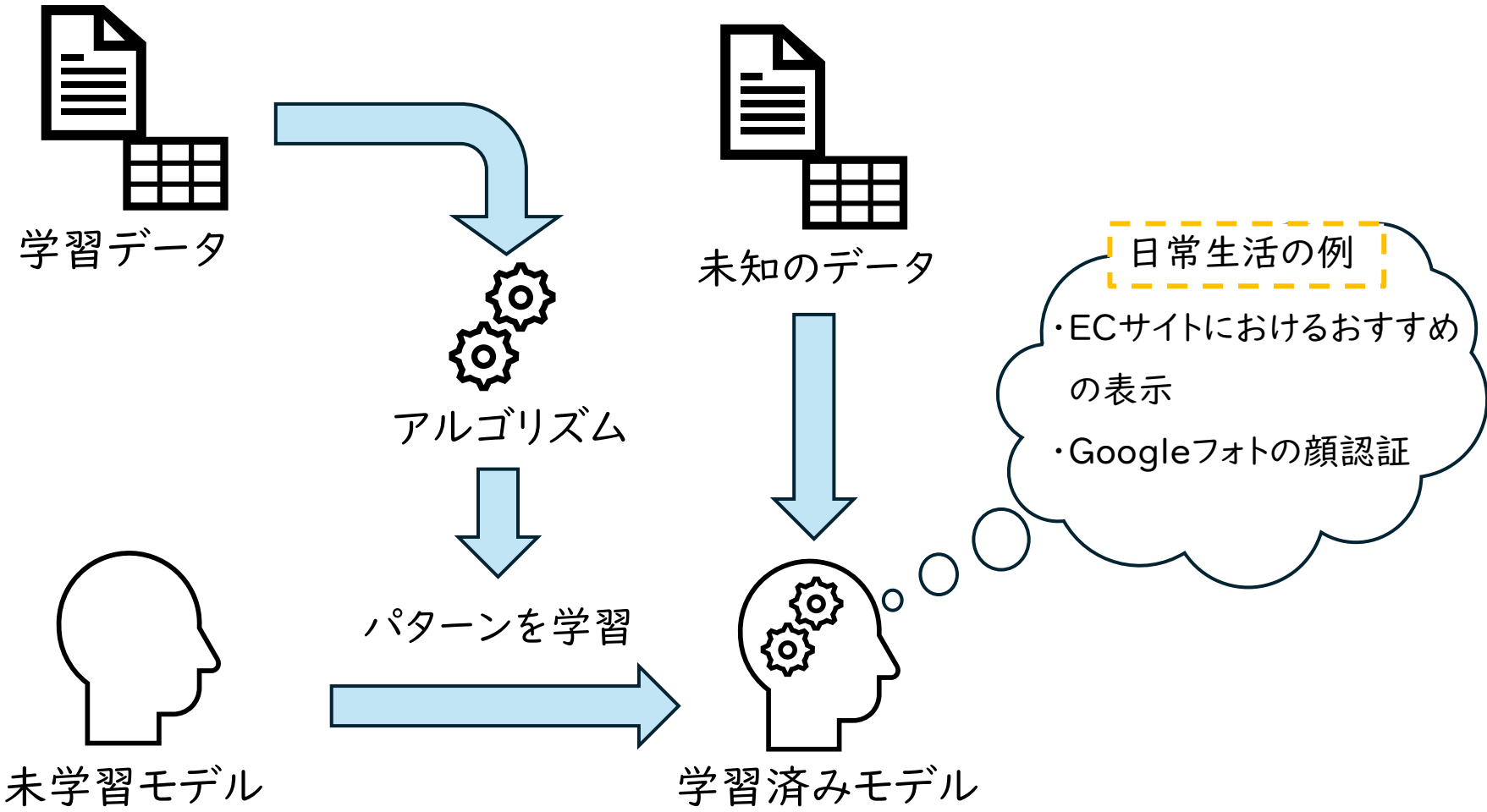
公衆衛生分野における機械学習活用の可能性

～自殺リスク予測モデルを参考に～

中川優馬、近藤文乃、持永展孝
宮崎県福祉保健課

はじめに(機械学習とは)

大量のデータの中からパターンや法則を見つけて、予測や判断を自動化する技術



はじめに(諸外国における機械学習を用いた自殺リスク予測)

	対象者	機械学習に用いたデータ
デンマーク（2020）	自殺により死亡した14,103人 自殺と無関係な265,183人	家族構成、自殺企図、収入、教育歴、就労状況、精神病既往、飲酒歴、処方歴
韓国（2019）	大うつ病性障害と診断されている自殺企図者56人 大うつ病性障害と診断されているが自殺企図のない39人 健常者87人	自殺企図・病歴に関するアンケート、ハミルトン評価尺度、Scale for Suicide Ideation、血中メチローム及びトランスクリプトーム
アメリカ（2019）	自殺企図のある入院患者27人 自殺企図のない入院患者46人	入院患者へのアンケート，臨床医の経過記録，家族，治療チームの他のメンバー，学校，地域サービス機関との連携に関する電子健康記録の情報
イギリス（2018）	自殺直前の2か月間に書かれた46の手紙や日記等の文章 他の期間からランダムに選択された54の文章	ヴァージニアウルフによって書かれた手紙・日記について、自殺直前の2か月間のものと、他の期間からのもの
アメリカ（2017）	自殺企図者130人 自殺企図のない精神病患者126人 健常者123人	音声を記録しながら 自由回答形式の半構造化インタビューを実施し、発言内容の単語・単語ペアを文字データとして抽出
アメリカ（2017）	希死念慮のある参加者17人 希死念慮のない参加者17人	参加者に対して、deathやpraiseなどの死および生に関する30種類の単語をランダムに提示した際に脳のどこが活動するかを機能的磁気共鳴画像法によるスキャンによって分析
中国（2019）	ソーシャルメディア上の投稿27,007件 抽出されたインターネットユーザー12,486人	ソーシャルメディア上の投稿に対し，死，自殺，これから，などの自殺のリスクが高いと専門家が分析した単語

「機械学習を用いた自殺予防の可能性とその問題点」 木下翔太郎，日本生物学的精神医学会誌31（3）：141-146,.2020より抜粋

方法(学習に用いたデータ)

- 平成29年から令和5年までの間に県警本部から情報提供のあった自殺未遂者410名

説明変数(特徴量)

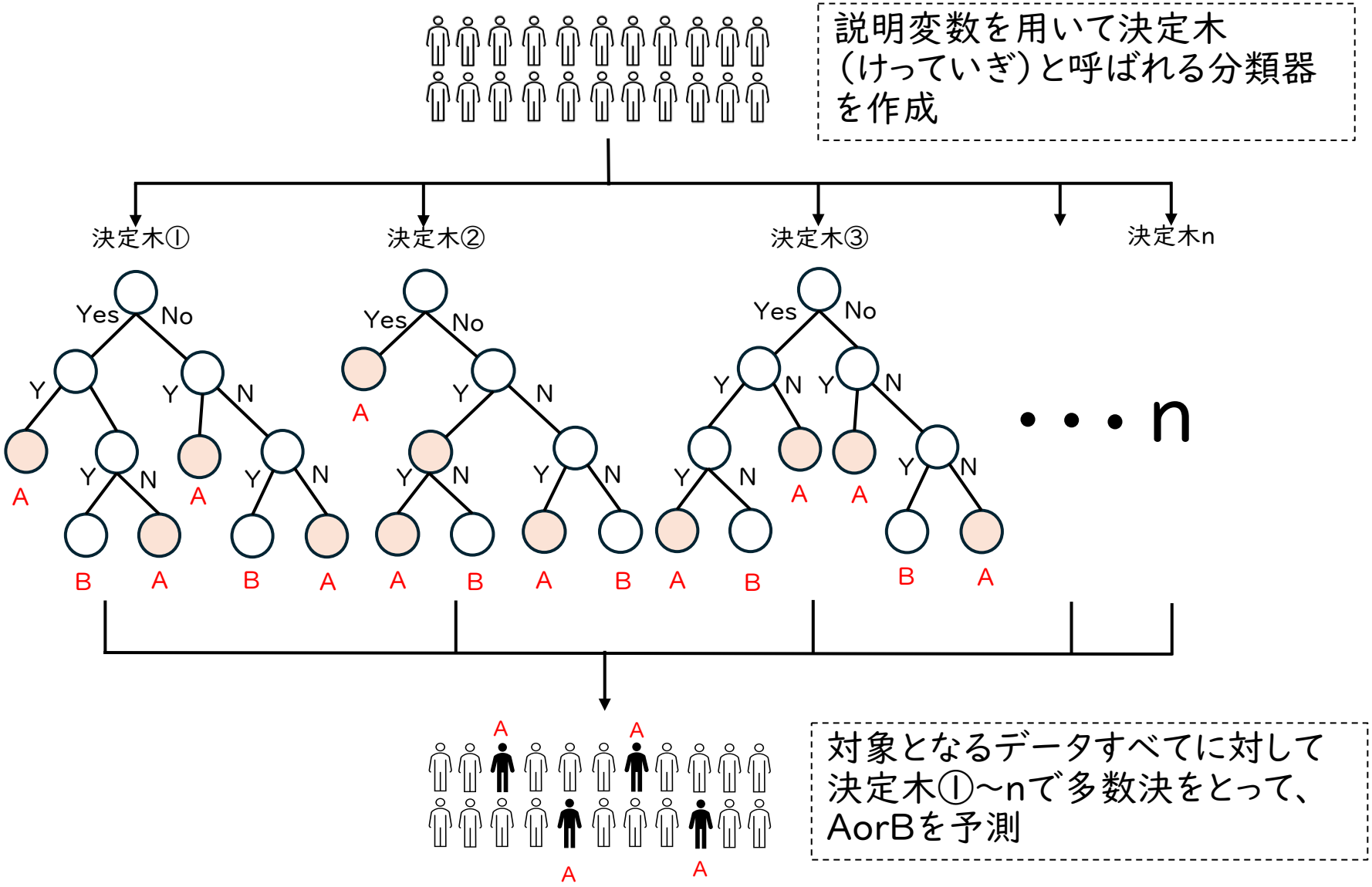
・年齢　・市町村　・性別　・自殺未遂時の手段
・自殺未遂に至った要因　・要因数
・情報提供書への「うつ病」の記載の有無
・職業(有職・学生生徒等・無職)

目的変数

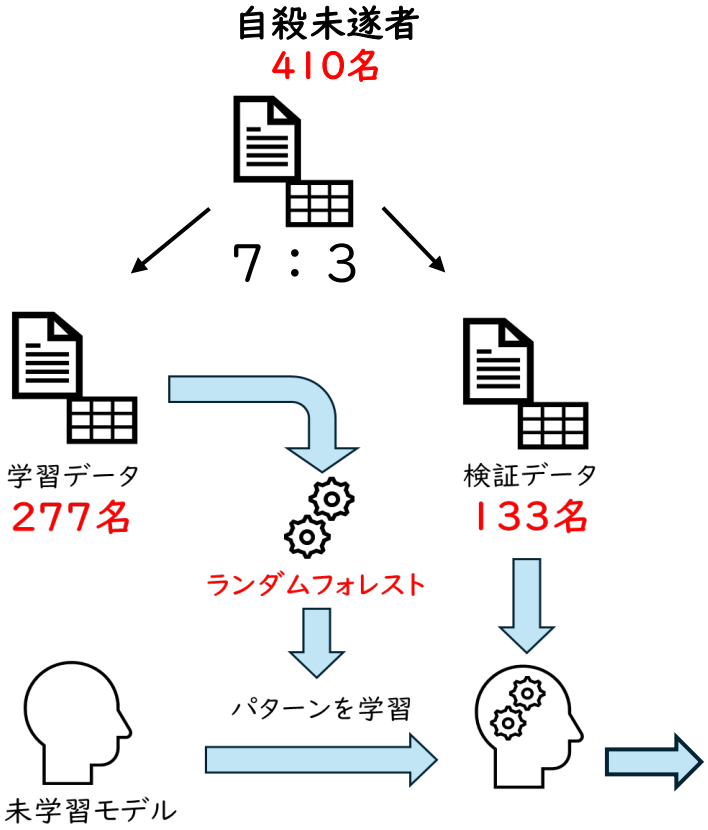
・平成31年～令和5年までの人口動態統計の死亡個票のうち死因が「自殺」である死亡個票と一致した者(10名)を【既遂】、それ以外を【生存】として正解ラベルを付与

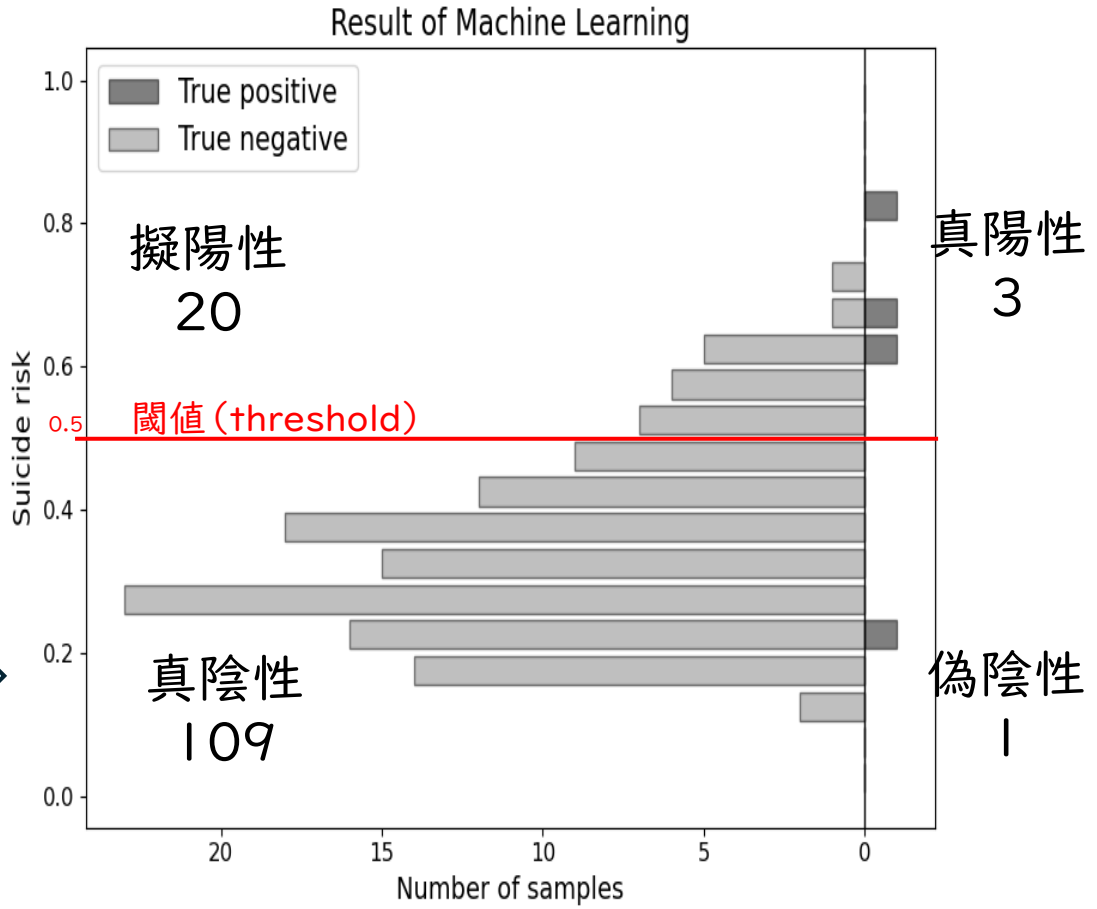
方法(学習に用いたモデル)

ランダムフォレストのイメージ



結果（ホールドアウト検証）



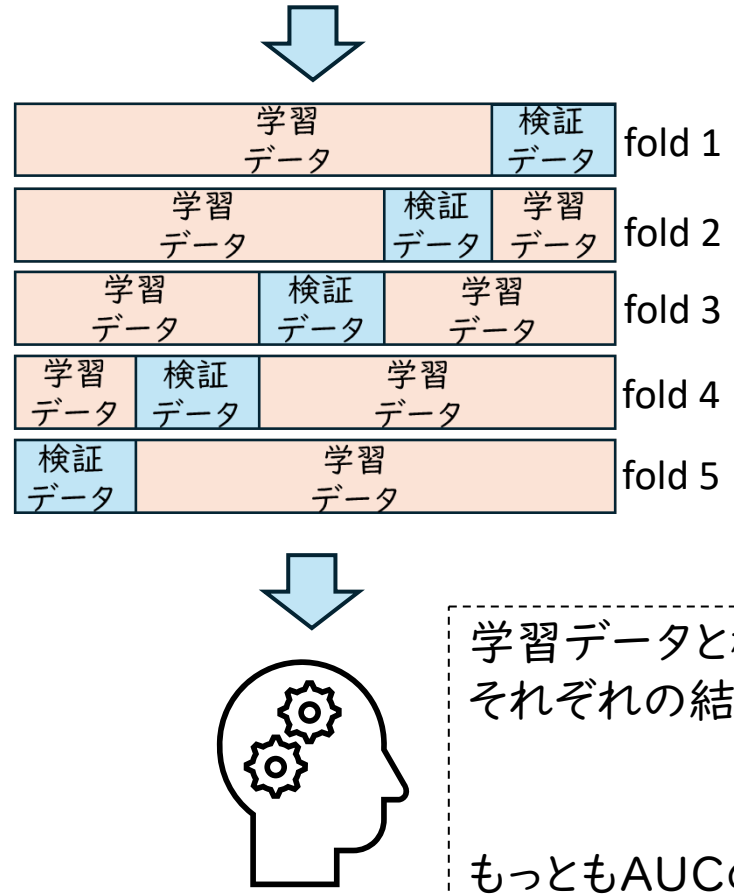


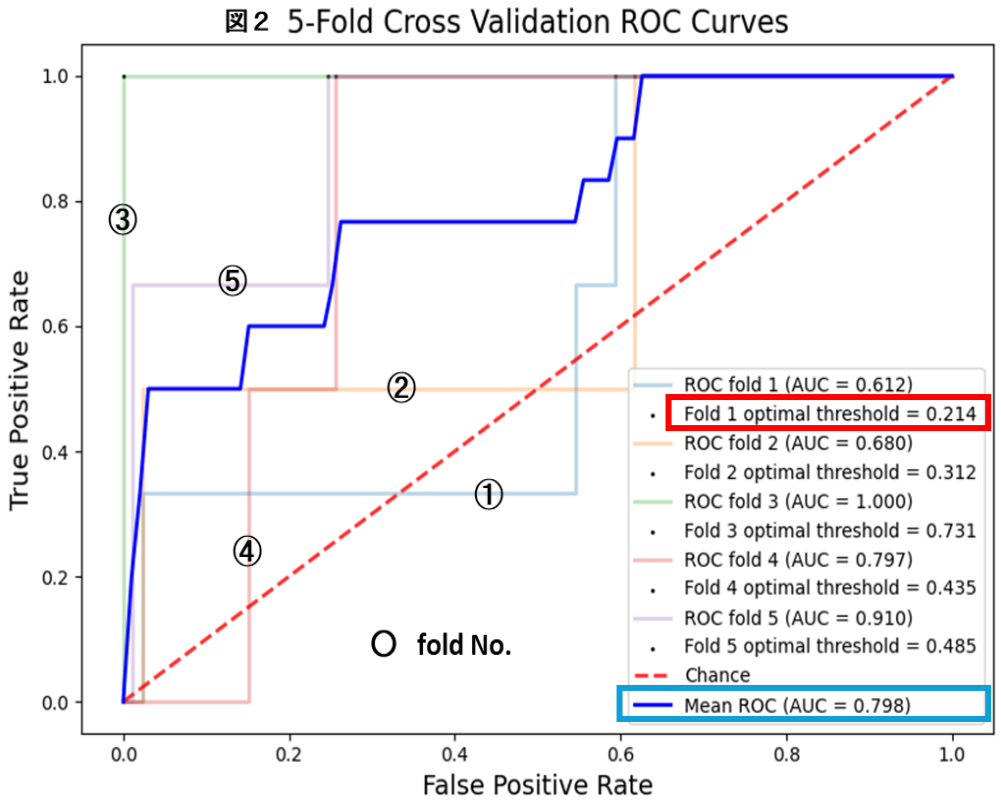
AUC（Area Under the Curve）※　＝ 0.80

結果（交差検証）

層状K分割法

自殺未遂者 410名



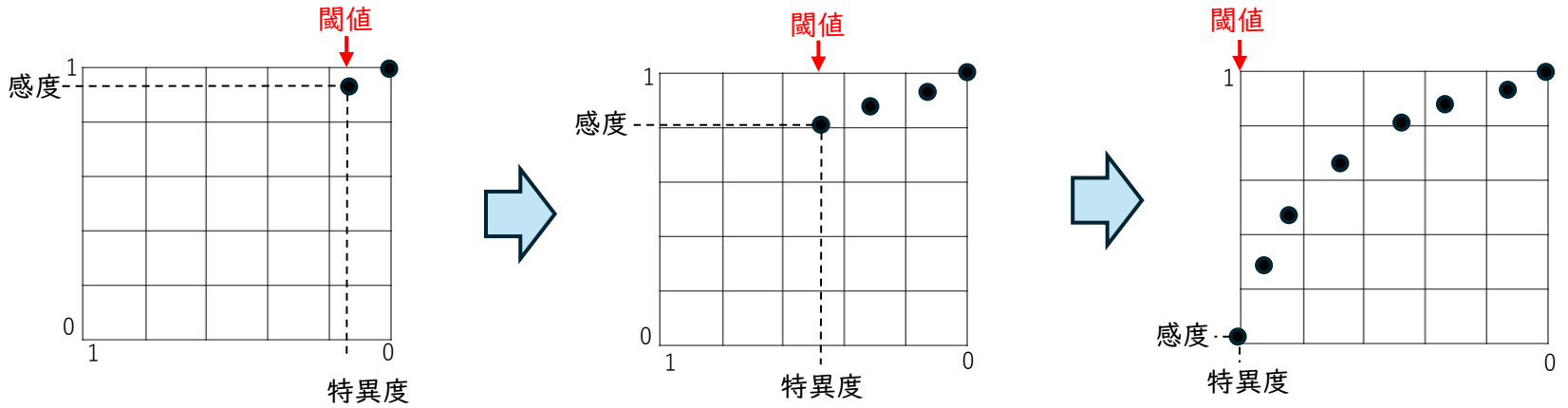


学習データと検証データを5つのパターン（hold out）に分け、
それぞれの結果を平均することでモデル全体の性能を評価
→平均AUC（Area Under the Curve）は、0.798

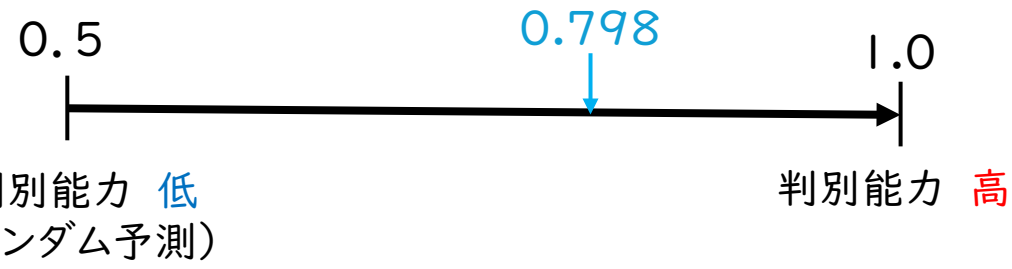
もっともAUCの低いfold 1 において、「既遂を必ず検出し、偽陽性をできるだけ少なくするような閾値（optimal threshold）」は、0.214であった。

考察（モデルの解釈）

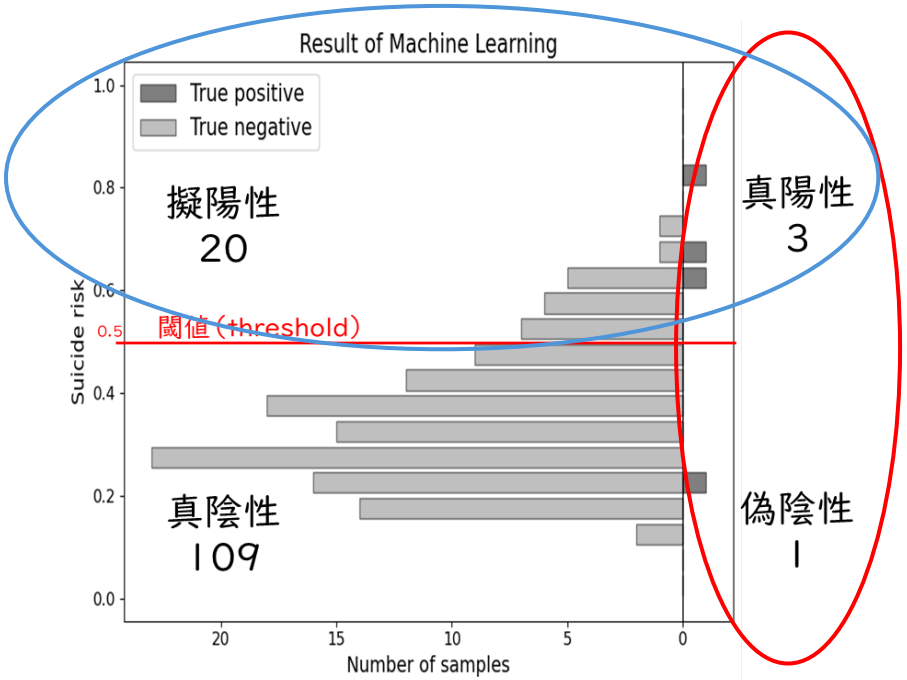
AUC（Area Under the Curve）とは「予測モデルの正確さ」を表す指標の一つ
ある閾値に対する感度、特異度をプロットしていった曲線（ROC曲線）の下部分の面積を指す



AUCの評価・・・
0.5～1.0の間をとり、1.0に近づくほど高い予測性能を持つ



考察（モデルの解釈）



※検証データ133件に対する結果

$$\text{陽性反応的中度} = \frac{\text{真陽性}}{\text{真陽性}+\text{偽陽性}} = \frac{3}{23} = 0.13$$

➢モデルが自殺リスクが高いと判断したケースの内、どれだけ正しく正例【既遂】を検出できたか。

→13%

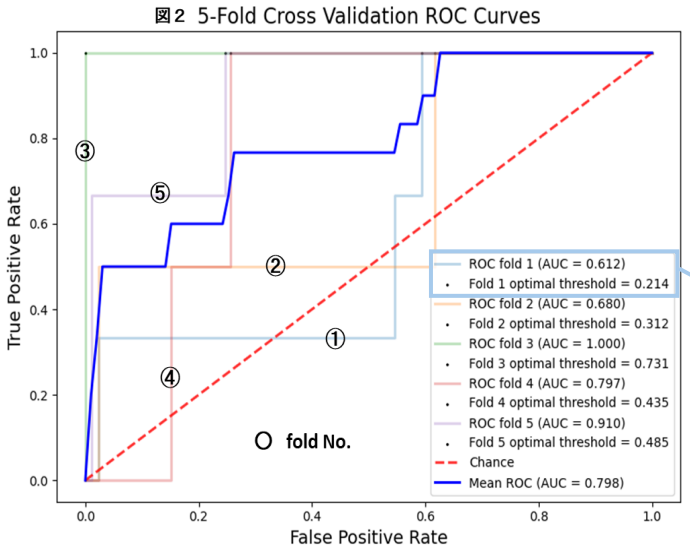
$$\text{感　度} = \frac{\text{真陽性}}{\text{真陽性}+\text{偽陰性}} = \frac{3}{4} = 0.75$$

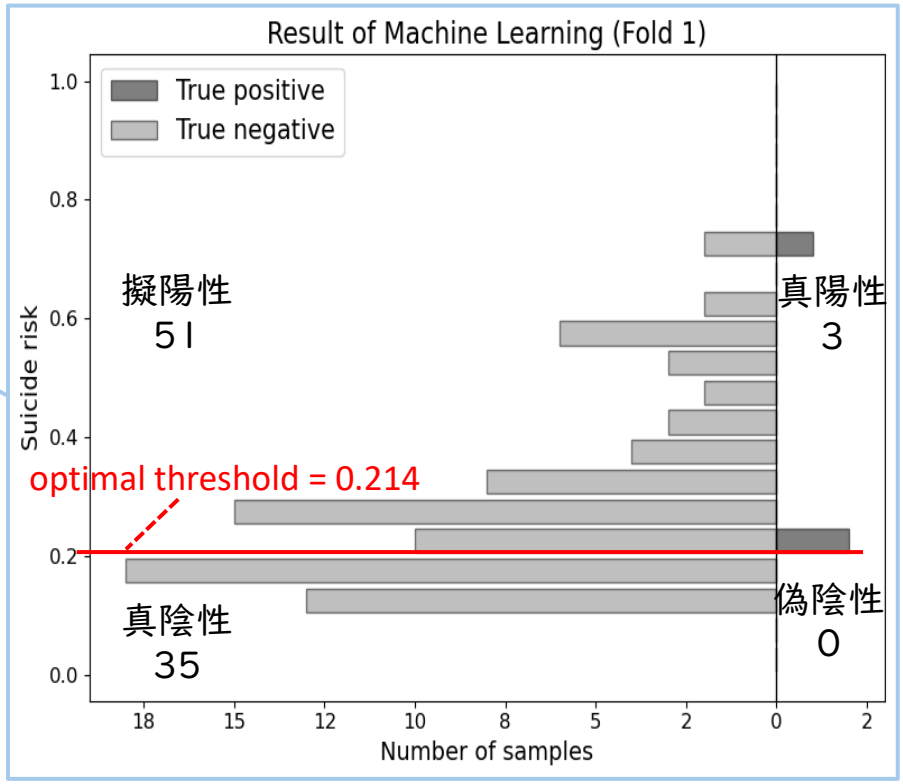
➢正例【既遂】のうち、モデルがどれだけ自殺リスクが高いと判断できたか。

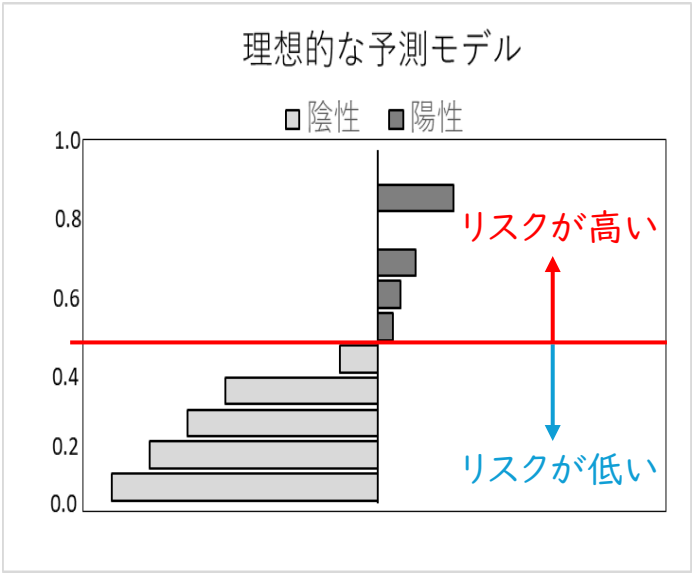
→75%

予測精度（陽性反応的中度）は、学習データの量に比例して向上。

考察（閾値の設定）







人命に関わる予測については、できるだけ正例（今回で言えば「既遂」）を拾うように設定が必要

予測精度が上がれば、より高い閾値での予測が可能

考察（実務への適用と公衆衛生分野への活用）

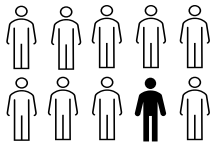
例えば・・・

・通常であれば、月１回の電話連絡
→2週間に１回に変更（徐々に介入頻度を減らしていく）

・通常であれば、半年間のフォロー
→１年間フォロー

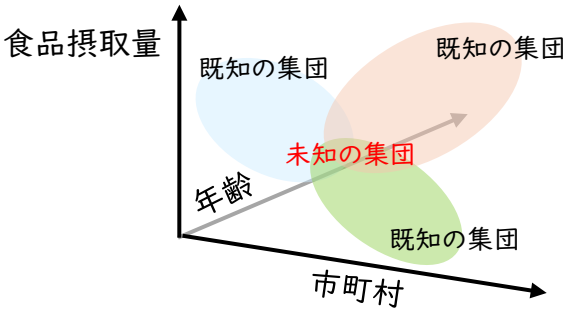
結核治療終了者

・発病年齢
・病型（I,r,b, Ⅰ, Ⅱ, Ⅲ）
・現病歴
・治療状況（完了、脱落、12か月以上）
・最大ガフキー号数
・薬剤耐性の有無



管理期間中や終了後の保健指導への活用

県民健康・栄養調査



新しい傾向を持つ集団の発見
新規事業への活用

まとめ

・自殺未遂者410名の情報を用いて自殺の予測モデルの構築を試みた

・モデル全体の性能としては、中程度（AUC=0.798）であったが、
実用にあたっては、より多くのデータの蓄積が必要

・公衆衛生分野において、機械学習を活用していくことで、
これまでにないアプローチでの保健指導や施策の展開が可能

・今後、活用が加速していく分野であり、
行政職員も最低限の理解と知識を持っておく必要がある。

※機械学習による予測を活かしつつ、最終的な判断は、人間が行うことが重要

（参考　今回使用したPythonコード）

```
import numpy as np
import pandas as pd
import sklearn
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score, accuracy_score, recall_score, precision_score, f1_score, confusion_matrix
from sklearn.model_selection import cross_val_score
from imblearn.ensemble import BalancedBaggingClassifier

df = pd.DataFrame(xl("'自殺企図者（県警）(python)'!A1:CC442", headers=True))
test = pd.DataFrame(xl("Sheet5!A1:Y2", headers=True))
df = df[['年齢','市町村','性別','首つり','有機溶剤吸入','服毒（医薬品）','服毒（医薬品以外）','練炭等','排ガス','その他のガス','刃物',
         '入水','飛び降り','飛び込み','その他（手段）','家庭問題','健康問題','経済・生活問題','勤務問題','交際問題','学校問題',
         'その他（要因）','要因数','うつ病','有職・無職','既遂']]
X = df.drop('既遂', axis=1)
y = df['既遂']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=100, stratify=y)

base_estimator = RandomForestClassifier(n_estimators=100,max_depth=8,random_state=42,class_weight='balanced')
model=BalancedBaggingClassifier(estimator=base_estimator,n_estimators=10,sampling_strategy='auto',replacement=False,
                                random_state=42)
model.fit(X_train, y_train)

train_cols = X.columns
test = test[train_cols]
pred_proba = model.predict_proba(test)[:, 1]
print("Test set prediction probabilities:", pred_proba)
```

年齢	市町村	性別	首つり	有機溶剤吸入	服毒（医薬品）	服毒（医薬品以外）	練炭等	排ガス	その他のガス	刃物	入水	飛び降り	飛び込み	その他（手段）	家庭問題	健康問題	経済・生活問題	勤務問題	交際問題	学校問題	その他（要因）	要因数	うつ病	有職・無職
60	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	3

ndarray
1行×1列
0.420666666666667