DEIM2020 I2-2(day1 p58)

健康診断データを用いた疾患予測における解釈可能なモデルの構築

大場　勇貴† 　手塚　太郎†† 　讃岐　勝††† 　我妻ゆき子†††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
††† 筑波大学 医学医療系　〒 305–8577 茨城県つくば市天王台 1–1–1
E-mail: †s1611488@klis.tsukuba.ac.jp, ††tezuka@slis.tsukuba.ac.jp, †††{sanuki,ywagats}@md.tsukuba.ac.jp

あらまし　より早期の段階で疾患の前兆をとらえるために健康診断データを用いた疾患の予測研究が行われているが、それらの研究では疾患の予測性能の向上や新たな手法の提案に注力しているため、予測モデルの解釈性については大きく触れていない。そのため、本研究では、疾患の進行について予測を行い、予測された理由が解釈可能なモデルの構築を行うことを目的とする。本研究では、健康診断の予防対象である糖尿病の判定について予測する多層ニューラルネットワークによるモデルの構築を行った。また、予測モデルを解釈するためにモデルにおける属性の重要度を算出する手法である Permutation importance と、入力を摂動させることにより入力と出力の関係を解釈する手法である Sensitivity analysis の 2 つの手法を用いた。実験の結果、関連研究で用いられている他の機械学習手法と比較して、高い性能を持つ予測モデルを構築することができた。Permutation importance では、構築した予測モデルが糖尿病の診断や判定に用いられる属性を重視しており、またそれら以外の利用されていない属性も重視した予測を行っていることが分かった。Sensitivity analysis による質問票の質問項目の回答に対する摂動による予測結果の比較では、多くの質問項目において、回答の変化と予測結果の変化が質問の意図と一致した一方で、一致しない質問項目もあった。

キーワード　機械学習、健康診断データ、疾患予測、機械学習における解釈性、Sensitivity analysis

1　はじめに

1.1　背景と目的

健康診断は、検査による測定や問診を通して現在の健康状態を把握し、疾患の前兆を捉え疾患を早期発見する目的を持つものである。疾患の早期発見は、疾患の重症化を防ぐことができる。それにより、人々の健康寿命を伸ばすことができ、増え続ける医療費の削減にもつながる。そのため、より早期の段階で疾患の前兆をとらえるために、医療データを用いた疾患の予測研究が行われている。それらの研究には、機械学習による手法が取り入れられており、様々な観点から研究が行われている。しかし、それらの研究では、疾患の予測性能の向上や手法の提案に注力している。それらの予測モデルが、なぜその予測結果を導いたのかといった解釈性については、大きく触れられていない。

また、健康診断データを用いた予測研究では、糖尿病といった疾患の発症予測が試みられている。しかし、健康診断が主な対象としている生活習慣病は、感染症などとは異なり発症までに時間がかかる。したがって、ある時点での健康診断データを用いて発症を予測することができたとしても、すでにその受診者は、生活習慣の改善だけでは予防できない段階まで病状が進行している可能性がある。特に、糖尿病は生活習慣が主な要因となる疾患である。そのため、事前に悪化を予測することができれば、健康指導により生活習慣の改善を促すことができ、より早期の段階で発症を防ぐことができる。

以上の点から、本研究では、健康診断データを用いて疾患を発症する前の段階である未病段階における疾患の進行について予測を行う機械学習による解釈可能な予測モデルの構築を目的とする。本研究では、健康診断が対象としている生活習慣病のうち糖尿病を予測対象とする。また、予測モデルを解釈可能とするために、次の 2 つの手法を用いる。1 つ目は、予測モデルが重視している属性を明らかにするために予測モデルにおける各属性の重要度を算出する手法である。2 つ目は、受診者の生活習慣や既往歴を尋ねた質問票に着目し、それらの回答に摂動を与えて予測モデルに入力し、摂動を与える前と与えた後のデータによる予測結果を比較する手法である。

2　糖尿病とその診断

2.1　糖尿病とは

糖尿病は、生活習慣病として挙げられる疾患の一つである。日本糖尿病学会では [1]、「インスリン作用不足による慢性の高血糖状態を主徴とする代謝疾患群」と定義されている。インスリンは、膵臓のランゲルハンス島によって分泌され、各臓器に対して影響することにより、血中の血糖値を低下させる作用を持つ物質である。糖尿病は、主にこのインスリンの分泌低下の要因によって、1 型糖尿病と 2 型糖尿病に分類されている。このうち 2 型糖尿病は、複数の遺伝的要因などによるインスリンの分泌低下に加えて、加齢や肥満などの生活習慣から生まれる要因などが組み合わさることによって発症する疾患である。生活習慣病としての糖尿病は、この 2 型糖尿病を指す。また、糖

尿病は、多くの合併症を引き起こす疾患として知られ、癌として知られる悪性新生物や心疾患、脳疾患、認知症といった疾患にも発展する可能性を持つ。このような重大な疾患にもかかわらず、急速に病態が悪化する感染症とは異なり、生活習慣による影響が大きいために病状は徐々に進行し、健康診断や医療機関で診察を受けた際にはすでに病状が進行していることが多い。

2.2 健康診断における判定基準

健康診断における糖尿病に対する判定には、ヘモグロビン A1c と早朝空腹時血糖値が用いられている。日本人間ドック学会 [2] では、この 2 種類の測定値について一定の基準値を設け、それぞれ基準値の段階により判定区分を決定している。これらの判定に基づき異常が認められた場合には、精密検査を受診させるなどして糖尿病の正確な診断を行う。

3 機械学習の解釈に用いる手法

3.1 Permutation importance

Permutation importance [3] は、機械学習によるモデルにおける属性の重要度を算出する手法である。Permutation importance は、以下の手順で算出される。

（1） 算出したいデータセットの属性に対して、その属性に含まれる値をランダムに並び変える。

（2） 変化させたデータセットを予測モデルに入力し予測させ、評価する。

（3） 並び替える前後のデータセットでの予測結果による評価指標の値の差を計算する。

モデルにおいて重要な属性であれば、予測対象との結びつきが強い。その場合、その属性をデータセット中で並び替えたときには結びつきが失われるため、予測結果における評価指標の値が大きく減少する。したがって、ある属性に含まれる値をランダムに並び変えて入力したときに、予測結果における評価指標の値が大きく減少した場合に、その属性が予測モデルにおいて予測性能に大きな影響を与えていると解釈することができる。Permutation importance では、並び替える前後での予測結果における評価指標による値の差を重要度とする手法である。

Permutation importance は、ある属性に含まれる値をランダムに並び変えて、入力したときの予測結果における評価指標の値が大きく減少したときに、その属性が予測モデルにおいて予測性能に大きな影響を与えていると解釈することができる。

3.2 Sensitivity analysis

Sensitivity analysis は、機械学習によるモデルにおける入力と出力の関係を解釈するための手法である。機械学習を用いた分類や予測を行うモデルでは、入力に対応した出力を行うが、なぜその入力から出力を行うのかを理解することは難しく、また、入力データに含まれる属性の変化が、どのように出力に影響を与えるのかを理解することは難しい。Sensitivity analysis は、入力する属性の値に対して摂動を与え、摂動を与える前の値と摂動を与えた値による予測結果の違いを分析することにより、属性ごとの入力と出力の関係を明らかにする手法である。それにより、ある属性について摂動を与えたときの予測結果の変化の仕方から、その属性の増減などの変化がどのような予測結果に結びついているかを解釈することができる。

Sensitivity analysis を用いた研究には、Mussone ら [4] の研究が挙げられる。Mussone らは、構築した交通事故における重症度を予測する多層ニューラルネットワークによる予測モデルを用いて、入力された道路状況や運転者の属性について摂動させ、それらの属性が重症度の予測に影響しているのかを分析している。

4 関 連 研 究

4.1 健康診断データを用いた生活習慣病の発症予測に関する研究

恒川ら [5] は、健康診断データを入力とした悪性新生物を除いた、生活習慣病を 1 年以内に発症するかについてランダムフォレストを用いて、発症した例を正例、発症しなかった例を負例とした 2 値について予測を試みている。この研究では、健康保険組合が保有する医療機関を受診した際に発行される診療報酬明細書などのレセプト情報を活用している。正例に用いられているデータは、レセプト情報から生活習慣病が重症した患者を抽出し、レセプト情報に記載された時点から直前 1 年間以内にその患者が受診した健康診断でのデータを取り出したものである。一方、負例に用いられているデータは、全体の健康診断データのうちそれぞれの健康診断を受診した人が、その健康診断を受診してから 1 年以内にレセプト情報が存在する場合のみを対象として取り出している。

4.2 電子カルテデータを用いた糖尿病の発症予測に関する研究

Garske [6] は、電子カルテデータから抽出されたデータを入力とする、深層学習を用いた糖尿病の発症予測を試みている。この研究では、6 か月、12 か月、18 か月、24 か月後にそれぞれの段階で、糖尿病を診断されるか、否かとした 2 値について予測を試みている。学習に用いたデータセットは、電子カルテデータから抽出した、年齢、性別、人種や家族に糖尿病の患者がいたか、異常血圧であるか、ヘモグロビン A1c に関する値、BMI に関する値、収縮期血圧及び拡張期血圧に関する値からなる 22 種の属性から構成されている。データセットに含まれる患者からは、すでに 2 型糖尿病または妊娠糖尿病と診断されている、もしくは代謝性疾患の履歴を持つ者を除外している。この研究では、6 か月、12 か月、18 か月、24 か月後について予測を行う学習モデルをそれぞれ構築し、比較している。

4.3 リカレントニューラルネットワークを用いた健康診断における測定値の予測に関する研究

Kim ら [7] は、複数年にわたる健康診断データを用いて健康診断で測定される測定値について予測を試みている。この研究では、人為的要因などよって特定の検査項目が欠損していた場合に、その検査項目に関連した疾患について診断を下すことができないという問題に着目している。健康診断を受診したとし

ても、検査による測定値が失われずに取得されなければ、疾患の判定や診断を行うことはできない。そのため、この研究ではそれらの欠損を持つ健康診断データに対しても診断を行うことできるように、欠損した測定値について予測を試みている。そのデータセットを、時系列性を考慮して学習させることができる単純なリカレントニューラルネットワーク、LSTM、ロジスティック回帰を用いてそれぞれ予測させ、比較を試みている。

4.4 健康診断データを用いた次年度の健康診断の非受診者予測に関する研究

下田ら [8] は、ある年の健康診断データと、その年及び過去数年間の健康診断受診の有無を用いて、ある年の健康診断を受診した人が翌年に健康診断を受診するか否かについて予測を試みている。健康診断は、受診した結果による診断と、それに対しての健康指導を行うことによって、疾患の発症予防や早期発見を目的とするものである。そのため、受診しなければその効果を発揮することはできない。そのためこの研究では、健康診断の受診率を向上させるために健康診断データからある受診者が次年度に健康診断を受診するか、否かの 2 値について予測を試みている。

5 実験方法

5.1 構築する予測モデルの探索方法

予測モデルの構築には、複数の層からなるニューラルネットワークを組み合わせた機械学習手法である多層ニューラルネットワークを用いた。多層ニューラルネットワークは、機械学習手法において高い予測性能とモデル設計の自由度の高さから近年、多様な研究が行われている。本研究では、多層ニューラルネットワークによる予測モデルの構築にあたって、以下の 4 点についてパラメータを変更し最良のモデルの探索を行った。モデルを構成する多層ニューラルネットワークの層は、すべて全結合層とした、各層の活性化関数には、ReLU 関数を用い、出力層の活性化関数には、シグモイド関数を用いた。また、損失関数には、Binary Cross Entropy を用いた。

- モデルを構成する層数：1 層から 10 層
- 各層のユニット数: 2, 4, 8, 16, 32, 64, 128, 256
- Optimizer：SGD または Adam
- Optimizer の学習率：0.01, 0.001, 0.0001, 0.00001

以上の変化のみによる探索だけでなく、それらの変化に加えてモデルの各層を構成している全結合層間に

- Batch Normalization 層
- Dropout 層 (Dropout 率 0.2)

をそれぞれ挿入して探索を行った。

モデルごとに訓練データを用いて学習を行い、テストデータを用いて評価を行った。

モデルの学習では、200 エポックずつ学習を行った。それぞれのエポックごとに検証データを用いて性能の検証を行い、検証データを用いた損失関数による値が 200 エポックの間で最小となったモデルを採用した。また、連続して 20 エポックの間に損失関数による値が改善しなかった場合には、Early Stopping により学習を打ち切った。

5.2 予測モデルの評価指標

予測モデルの評価指標として、以下の指標を計算した。

- 正確度 (Accuracy)
- 適合率 (Precision)
- 再現率 (Recall)
- F1 値 (F1 Score)
- マシューズ相関係数 (Matthews Correlation Coefficient, MCC) [9]

マシューズ相関係数は、それぞれの正解と予測結果の関係が表 1 のような関係で表されるとき、

表 1　正解と予測結果の関係

		予測	
		Positive	Negative
正解	Positive	TP	FP
	Negative	FN	TN

次の式で導出される指標である。

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TP + FP) \times (TN + FN)}}$$

本研究で扱う「健康診断」では、その性質上、健康診断を受診したすべての受診者に異常が見つかるわけではない。よって、受診者のうち、健康な受診者よりも、異常が見つかる受診者の割合が少ない。そのため、本研究で扱う健康診断データによるデータセットは、健康な受診者と異常が見つかる受診者の人数差が大きい不均衡なデータセットである。ゆえに、正確度だけでは正しい予測性能を持ったモデルかどうか判断することはできない。すべての予測結果を健康であるとしたときであっても、高い正確度となるためである。

一方、マシューズ相関係数は、−1 から 1 の間の値をとり、表 1 における TP と TN がそれぞれ多いほど値が 1 に、それぞれ少ないほど値が −1 に近づく指標である。また、マシューズ相関係数では、予測対象の 2 値どちらを正例・負例としても同じ値が導かれるため、正例・負例の 2 値に対して、正しく予測できているかを評価することができる。本研究ではマシューズ相関係数による値によって比較を行い、最良なモデルを選択した。

5.3 データセットの概要

本研究に使用しているデータセットは、「筑波大学附属病院 水戸地域医療センター JA 茨城県厚生連総合病院 水戸協同病院」で収集された健康診断データを用いた。

データセットに含まれる年度は、2016 年度、2017 年度、2018 年度の 3 年度である。

データセットに含まれている属性のうち、含まれる値のうち 95%以上が欠損している属性については、学習に用いる属性から除外した。また、糖尿病以外の測定値についての判定については、学習に用いる属性から除外した。

学習に用いた属性のうち、質問票の質問項目以外の属性は、

年齢、性別、身長、体重、胴囲、BMI、収縮期血圧、拡張期血圧、総コレステロール、LDL コレステロール、HDL コレステロール、空腹時血糖（FBS）、ヘモグロビン A1c(HbA1c)、糖尿病の判定、ヘモグロビン、赤血球数、ヘマトクリット、白血球数、尿酸、尿検査：潜血、尿検査：タンパク質、尿検査：グルコース、便潜血：1 日目、便潜血：2 日目、中性脂肪、コリンエステラーゼ（ChE）、クレアチニン、アルブミン、アラニントランスアミナーゼ、アスパラギン酸トランスアミナーゼ、γ-グルタミルトランスペプチダーゼ、C 反応性タンパク質、電気心電図についての判定、腹部エコーについての判定、胸部 X 線についての判定、上部消化管 X 線についての判定、眼科についての判定、血清についての判定、肝炎についての判定からなる 39 種類である。各判定は、「特になし、軽度異常、経過観察、要治療、要精検、治療中」の 6 段階である。

また、質問票の質問項目を表 2 に示す。

表 2　データセットに含まれる属性 (質問票)

質問内容
Q1：血圧を下げる薬の使用の有無
Q2：インスリン注射又は血糖を下げる薬の使用の有無
Q3：コレステロールを下げる薬の使用の有無
Q4：医師から、脳卒中（脳出血や脳梗塞等）にかかっているといわれたり、治療を受けたことがありますか。
Q5：医師から、心臓病（狭心症や心筋梗塞等）にかかっているといわれたり、治療を受けたことがありますか。
Q6：医師から、慢性の腎不全にかかっているといわれたり、治療（人工透析）を受けたことがありますか。
Q7：医師から、貧血といわれたことがある。
Q8：最近 1ヶ月間、たばこを吸っている。
Q9：20 歳のころより体重は 10kg 増えた。
Q10：1 回 30 分以上の軽く汗をかく運動を週 2 日以上、1 年以上実施している。
Q11：日常生活において歩行又は同等の身体活動を 1 日 1 時間以上実施している。
Q12：ほぼ同じ年齢の同性と比較して歩く速度は速い。
Q13：この 1 年間で体重の増減が± 3kg 以上あった。
Q14：人と比較して食べる速度が速い。
Q15：就寝前の 2 時間以内に夕食をとることが週に 3 回以上ある。
Q16：夕食後に間食 (3 食以外の夜食) をとることが週に 3 回以上ある。
Q17：朝食を抜くことが週に 3 回以上ある。
Q18：お酒（清酒、焼酎、ビール、洋酒など）を飲む頻度
Q19：飲酒日の 1 日当たりの飲酒量
Q20：睡眠で休養が十分とれている。
Q21：運動や食生活等の生活習慣を改善してみようと思いますか。
Q22：生活習慣の改善について保健指導を受ける機会があれば、利用しますか。

6　データセットへの前処理

5.3 節で述べたデータセットに対して、以下の前処理を行った。

6.1　欠損値の処理

値が一部欠損している属性のうち、検査による測定値など連続値による属性については、欠損値をその属性の平均値に置き換えた。また、質問票の質問項目の回答などの離散値による属性については、欠損値を欠損であることを示す値に置き換えた。

6.2　連続値の標準化

データセットに含まれてる連続値である収縮時血圧や空腹時血糖といった属性は、大小の範囲が異なる。そのため、それぞれの属性における大小の範囲の差異を除くために、属性ごとに標準化を行った、標準化では、その属性の平均値を 0、分散を 1 となるように scikit-learn の StandardScaler を用いて行った。まず訓練データを標準化し、次にテストデータに対して訓練データの平均値及び分散となるように標準化を行った。値 x について、標準化による変換は、属性の平均値を μ、標準偏差を σ、標準化後の値を z としたときに、以下の式で表される。

$$z = \frac{x - \mu}{\sigma}$$

6.3　離散値の One-Hot 表現化

データセットには、検査による測定値などの連続値と、質問票の質問項目の回答といった離散値がどちらも含まれている。離散値の値の大小には意味がない。そのため、連続値及び離散値の属性を組み合わせて学習に利用できるように離散値の属性を One-Hot 表現とした。

6.4　正解ラベルのラベリング

データセットの属性に含まれている、測定値から判定された糖尿病の判定には、「特になし、軽度異常、経過観察、要精検、要治療、治療中」の 6 段階である。この 6 段階の判定について年度間の判定を比較し、入力年度の次年度における糖尿病の判定の変化が「維持・改善」及び「悪化」の二種類となるようにラベリングを行った。ラベリングでは、年度間の糖尿病の判定を比較し、入力年度よりも次年度の判定が改善または変化しなかった場合には、「維持・改善」、入力年度よりも次年度の判定が悪化方向へ変化している場合には、「悪化」とした。本研究では、入力年度の判定が「特になし」、「軽度異常」、「経過観察」である受診者のデータのみを使用し、入力年度の判定がそれら以外の受診者のデータは使用しなかった。

正解ラベルのラベリング例を表 3 に示す。

表 3　正解ラベルのラベリング例

入力年度の判定	次年度の判定	正解ラベル
特になし	軽度異常	悪化
軽度異常	軽度異常	維持・改善
経過観察	軽度異常	維持・改善

このうち、次年度の糖尿病の判定のうち「治療中」は、その受診者が確実に治療中であると判断するために、質問票の質問項目である「Q2：インスリン注射又は血糖を下げる薬の使用の有無」に「はい」と回答している受診者を「治療中」として扱い、糖尿病の判定が「治療中」にもかかわらず、「いいえ」と回

答している場合には、学習に用いるデータセットから除外した。次年度の糖尿病の判定が「治療中」にもかかわらず、「Q2：インスリン注射又は血糖を下げる薬の使用の有無」に「いいえ」と回答していた受診者は、2 人だった。

6.5 データセットの分割

予測モデルの学習及び評価に用いるために、前処理を行ったデータセットに対して訓練データ・検証データ・テストデータの比が「8:1:1」となるように分割を行った。分割したデータごとのサイズは表 4 に示す。

表 4　分割したデータセット

	訓練データ	検証データ	テストデータ	合計
維持・改善	2764	346	346	3456
悪化	506	85	86	677

6.6 質問票の回答に対する摂動

本研究に用いている健康診断データによるデータセットには、受診者が回答した質問票の質問項目が 22 項目含まれている。これらの質問には、現在の使用している薬の有無や既往歴を質問した項目、現在の生活習慣を質問した項目、現在までの体重の変化を質問した項目などがある。これらの質問項目は、健康診断データに含まれる測定値とは異なり、その回答が直接疾患の診断にはつながらない。しかし、これらの回答は、生活習慣といった、現時点の検査では分からない受診者の状態を把握することができる。また、健康診断によりこれらの生活習慣を把握し、それらを改善するように指導することによって、受診者の健康状態を現在よりも良い状態へ導くことができる。健康診断が予防対象としている疾患である生活習慣病は、感染症のように急性的に発症するわけではなく、不摂生な生活習慣などを続けた結果に待ち受ける疾患である。それゆえに、現時点で健康診断によって測定される値によって、異常が見つかるということは、過去の生活習慣の結果が反映された結果ともいえる。

そのため、本研究では、この質問票の質問項目の回答について着目する。入力された値と予測モデルによる予測結果の関係を解釈するために、他の属性を固定し、それぞれの質問項目の回答に対して摂動を与えて予測モデルに予測させる。そして、摂動を与える前のデータと与えた後のデータでは予測結果が変化するかを実験し、その結果が変化する場合には、その結果が人間にとって理解できる変化であるかを考察する。

質問項目以外の属性は、現時点においてすでに確定している結果であり、受診者が変えることはできない。しかし、個人の努力によって生活習慣などは変化させることができ、また予測結果の変化が客観的に理解できるため、質問票の質問項目の回答に着目した。

質問票の質問項目の回答に対する摂動を与える実験の手順を以下に示す。

（1） 摂動させる質問項目を 1 つ選択する。

（2） その質問項目の回答について、摂動を与える。

（3） 質問項目の回答について摂動を与えなかった健康診断データを予測モデルに入力し、予測させる。

（4） 質問項目の回答について摂動を与えた健康診断データを予測モデルに入力し、予測させる。

（5） 摂動を与えなかったデータと摂動を与えたデータの間で予測結果が変化したかそれぞれ集計する。

（6） 2 から 4 を回答の種類だけ繰り返す。

実験にはテストデータを使用し、それぞれの質問項目ごとにその質問項目の回答を変化させることができるデータのみを抽出した。例えばある質問項目の回答を「はい」から「いいえ」へ変化させる場合には、テストデータから「はい」と回答しているデータのみを抽出した。

7 実験結果

7.1 実験設定

学習に用いるデータセットに含まれる正解ラベルの比率が不均衡であるため、正解ラベルの「維持・改善」及び「悪化」に対して、それぞれ表 5 の重みを設定し学習を行った。各ラベルの重みの算出には、scikit-learn の compute_class_weight メソッドを用いた。

表 5　正解ラベルごとの重み

正解ラベル	重み
維持・改善	0.62409551
悪化	2.51457726

7.2 構築した予測モデル

探索した結果、最もマシューズ相関係数が高い予測モデルは、8 層からなる図 7.2 の構造を持つモデルとなった。なお、この予測モデルの全結合層間には、Batch Normalization 層及び Dropout 層は挿入されていない。

構築した予測モデルの Optimizer と Optimizer の学習率を以下に示す。

- Optimizer: SGD
- Optimizer の学習率: 0.01

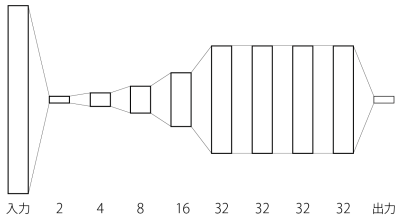

図 1　構築した予測モデルの構造

7.3 機械学習手法間での予測性能の比較

関連研究で用いられていた手法のうち、以下の手法について同様のデータセットを用いて学習を行い、テストデータを用いて評価を行った。構築した予測モデルの学習と同様に、学習時に正解ラベルの「維持・改善」及び「悪化」に対して、表 5 の重みを設定した。

- 勾配ブースティング決定木 (XGBoost) [10]
- ランダムフォレスト
- ロジスティック回帰
- サポートベクターマシン (SVM)

それぞれの手法における評価指標ごとの値を表 6 を示す。Precision、Recall、F1 Score は、「悪化」ラベルに対して算出した値を示す。

表 6　機械学習手法間の比較

手法名	Accuracy	Precision	Recall	F1 Score	MCC
構築したモデル	0.715	0.390	**0.788**	**0.521**	**0.394**
XGBoost	0.787	0.462	0.506	0.483	0.350
ランダムフォレスト	0.606	0.351	0.612	0.446	0.280
ロジスティック回帰	**0.812**	**0.540**	0.318	0.400	0.312
SVM	0.696	0.363	0.718	0.482	0.343

7.4　Permutation importance

構築した予測モデルに対して、テストデータを用いてモデルに入力された各属性の Accuracy に基づく Permutation importance を算出した。各属性の Permutation importance のうち、上位 25 属性を図 2 に示す。

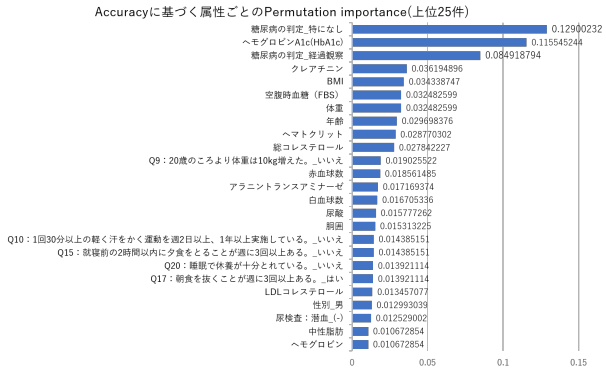

図 2　予測モデルにおける Permutation importance

8　考　　察

8.1　予測モデルの性能

今回構築した予測モデルでは、比較した機械学習手法よりも、マシューズ相関係数による評価に基づき高い予測性能を持つ予測モデルを構築することができた。本研究では、データセットの都合上、予測対象を「維持・改善」及び「悪化」の 2 値とした。しかし、予測対象を「維持」、「改善」及び「悪化」の 3 値分類や 6 段階の糖尿病の判定そのものを予測対象とすることができれば、より糖尿病の進行を予測する予測モデルとしての有用性が高まると考えられる。本研究で使用したデータセットは、一般に多層ニューラルネットワークによる予測モデルの学習に用いられるデータセットのサイズよりも比較的小規模であった。個人を特定可能な高度な個人情報を含む健康診断データを大規模にかつ継続的に収集することは難しいが、より大規模なデータセットを使用することができれば、予測性能の向上や予測モデルについてより良い評価を行うことができると考えられた。

8.2　Permutation importance

図 2 の Permutation importance の算出結果を見ると、「糖尿病の判定」や「ヘモグロビン A1c」、「空腹時血糖値」といった糖尿病の診断や判定に用いられる属性が上位に位置していることが分かる。この結果から今回構築した予測モデルでは、既存の糖尿病の判定や診断に用いられる指標を考慮した学習及び予測を行うことができていると考えられる。また、糖尿病の診断や判定に用いられる属性につづいて、「年齢」や「体重」、「BMI」といった属性も上位に位置していることが分かる。糖尿病は、加齢や肥満などにより発症する可能性が高まる疾患であるため、これらの属性についても考慮した予測を行うことができていると考えられる。

また、腎臓の状態を把握するために測定されるタンパク質である「クレアチニン」や血中の状態を把握するための血液中の赤血球の体積である「ヘマトクリット」や「総コレステロール値」が上位に並んでいることが分かる。これらの結果から、今回構築した予測モデルは、糖尿病に関する指標だけでなく、他の指標の関係を考慮した予測を行うことができていると考えられる。本研究では触れることができなかったが、これら属性の値の変動や相互関係が予測結果に影響を与えているか分析することによって、糖尿病の判定及び診断には、現在直接用いられていない属性と糖尿病の関係性を明らかにすることができると考えられる。

8.3　質問票の回答に対する摂動

質問票の質問項目には、受診者の生活習慣などから診断や健康指導を行うことができるようにそれぞれ意図が設定されている。厚生労働省健康局の標準的な健診・保健指導プログラム [11] において示されている質問の意図に従い、その予測結果の変化がその質問の意図に従っているか判断した。

既往歴の有無や薬の使用の有無を尋ねた質問では、既往歴がある、薬を使用しているとした回答を、既往歴がない、薬を使用していないと回答を変化させたときに「維持・改善」となり、既往歴がない、薬を使用していないとした受診者の回答を、既往歴がある、薬を使用しているとしたときに「悪化」となるか判断した。

生活習慣について尋ねた質問では、生活習慣としてより好ましい回答へ変化させたときに「維持・改善」となり、生活習慣として好ましくない回答へ変化させたときに「悪化」となるか判断した。

飲酒及び喫煙について尋ねた質問では、現在の回答から飲酒及び喫煙を控えた回答へ変化させたときに「維持・改善」となり、現在の回答から飲酒及び喫煙する回答へ変化させたときに「悪化」となるか判断した。

生活習慣の改善への意識について尋ねた質問では、現在の回答より改善への意識が高い回答へ変化させたときに「維持・改善」となり、現在の回答よりも改善への意識が低い回答へ変化させたときに「悪化」となるかを判断した。

それぞれの回答の変化及び予測結果の変化と質問意図との比較を表 7 に示す。

表 7　回答の変化及び予測結果の変化と質問意図との比較

質問内容	質問意図と一致	質問意図と相違
Q1	○	
Q2	○	
Q3		○
Q4	○	
Q5		○
Q6	○	
Q7	○	
Q8		○
Q9		○
Q10	○	
Q11		○
Q12		○
Q13	○	
Q14		○
Q15	○	
Q16		○
Q17		○
Q18		○
Q19	○	
Q20	○	
Q21	○	
Q22	○	

また、回答の変化と予測結果の変化が質問の意図と一致した質問項目と一致しなかった質問項目の摂動による予測結果の変化をそれぞれ 2 項目ずつ示す。それぞれの円グラフ上の値は、「維持・改善」と「悪化」ラベルごとに摂動を与えた後のデータにより予測された件数を示している。

8.3.1　回答の変化と質問の意図が一致した質問項目

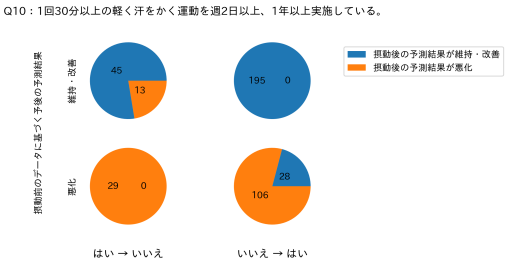

図 3　Q10 における摂動による変化

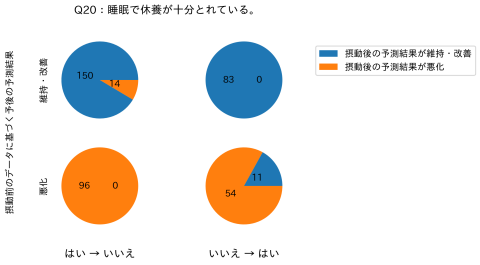

図 4　Q20 における摂動による変化

8.3.2　回答の変化と質問の意図が一致しなかった質問項目

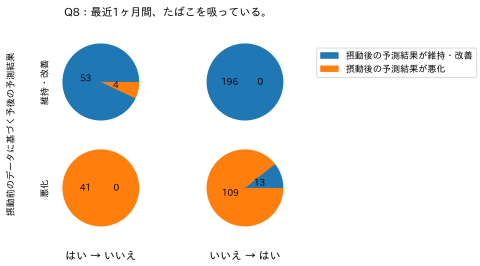

図 5　Q8 における摂動による変化

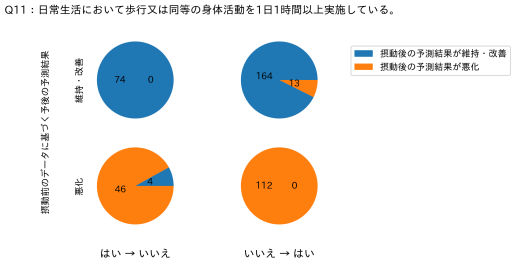

図 6　Q11 における摂動による変化

表 7 より、質問票の質問項目の回答を摂動させたときに、回答の変化による予測結果の変化が質問の意図と一致したものは、12 項目である。一方、回答の変化と予測結果の変化が質問の意図と一致しなかったものは、10 項目である。多くの質問項目において、回答の変化による予測結果の変化が質問の意図と一致していた。そのため、構築した予測モデルでは、これらの質問項目の意図を考慮した予測を行うことができていると考えられる。一方で、回答の変化による予測結果の変化質問の意図と一致しなかった質問項目もある。例えば、図 5 の「Q8：最近 1ヶ月間、たばこを吸っている。」の質問では、喫煙していた受診者を「喫煙しない」と変化させると一部の「悪化」と予測されていた受診者において、「維持・改善」に予測結果が変化した。また、喫煙していなかった受診者を「喫煙する」と回答を変化させたときにも、一部の「維持・改善」と予測されていたの受診者において、「悪化」と予測結果が変化した。

喫煙は、血糖値を上昇させ、糖尿病の発症を促進するとされているため、回答の変化による予測結果の変化が質問の意図と一致していない。一方で、福井ら [12] は、運動や適度な飲酒習慣は、インスリン抵抗性を低下させ、一方で喫煙習慣はインスリン抵抗性については影響を与えないと報告している。インスリン抵抗性は、臓器がインスリンの影響を受けづらくなり体内でのインスリンの作用が低下している状態を指す。また、喫煙者が禁煙した場合、食欲が増し体重が増加することがあり、生活習慣を意識しなければ肥満につながることがある。肥満もまた糖尿病の発症原因の 1 つである。今回構築した予測モデルは、データセットからこれらの関係を学習したと考えることもできる。

本研究で構築した予測モデルでは、入力年度の健康診断デー

タを用いて次年度の糖尿病の判定を予測している。したがって、受診者の生活習慣の変化を踏まえない短期的なデータを用いた予測となっている。また、予測性能もマシューズ相関係数による評価により他の機械学習手法と比較して優位であっても、単純な正確度の観点から見るとテストデータの 7 割しか正しく予測することができていない。そのため、モデルの学習に起因し、喫煙しない受診者を「喫煙する」と変化させた場合に、「維持・改善」と予測が変化するという結果につながったとも考えられる。

受診者の生活習慣の変化を考慮するためには、複数年の健康診断データを時系列データとして扱い、生活習慣の変化を踏まえた予測モデルの構築を行う必要があると考えられる。それにより、今回の摂動による予測結果の変化が、予測モデルの性能によるものなのか否かを正確に検証することができると考えられる。

9 おわりに

本研究では、医療データのうち健康診断データを用いて、次年度の健康診断における糖尿病の判定を予測するモデルを構築した。その予測モデルを解釈可能とするために、入力した健康診断データにおけるどの属性を重視して予測しているのかを Permutation importance を用いて示した。また、質問票の質問項目の回答と予測結果の関係について着目し、Sensitivity analysis によって入力する質問表の質問項目の回答を摂動させることにより、その回答の変化が「維持・改善」及び「悪化」の予測に影響するのかを示した。

パラメータを変化させモデルの探索を行った結果、本研究では他の機械学習手法よりも高い予測性能を持つモデルを構築することができた。多層ニューラルネットワークは、様々なモデル設計を行うことができるが、今回は全結合層から構成されるモデルのみしか扱うことができなかった。また、予測モデルにおける属性の重要度である Permutation importance の算出結果からは、入力時点での糖尿病の判定や糖尿病の診断や判定に用いられる指標を重視した予測を行っていることが分かった。この点から、糖尿病について予測する予測モデルとして正しい指標をもとに予測できていると考えられた。また、糖尿病の診断や判定に用いられる指標以外も、重視した予測を行っていることがわかり、これらの値を変化させ予測し、入力と出力の関係を細かく分析することによって新たな関係性を見つけられると考えられた。

質問表の質問項目の回答に摂動を与える Sensitivity analysis では、多くの質問項目において回答の変化と予測結果の変化が質問の意図と一致した。本研究で構築した予測モデルでは、一致した質問項目について、これらの質問項目の意図を考慮した予測を行うことができていると考えられる。一方で、一致しなかった質問項目もあり、それらの回答の変化と予測結果の関係は、予測性能に起因するものか否かを検証する必要がある。本研究で構築した予測モデルでは、ある 1 年度の健康診断データを入力し、その次年度の糖尿病の判定について予測を試みた。したがって、予測モデルの学習及び予測では、入力年度時点での健康状態や生活習慣しか考慮することができず、それまでの健康状態の変化や生活習慣を考慮することができていない。生活習慣や健康状態は、日々変化しているため、それらを考慮した予測モデルを構築することができれば、より高い予測性能を持つ予測モデルを構築することができると考えられる。今後の課題として、生活習慣や健康状態の経年変化を考慮した予測モデルを構築する必要があると考えられる。

謝　　辞

本研究は、JSPS 科研費 JP16K00228，JP16H02904 の助成及び JST、COI、JPMJCE1301 の支援を受けたものである。

文　　献

[1] 清野裕, 南條輝志男, 田嶋尚子, 門脇孝, 柏木厚典, 荒木栄一, 伊藤千賀子, 稲垣暢也, 岩本安彦, 春日雅人, 花房俊昭, 羽田勝計, 植木浩二郎. 糖尿病の分類と診断基準に関する委員会報告 (国際標準化対応版). 糖尿病. 2012, vol. 55, no. 7, p. 485-504.

[2] 日本人間ドック学会. ”2019 年度判定区分表”. 日本人間ドック学会. https://www.ningen-dock.jp/wp/wp-content/uploads/2013/09/ac338c5caeb772b77a32f56332791305.pdf, (参照 2019-12-16).

[3] Altmann, André; Toloşi, Laura; Sander, Oliver; Lengauer, Thomas. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010, vol. 26, no. 10, p. 1340-1347.

[4] Lorenzo, Mussone; Bassani, Marco; Masci, Pietro. Analysis of factors affecting the severity of crashes in urban road intersections. Accident Analysis and Prevention. 2017, vol. 103, p. 112-122.

[5] 恒川充, 岡夏樹, 荒木雅弘, 新谷元司, 吉川昌孝, 谷川武. ”健診データを用いた生活習慣病の発症予測”. 人工知能学会全国大会論文集. 新潟市, 2019-06-04/07, 人工知能学会. 2019, p. 4D3E205.

[6] Garske, Thomas. Using Deep Learning on EHR Data to Predict Diabetes. University of Colorado at Denver, 2018, Ph.D. thesis.

[7] Kim, Han-Gyu; Jang, Gil-Jin; Choi, Ho-Jin; Kim, Minho; Kim, Young-Won; Choi, Jaehun. ”Recurrent neural networks with missing information imputation for medical examination data prediction“. 2017 IEEE International Conference on Big Data and Smart Computing (BigComp). Jeju, 2017-02-13/16, IEEE, 2017, p. 317-323.

[8] Shimoda, Akihiro; Ichikawa, Daisuke; Oyama, Hiroshi. Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. Computer Methods and Programs in Biomedicine, 2018, vol. 163, p. 39-46.

[9] Matthews W Brian. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975, vol. 405, no. 2, p. 442-451.

[10] Chen, Tianqi; Guestrin, Carlos. ”XGBoost: A Scalable Tree Boosting System”. KDD ’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, 2016-08-13/17, ACM, ACM, 2016, p. 785-794.

[11] 厚生労働省健康局. ”標準的な健診・保健指導プログラム (改訂版) 第 2 編: 健診”. 厚生労働省. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/kenkou/seikatsu/dl/hoken-program2.pdf, (参照 2019-12-18).

[12] 福井敏樹, 吉鷹寿美江, 山本由美子, 綾田陽子, 安田忠司. 生活習慣は本当にインスリン抵抗性に影響を与えるか. 人間ドック (Ningen Dock). 2007, vol. 22, no. 1, p. 51-58.