くても糖尿病になる）※2※3により、リスクが段階的ではなく連続的に変化している可能性が示唆される。また、判定閾値の最適化が不十分である可能性を示唆する。ただし、サンプルサイズの差も影響している可能性があり、さらなる検証が必要である。

次に、全データ（から妊娠糖尿病と1型糖尿病を除外したもの）で同様の機械学習モデルを用いて検証した。コードは省き結果だけ記載する。

	precision	recall	f1-score	support
0	0.83	0.89	0.86	916
1	0.92	0.88	0.9	1382
accuracy			0.89	2298
ROC-AUC:	0.9345167753			

表4. 全データのうちアジア人テストデータに対する予測精度

White

	precision	recall	f1-score	support
0	0.83	0.89	0.86	17392
1	0.93	0.88	0.9	26216
accuracy			0.88	43608
ROC-AUC:	0.9333101293			

Black

	precision	recall	f1-score	support
0	0.83	0.9	0.86	6949
1	0.93	0.88	0.9	10490
accuracy			0.89	17439
ROC-AUC:	0.9362897954			

Hispanic

	precision	recall	f1-score	support
0	0.83	0.9	0.86	7908
1	0.93	0.87	0.9	11578
accuracy			0.88	19486
ROC-AUC:	0.9316559702			

表5. 全データのうちアジア人以外のデータに対する予測精度

アジア人と、アジア人以外の人種ではモデルの性能の差は誤差の範囲内であった（0.01）。また、サンプルデータで観察されたアジア人における性能の僅かな低下は、全データを用いた検証では再現されなかった。全件データにおいては、再現率および ROC-AUC ともに高い値を示しており、アジア人に対する糖尿病予測が特別に困難であるとは言えない結果となった。

ハイパーパラメータ（ C,penalty,solver）の調整を行った。C: [0.01, 0.1, 1, 10]、solver: ["lbfgs", "liblinear"]、penaltyは”l2”を用いて探索を行い、C:10、solver:”lbfgs”、penalty:”l2”と決定し、そのパラメータでアジア人データ、その他の人種データを予測、評価をおこなったが、主要な評価指標（Recall、ROC-AUC）に大きな変化は見られなかった。

また、生活習慣指標の一つとして screen_time_hours_per_day や alcohol_consumption_per_weekを追加したモデルも検討したが、主要な評価指標（Recall、Precision、ROC-AUC）に顕著な改善は見られなかった。そのため、解釈性と簡潔性を重視し、ベースラインの特徴量構成を最終モデルとして採用した。

---

3. 考察・反省

アジア人データのみで学習したモデルは、本データセットにおいて、他人種に対しても同等の予測性能を示した。本データセットにおいては、アジア人で学習したモデルが他人種に対しても汎用的に適用可能であることが示唆された。

また、医療においては、診断のために閾値を設定することが一般的であるが、実際にはその閾値によって明確に区分できない症例が一定数存在する。本分析においても、糖尿病リスクが連続的に分布している可能性が示唆されており、単一の閾値による二値分類では境界付近の症例を完全に捉えることは困難である。このような特性は、特定の人種に限らず、糖尿病という疾患そのものの病態的特徴を反映していると考えられる。

---

4. 今後の展開と活用

本モデルは、血液検査項目を含む基本的な健康指標および生活習慣情報を入力することで、糖尿病リスクを推定できる点に特徴がある。今後は本モデルを基盤として、患者自身が医療機関で得られた検査結果や日常的な生活習慣情報を入力し、自らのリスクを振り返るための支援ツールとしての活用が考えられる。例えば、Webアプリケーションやモバイルアプリとして実装することで、定期健診後のセルフチェックや、生活習慣改善への動機付けにつなげることが期待される。なお、本ツールは医師による診断を代替するものではなく、あくまで受診や生活改善を促す補助的手段として位置付けることが重要である。

本分析を通じて、データサイエンスが、病院を受診する前の段階において、受診を促したり生活行動を見直すきっかけとなり得る根拠を示すことができると分かった。特に、複数の健康指標や生活習慣データを統合的に扱うことで、単一の数値では捉えきれないリスクを可視化できる点に、データサイエンスの有用性があると考えられる。今後は、こうした分析手法を活かし、健診結果や日常的な生活習慣情報をもとに、個人が自身の健康状態を振り返りやすくする支援ツールの開発や、早期受診・行動変容を後押しする仕組みづくり等にデータサイエンスを活用していきたい。

おわりに

データ分析はモデル構築や精度の評価だけでなく、出てきた数値をどう捉えるのかという部分がさらに難しいと感じた。また、今回のデータセットは欠損値がなかったり、カテゴリカル変数が少なかったため前処理が少なくて済んだ。もっと前処理が煩雑だと予想していたので、他のデータセットなどで前処理をする練習をした方が自信がつくと思う。

出典

※1:https://dmic.jihs.go.jp/general/about-dm/010/010/01.html,国立健康危機管理研究機構糖尿病センター,糖尿病とは,2025-1-12検索.

※2:https://dm-net.co.jp/calendar/2024/038228.php,糖尿病ネットワーク,日本人はやせていても糖尿病に「肥満」と「やせ」の両方が課題に「やせることだけが健康的」は誤解,2025-1-12検索.

※3:https://dm-rg.net/news/3b00359e-e45f-46e4-8631-8eb6211c5fda,糖尿病リソースガイド,アジア人は白人よりも低体重で糖尿病を発症　BMIだけでは糖代謝異常の多くが見逃される,2025-1-12検索.