東京大学大学院新領域創成科学研究科
国際協力学専攻

2021 年度
修士論文

機械学習による水道水質内の大腸菌の存在予測-ネパールを事例に

(Presence prediction of E. coli in water taps using machine learning in Nepal)

2022 年 1 月 17 日提出
指導教員　坂本　麻衣子　准教授

黒木　颯

目次

第1章　序章

1.1.　背景

水は日常生活に欠かせない要素である。誰もが清潔な水にアクセスする権利を持っており、共有財としてその権利は侵害されてはならない。しかし、世界では 7 億 8500 万人が基本的な基準を満たす水へアクセスすることができずにいる(WHO/UNICEF, 2019)。また、発展途上国では、水に関連する原因によって年間 220 万人の命が失われている(WHO & Unicef, 2000)。同様にネパールでも、約 550 万人が安全な水にアクセスできていない(Koju et al., 2015)。また、水を原因とする病気で年間 10,500 人の子どもたちが命を落としているのが現状である(*Wateraid, 2011*)。近年、ネパールでは水道を利用できる人口が大幅に増加しており、1990 年に水道を利用できる人口は推定 36％であったのに対して(Budhathoki, 2019)、2016 年には 95％に増加した(Ministry of Health Ramshah Path et al., 2016)。しかし、水道を利用できる人が増えたにもかかわらず水質は十分に安全ではない。カトマンズ渓谷で行われた調査では、水道や地下水から取得した水の 72%から大腸菌が検出された(Warner et al., 2008)。また、ネパール政府が 2014 年から 2016 年にかけて実施した調査では、ネパール 26 都市の約 55％の水道水から大腸菌が検出された(Ministry of Health, 2016)。こうした水質問題に効果的に対処するためには、できるだけ広い範囲の詳細な水質情報が必要である。WHO と UNICEF による合同モニタリングプログラム Joint monitoring programme (JMP)は、世界規模で安全に管理された飲料水、衛生設備、衛生サービスへの家庭単位でのアクセス状況を推計して発表している。しかし JMP は、すべての国における安全に管理された飲料水にアクセスできている家庭を推定するために必要となる、推計対象人口の 50％のデータを収集することができていない(WHO/UNICEF, 2019)。したがって、必要データの収集が優先事項となるが、既に収集されているデータから未収集のデータを高い精度で予測できるモデルが構築できれば、水道事業の運営やデータ収集を効率的に進められると考えられる。

近年、様々な分野において機械学習の応用が盛んであり、高い予測精度を示している。その一例として、120 万枚の画像を用いて 1000 種類のクラス分類を行い、予測精度を競う ImageNet Large Scale Visual Recognition Challenge(ILSVRC)があげられる。2014 年度の優勝モデルである GoogLeNet は約 93%の予測精度を示した(Russakovsky et al., 2015)。機械学習を用いた水質予測の研究として、Ahmed らは(2019)、パキスタンの水道ネットワークから温度、濁度、pH、総溶解固形物の 4 パラメータについて 663 個のサンプルを収集し、水質クラス Water quality class (WQC)を予測した。彼らは、logistic regression (LR)、ranfom forest(RF)、support vector machine(SVM)、多層パーセプトロン、およびその他の機械学習と深層学習モデルを使用した。多層パーセプトロン は 85% という最高の精度を示し、他のモデルを凌駕した。水質予測の分野では、Singha らは(2021)、インドの農業集約地域から 226 の地下水サンプルを集め全溶解固形物、硬度、カルシウム濃度、マグネシウム濃度、ナトリウム濃度、カリウム濃度、炭酸水素塩濃度、塩素濃度、硫酸濃度、硝酸塩濃度、フッ素濃度、リン酸濃度を計測し、これらの水質指標を使用して entropy-based groundwater quality index (EWQI)を予測した。EWQI は複数の水質指標から求

められ、エントロピーにより水質を測定する手法である。機械学習予測モデルとして RF、XGBoost、および、Artificial Neural Network (ANN)が採用され、RF、XGBoost、ANN でそれぞれ決定係数が 0.888、0.927、0.917 と高い予測性能を発揮した。統計学が目的変数と説明変数の関係の説明に主眼を置いているのに対し、機械学習は学習データからパターンを見つけ出し、未知のデータを予測することに主眼を置いている。機械学習が非線形性を柔軟に捉えられることを踏まえると、予測精度が分析の目的である場合は機械学習が適していると考えられる。発展途上国における水質に関連するデータの不足は大きな課題である。しかし、水道情報を用いた機械学習による水道水質予測は、特に発展途上国においてはほとんど行われていない。発展途上国では、水道水の汚染は、水源の汚染、浄水処理の不備、水道管の破損による汚染など、水源から水道までの過程で複数の問題が発生する(Corporation, 2015)。水源から水道までの変数を組み込んだ機械学習モデルを作成することで、水道水の水質予測が可能になる可能性がある。

1.2. 目的

本研究では、ネパールの都市における水道水の大腸菌汚染を予測する機械学習モデルの作成を試みる。発展途上国でよく見られることだが、入手可能な水質データや関連データは限られている。このような制約の中で、どのようにして正確な予測モデルを作成することができるかを検討し、また、作成した機械学習モデルをどのようにして現地の状況下で有効に活用することができるかを考察する。2 章では機械学習、先行研究の既存の知見を整理する。3 章では使用する目的変数・説明変数、予測モデルのアルゴリズムの説明を行う。4 章、5 章では取得した変数の記述統計、予測精度の分析の結果を示し、6 章で結論をまとめる。

第2章　機械学習の概要

2.1.　機械学習の分類

機械学習は人工知能の一つの要素であり、大量のデータをコンピューターに反復的に学習させることで規則性を見つけ出し、未知のデータの判別や予測を行うことができる。機械学習は大きく教師あり学習、教師なし学習、強化学習の 3 つに分類することができる。以下それぞれについて説明する。

2.1.1.　教師あり学習

教師あり学習は、説明変数に対して目的変数が与えられている際に用いられる。説明変数と目的変数から規則性を学習し、目的変数を未知の説明変数から目的変数の判別、予測をすることができる。例えば、商品のリコメンデーションシステム、工場での製品の故障予測などに用いられる。本研究では水道水内の大腸菌の存在有無を目的変数とした教師あり学習を行う。教師あり学習の流れは図 2-1 のように表すことができる。全データを説明変数と目的変数に分けたのち、それぞれをトレーニングデータとテストデータに同じ比率で分割する。トレーニングデータからモデルを構築し、テストデータの説明変数を入力することでテストデータの目的変数の予測値を計算する。予測値と本来の目的変数のテストデータを照合することでモデルの性能を評価する。

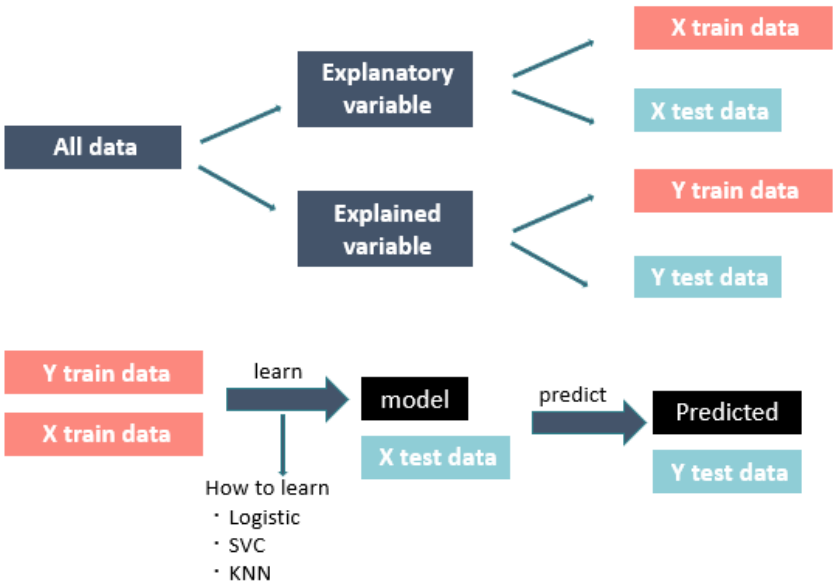


図 2-1 教師あり機械学習の流れ

2.1.2.　教師なし学習

教師なし学習は、説明変数に対して目的変数が与えられていない際に用いられる。クラスタリングによるグルーピング等が教師なし学習に当たり、説明変数間の類似性を分析するなど、データを構造化し解釈をすることで情報を得ることができる。具体的な使用例としては、迷惑メールの検出

に使用されている。迷惑メールに使用される文章や単語には規則性があり、クラスタリングを行うことで迷惑メールとそうでないメールを別クラスに分類することができる。

2.1.3. 強化学習

強化学習は、コンピューターが目的と設定された報酬を最大化するために学習する。代表例としては、ロボット歩行があげられ、例えば「歩行距離」を報酬に設定することで、コンピューターはそれを最大化するために学習を行う。2016 年にそれまで機械には難しいとされていた囲碁で、deepmind 社が開発した AlphaGo がプロ棋士を破ったが、AlphaGo も強化学習のアルゴリズムを用いている。AlphaGo はまずプロ棋士の譜面から教師あり学習をしたモデルを作成した後に、対極に勝利することを報酬に作成したモデル同士の対局を繰り返し行うことで学習を行った。

2.2. 先行研究

2.2.1. Water quality index (WQI)、Water quality class (WQC)の予測

Ahmed らは(2019), パキスタンに位置する水源から 663 個のサンプルを収集した。663 のサンプルは、13 の異なる水源からそれぞれ 51 サンプルずつ取得し、かつ各サンプルごとにアルカリ度、透明度、カルシウム濃度、塩素濃度、電気伝導度、糞便性大腸菌群、硬度、亜硝酸濃度、pH、温度、全溶解固形物、濁度の計測を行った。取得した変数に対して Box plot analysis と Z-score normalization の前処理を行っている。Box plot analysis は外れ値の検出に使用しており、閾値を超える値を閾値内の最大値に、閾値を下回る値を閾値内の最小値に変換している。また、取得した変数から WQI、WQC を計算し目的変数としている。説明変数としては温度、濁度、pH、全溶解固形物を用いている。また精度の評価は Mean Absolute Error (MAE)、Mean Square Error (MSE)、Root Mean Squared Error (RMSE)、決定係数を用いている。表 2-1 に示す機械学習のアルゴリズムのうち、Gradient Boosting が MAE、MSE、RMSE、で最小値、かつ決定係数で最大値を示しており、最も良い精度で WQI を予測した。また、WQC のクラス分類では Multilayer perceptron (MLP) が最も高い 85% の予測精度を示した。

表 2-1 WQI の機械学習の結果

Algorithm	MAE	MSE	RMSE	決定係数
Linear Regression	2.6312	11.7550	3.4286	0.6573
Polynominal Regression	2.0037	7.9467	2.8190	0.7134
Random Forest	2.3053	9.5669	3.0930	0.6705
Gradient Boosting	1.94642	7.2011	2.6835	0.7485
SVM	2.4373	10.6333	3.2609	0.3458
Ridge Regression	2.6323	11.7500	3.4278	0.4971
Lasso Regression	3.5850	20.1185	4.4854	-2.9327
Elastic Net Regression	3.6595	20.9698	4.5793	-4.0050

2.2.2. 地下水質予測

Singha らは(2021)、インドの農業集約地域から 226 の地下水サンプルを集め、各サンプルごとに全溶解固形物、硬度、カルシウム濃度、マグネシウム濃度、ナトリウム濃度、カリウム濃度、炭酸水素塩濃度、塩素濃度、硫酸濃度、硝酸塩濃度、フッ素濃度、リン酸濃度を計測した。計測した変数から entropy-based groundwater quality index (EWQI)を計算し目的変数としている。説明変数データの前処理として min-max normalization を使用している。予測精度の評価は MSE, MAE, RMSE, RMSE-observations standard deviation ratio (RSR), Nash-Sutcliffe efficiency (NSE), mean absolute percentage error (MAPE), 決定係数を用いている。械学習予測モデルとして RF、XGBoost、および、ANN が採用され、RF、XGBoost、ANN でそれぞれ決定係数の値が 0.888、0.927、0.917 と高い予測性能を発揮した。

2.2.3. 河川水の pH 予測

Memon らは(2011) パキスタンの都市であるハイデラバード内の水源である河川、処理場内の処理水を含む 10 か所に測定ポイントを設置して、各ポイントから 1 週間に 2 回のサンプルの取得及び、鉄濃度、塩素濃度、カルシウム濃度、マグネシウム濃度、硬度、全溶解固形濃度、硫酸塩濃度、濁度、そして pH の測定を行った。河川水の pH の予測結果として(ANN)が決定係数 0.99 の値を示した。

2.2.4. 下下水処理場での有機汚濁の予測

Abyaneh は(2014)イランのテヘランにある下水処理場で有機汚濁(Biochemical oxygen demand (BOD) 及び Chemical oxygen demand (COD))の予測を行った。1998 年から 2002 年の間に得られたデータから pH, 全溶解固形濃度、全蒸発残留物の 4 つの変数が有機汚濁と関連があることが分かっていた。したがって、2003 年から 2009 年の 7 年の間、毎月に上記 4 指標と BOD, COD のデータを収集した。予測モデルは線形モデルである重回帰モデル、ANN を使用し、予測の評価指標は、相関係数及び RMSE を使用している。2 つの予測結果を比較したところ ANN が重回帰分析より高い精度を示した。ANN の RMSE が 25.1mg/L であるのに対して、重回帰分析の RMSE は 37.8mg/L であった。

2.3. 企業による取り組み

2.3.1. Fracta, Inc・栗田工業株式会社: 水道管の腐食予測

栗田工業株式会社は日本の東証 1 部に上場している大企業であり水処理薬品、水処理装置、メンテナンス・サービスなどのサービスを提供している企業である。2018 年に米国シリコンバレー発の企業である Fracta, Inc の株式を購入した(日本経済新聞, 2021)。Fracta, Inc は機械学習を用い、水道管の過去破損データを目的変数とし、水道管情報や土壌、気候、人口などの地理的データを組み合わせた説明変数から水道管の破損予測を行うソフトウェアを開発している。Fracta, Inc

は日本でのサービスも展開しており、地方自治体の水道管の劣化予測を行っている。また Fracta, Inc の傘下である Fracta leap では浄水場・処理場での運転最適化を目的に機械学習を含む技術を用いた開発を行っている。

2.3.2. TAIGER: ハイブリッド型

シンガポール発の企業である TAIGER は経理や総務部門の書類確認を高速化する低コストのソフトウェアの開発及び販売を行っている。TAIGER が特徴的なのは機械学習の予測精度のみではなく、人が持つ知見を組み合わせた、ハイブリッド型を提唱していることである(日本経済新聞, 2021)。機械学習は一般的に大量の学習データを必要とするが、必ずしも導入先の企業が大量のデータを既に保有しているわけではない。そこで、多少精度が落ちるものの少量データから得た予測結果と、導入先が持つ既存の知見とを合わせることで、企業を運営する上で十分な精度を従来より素早く得ることができる。さらに、一度導入が進めばその後データがたまるにつれて予測精度は向上していく。

第3章　データおよび分析手法

3.1.　対象地域

本研究では、分析対象データとして、2013-2014 年度にネパール政府が日本国際協力機構(JICA)の支援を受け、他の都市のベンチマークとなるような大都市や歴史ある 26 都市を対象に実施した、水源および水道水のサンプル採取と水消費者への聞き取り調査の結果を用いる。図 3-1 に選定された 26 都市の所在地を示す。当該調査では、カトマンズの 21 の主要水源と無作為に選択された 103 の水道水、および他の 25 都市のすべての水源と無作為に選択された水道水を採取し、それぞれのサンプルについてpH、濁度、遊離残留塩素、大腸菌、電気伝導度などの水質が測定された。また各都市で無作為に選ばれた 50 世帯に聞き取り調査が行われた。聞き取り調査では年間の無水給水日の日数、1 日で水供給が行われる時間等が質問された。

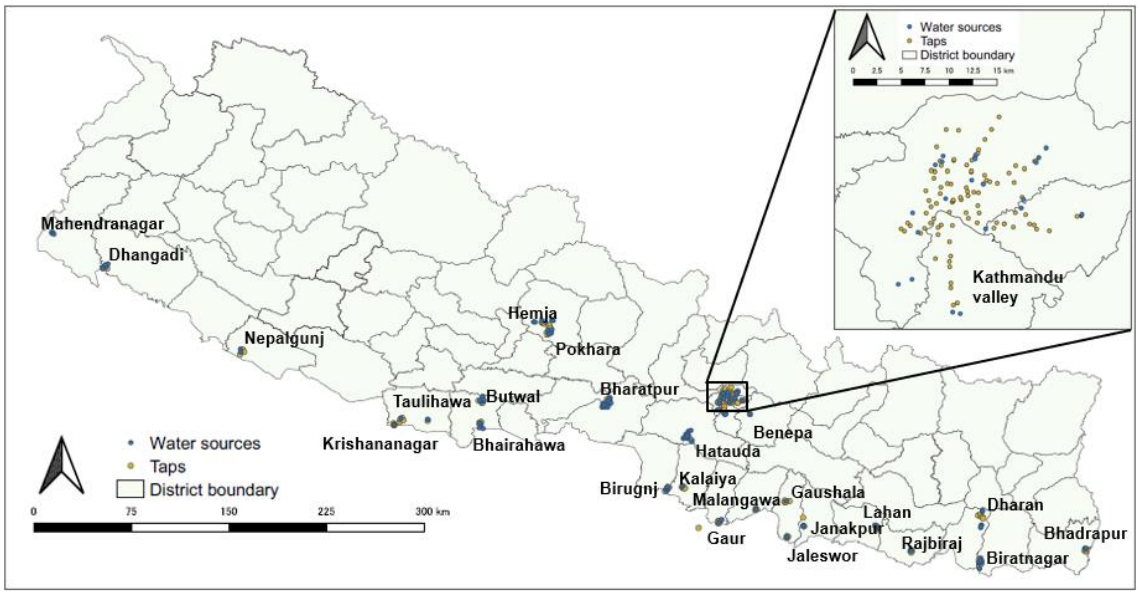


図 3-1　選定された 26 都市の所在地

3.2.　分析の流れ

まず本研究で使用する説明変数を検討し、データを取得した後、機会学習モデルの作成以前に前処理として標準化を行う。その後、モデルを作成し、最後にモデルの精度評価を行う。以下、各手順について説明していく。

3.3 目的変数

本研究では、水の安全性を示す重要な水質指標である、大腸菌の存在有無を予測する。水を介する感染症の主な原因は温血動物の糞便由来であり、大腸菌検査は水道糞便汚染を検知するのに効果的である。例えば水域に糞便汚染がある場合には、同時に赤痢菌、疫痢菌、チフス菌等の病原菌が存在する可能性があり、公衆衛生上大きな問題となる(環境省，2015)。また、人々はそ

の汚染を認識できないため、ろ過や煮沸などの適切な処理をせずにそのまま摂取してしまうという報告がなされている(Ogata et al., 2019)。ネパールの水道水質基準(NDWQS) において、大腸菌は検出されないことが基準となっている。したがった、本研究でも水質基準にのっとり、大腸菌を含まないことを基準にサンプルの分類をする。26 都市から取得した 218 の水道サンプルを大腸菌の検出有無で分類した結果、大腸菌を含まない水道が 98 個、大腸菌を含む水道が 110 個となった。大腸菌を含まない水道を 0、大腸菌を含む水道を 1 と扱い目的変数とすることで、2 クラス分類のモデルを作成する。

3.3. 3. 4 説明変数

まず、網羅的に検討した説明変数を含めたモデルを作成したところ(ステップ 1)、モデルが過学習し、また重要変数の解釈が困難である問題が発生した。これを受けて、説明変数の改善及び水道水中の大腸菌の存在に特に関連性が高いと思われる変数を選択した(ステップ 2)。以下では、各々のステップで用いた説明変数について説明する。

3.3.1. ステップ 1

まず、網羅的に検討した説明変数を表 3-1 に示す。各説明変数を考慮した理由を以下に示す。

(1)水源に関する情報

水源に関連する変数として、以下の 3 つを考慮した。

- 水源の大腸菌数
- 表流水・地下水割合
- 高度

水処理設備が十分に整っていない発展途上国では、水源の水が十分に処理されずに供給される可能性がある。そのため水源が汚染されているか否かは重要な情報である。水源の大腸菌数は、水源の大腸菌検出検査を行った際に検出された大腸菌数の結果を用いた。表流水・地下水割合は各都市毎の水源の表流水と地下水の割合を表している。一般的に表流水の大腸菌汚染率は高く、一方で地下水の大腸菌汚染率は低い。特に市街地を流れる河川(表流水)等は、野外排泄といった生活排水による汚染が起こっている可能性が高い。そのため、ネパールにおいては表流水を水源とする場合は多くが浄水場を経由してから給水が行われている。一方で地下水については、一般的には 20m 以上の深井戸は病原菌に汚染されていることが少ないとされ、ネパールにおいて塩素添加のみ、または塩素添加をせずに給水が行われている場合も多い。以上のような表流水と地下水の違いを捉えることを意図している。高度は表流水において水源が上流、中流、下流のどこに位置するかを表す指標となる。一般的に下流の水は上流より汚染が進んでいる可能性が高いと考えられる。

(2)浄水場で行われる処理

各都市別の浄水場で行われている処理方法を変数に含めた。浄水場で実施される処理方法としては、Flocculation(沈殿処理)、Sedimentation(凝集処理)、Rapid sand filter (急速濾過)、Pressure Filter(加圧濾過)、Roughing Filter(簡易濾過)がある。都市毎に採用されている処理方法が異なり、以上の項目をすべて行っている都市もあれば、地下水から直接水供給を行っており浄水場を介さない場合もある。以上のことから処理手法の組み合わせ、また処理を実施しているか否かは水道水質に密接に関連すると推察される。

(3)水道管に関する情報

浄水場での処理後、または地下水は水道管を通って水道に配水される。したがって水道管の情報は水道水質に関連する。水道管の古さ、長さ、また水道の数を変数に含めた。水道管の古さは、各都市別に最も古い水道管が埋設されてから測定時点までの年数を表している。水道管の古さは水道管の老朽化や水道管のメンテナンスが行われているかを測る指標となる。水道管の長さは、都市別に埋設されている水道管の総長を表している。水道管にかかる負荷、また水が水道管に留まる期間を表す指標となると考えられる。水道の数は各都市別の供給先の水道の数になるが、各都市の水道ネットワークの規模を表す指標となる。

(4)都市別基本情報

都市別の基本情報として 1 日で水供給が行われる時間、年間で水供給が行われない総日数、水道ネットワーク内の総人口・供給人口を変数として扱った。1 日で水供給が行われる時間、水供給が行われない総日数は、聞き取り調査の結果を使用した。水供給が行われない時間、日数はその間に水道管内に水が留まっている可能性が高く、汚染リスクは高くなる。水道ネットワークの総人口・供給人口は、水道ネットワークの規模、生活排水量と関連する。

(5)水道地点情報

水道地点情報として、水道からの河川への距離、最寄り駅への距離、主要道路への距離、起伏、曲率、傾斜、傾斜方向を変数として含めた。最寄り駅への距離は、同じ都市内でもその地点が比較的都市部に位置するかどうかを表すと考えられる。

表 3-1 網羅的に考慮した説明変数

カテゴリー	変数名	説明	データの出典
水源	altitude	高度	the United Nations Office for the Coordination of Humanitarian Affairs(UN OCHA)
	source_ecoli	水源大腸菌濃度	Ministry of Health - Nepal
	ratio	表流水・地下水の割	Ministry of Health - Nepal

		合	
処理方法	ST	沈殿処理	Ministry of Health - Nepal
	RSF	急速濾過	Ministry of Health - Nepal
	FL	凝集処理	Ministry of Health - Nepal
	PF	加圧濾過	Ministry of Health - Nepal
	RF	簡易濾過	Ministry of Health - Nepal
水道管	oldest_pipe_age	最も古い水道管の年代	Ministry of Health - Nepal
	dist_syorizyo	浄水場への直線距離	Ministry of Health - Nepal
	pipelength	水道管総長	Ministry of Health - Nepal
	number_taps	水道数	Ministry of Health - Nepal
	pipelength_per_tap	水道管総長/水道数	Ministry of Health - Nepal
都市別	total_population	総人口	Ministry of Health - Nepal
	population_served	給水人口	Ministry of Health - Nepal
	popu-served	無給水人口	Ministry of Health - Nepal
	served/pipes	水道管長当たりの給水人口	Ministry of Health - Nepal
	(popu-served)/pipes	水道管当たりの給水人口	Ministry of Health - Nepal
	Days of no water	無給水日数/年	Ministry of Health - Nepal
	supply_hours	給水時間/日	Ministry of Health - Nepal
水道地点情報	up_down	起伏	UN OCHA
	curvature	曲率	UN OCHA
	inclination	傾斜	UN OCHA
	title_derection	傾斜方向	UN OCHA
	disto_river	河川への距離	UN OCHA
	disto_stations	駅への距離	UN OCHA
	disto_mainroad	主要道路への距離	UN OCHA

3.4. ステップ 2

ステップ 1 のモデルが過学習を起こしたため、ステップ 2 では、水源から水道への配水ネットワークを想定して説明変数を作成し追加するなどの改善を行い、その上で水道水中の大腸菌の存在に特に関連性が高いと思われる 8 変数に絞り込んだ（ステップ 2）。なお、この際、どの都市別の

説明変数にも対応づけられない水道があったため、サンプル数が減少し、大腸菌を含まない水道が 92 個、大腸菌を含む水道が 110 個となった。表 3-2 に最終的な説明変数を示す。

表 3-2 最終的な説明変数

Variable	Unit	Descriptions
Type of the water source	0: ground 1: surface	Type of the nearest water source whether surface or ground
Number of processors	processing	Number of processors in the water purification plant by each city
Amount of E.coli in the water source	CFU/100ml	Amount of E.coli in the nearest source
Length	m	Length from the nearest source or upstream tap
Days of no water	days/year	No water days in a year by each city
Number of buildings	building	Number of buildings within 100m of each path
Population density	people/km²	Population density connected to each rode nodes
Agriculture area	m²	Area of agriculture within 100m of each path

Type of the water source (最も近い水源の種類)

ステップ 1 の表流水・地下水割合から修正を行った。表流水・地下水割合では、都市別の水源の地下水割合をその都市内の全水道と紐づけたが、Type of the water source は、Dijkstra アルゴリズムにより各水道と各水源と紐づけ、各水道ごとに水源が表流水または地下水であるかの情報を紐づけた。ネパールの特徴として、水源として地下水が用いられることが多く、本研究で扱ったサンプルでも水源の半数が地下水であった。表流水が生活排水といった外部からの汚染を受けやすく、高い確率で大腸菌に汚染されているのに対して、一般的に地下水は汚染が進んでいない。このような水源毎の特徴を捉えるために本変数を含めた。

Number of processors (浄水場で使用される処理方式の数)

ステップ 1 では都市別の浄水場の変数として沈殿処理、 凝集処理、急速濾過、加圧濾過、簡易濾過がそれぞれ実施されているか否かを変数としていたが、Number of processors では、都市別の浄水場で実施される処理プロセスの数を変数としている。これは説明変数の数を減らす目的で行った。特に水源が表流水である場合は供給前に浄水場を通す。給水以前にどの程度水から大腸菌が取り除かれているかは、給水水質に直接関連する。都市毎に使用されている処理方式は異なり、都市別に大腸菌の除去能力が異なることが推察される。都市別に行われている処理プロセスの数を変数とすることで都市別の処理能力を捉える。

Amount of E.coli in the water source (最も近い水源の大腸菌の数)

ステップ 1 の水源大腸菌濃度では、水源と水道間の直線距離によって水源と水道の紐づけを行ったが、Amount of E.coli in the water source では Dijkstra アルゴリズムから各水源と水道間の距離を計算し、最も近い水源の大腸菌数を各水源の変数とした。ネパールでは、特に地下水を水源とした場合は浄水場を通さず、塩素処理をしたのちに直接給水を行う場合もある。また、浄水場を通したとしても、浄水場の設備・管理状況によっては必ずしも完璧に大腸菌の除去がなされるわけではない。したがって、水源の大腸菌の汚染は、給水される水の水質と関連がある可能性が高い。水源の大腸菌数を変数とすることで、水源の大腸菌汚染を捉える。

Length (最も近い水源から水道、または上流の水道から水道までの距離)

新たに追加した変数である。各水源と各水道間の距離を Dijkstra アルゴリズムによって計算し、各水道から最も近い水源への距離を各水道の変数として用いた。水道管の破損部からの外部汚染を考慮した場合、水道管を通る時間が長くなるほど汚染の確率は高まる。物理的に通る水道管が長くなると給水にかかる時間も長くなり、逆に通る水道管が短くなると給水にかかる時間も短くなる。水道管を通る時間を表す変数として、通過する水道管の長さを変数として含める。

Days of no water (無給水日数)

ステップ 1 と同様の変数である。ネパールでは、定期的に水供給の一時停止、断水が起こる。その間、水は水道管内に留まっていると想定すると水道管の破線部からの外部汚染のリスクが高まる。水道管内に留まる時間を表す変数として、無給水日数を変数を求める。

Number of buildings (建物の数)

新たに追加した変数である。各水源と各水道間の経路を Dijkstra アルゴリズムより取得する。。その後、直近の水源または上流の水道からの経路に対して幅 100m のバッファーを作成し、バッファー上に位置する建物の数を各水道の変数として扱った。水道管の破損部から発生する外部汚染の要因として生活排水があげられる。水道管周辺の建物の数は、周囲に住む人口及び排出される生活排水と関連がある。水道管周辺の生活排水量を表す指標として、建物の数を使用する。

Population density (人口密度)

新たに追加した変数である。各水源と各水道間の経路を Dijkstra アルゴリズムより取得する。次に経路を構成する交叉点と 1km メッシュの人口密度情報を持つ GIS データを重ね合わせる。各交叉点ごとに対応するメッシュは一つに決まるので、各交叉点と人口密度を紐づける。その後、直近の水源または上流の水道からの経路を構成する交差点に紐づけられた人口密度を平均し、各水道の変数として扱った。取得理由は建物の数と同様で、本変数は人口及び生活排水量を捉える変数として含める。建物の数と異なる点は、建物の数が各径路から作成された 100m のバッファ上の建物数をカウントしているのに対して、人口密度データは 1km 四方で作成されたメッシュデー

タを各径路に紐づけており、建物の数と比べてより広範囲の人口データを捉えることができる。

Agriculture area (農地面積)

新たに追加した変数である。建物の数を求めた際と同様に直近の水源または上流の水道からの経路に対して幅 100m のバッファーを作成し、バッファー上の農地の総面積を各水道の変数として扱った。ネパールの農耕地では肥料として堆肥が使用されており、堆肥に含まれる大腸菌も外部汚染の原因となる可能性がある。また、農耕地での家畜の利用も盛んであり、作業中に動物から排泄される糞便等も外部汚染の要因となると推察される。

3.5. 疑似水道ネットワーク

3.5.1. 疑似水道ネットワークの取得

いくつかの説明変数のデータを得るためには、水源から水道までの水道管に沿った最短経路を知る必要がある。しかし、発展途上国では水道管の位置情報は容易に入手できない。ネパールも例外ではないため、道路の位置情報を用いて、水道管が道路の下に埋設されていると仮定し、擬似的な水道ネットワークのデータを作成した。 JICA (2020)によれば、ポカラでは、水道管は幹線道路の下に埋設されており、新設する水道管も公道の下に埋設されると報告されている。このことから、水道管の位置情報を道路の位置情報で代用することは合理的であると考えられる。図 3-4, 5, 6, 7 は、最短経路の推定と経路から作成したバッファによる変数取得の流れを示す。まず、図 3-4 は水源・水道の位置情報、また the United Nations Office for the Coordination of Humanitarian Affairs(UN OCHA)が公開している道路位置情報を表示している。図 3-5 では、各道路の交差点を取得しており、これを以下、道路ポイントと呼ぶ。次に、図 3-6 では、各水源から各水道までの道路位置情報に沿った最短道路ポイントのリスト情報を Dijkstra アルゴリズムを用いて求める。Dijkstra アルゴリズムの説明は次節は示す。この道路ポインタのリスト情報から各水源から各水道への最短経路を抽出する。得られた最短経路をもとに、各水源から各水道までの長さ、最も近い水源の種類、最も近い水源が含む大腸菌の数を得る。 図 3-7 では、最短経路から 100m のバッファーを作成し、建物データ(UN OCHA)を重ね合わせ、各バッファと重なる建物の数をカウントしている。また、図 37 の処理と同様に、各バッファーに ICIMOD (Uddin et al. 2015)が公開している土地利用図を重ねて、バッファー上の農地面積を算出する。

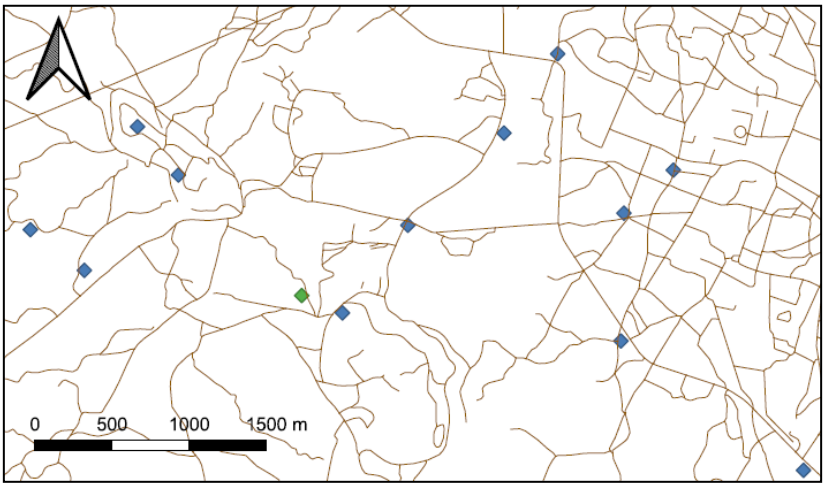


図 3-2 水道・水源、及び道路の位置情報

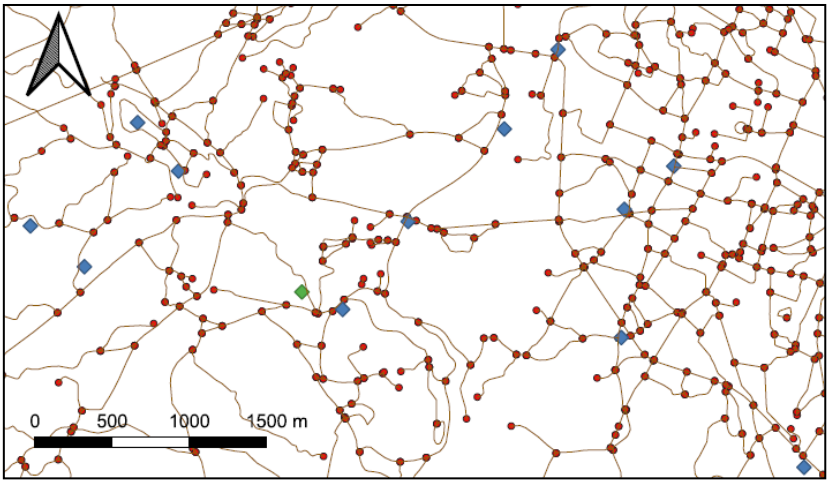


図 3-3 道路ポイントの取得

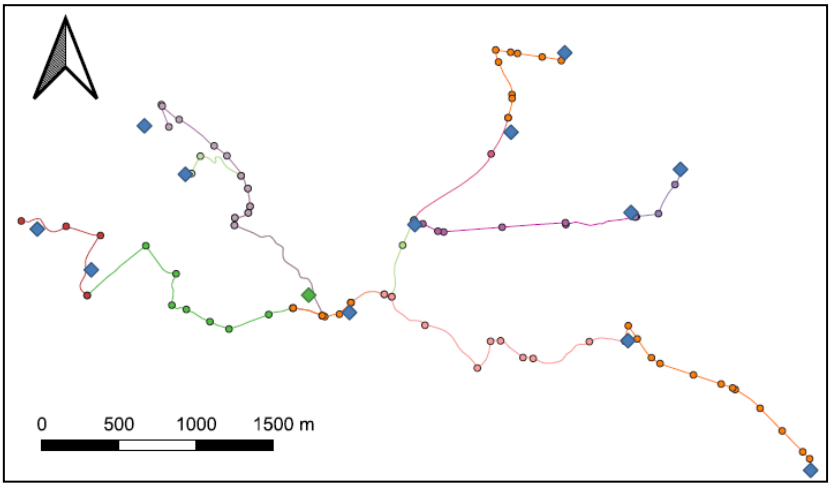


図 3-4 各水源と各水道を結ぶ最短経路の取得

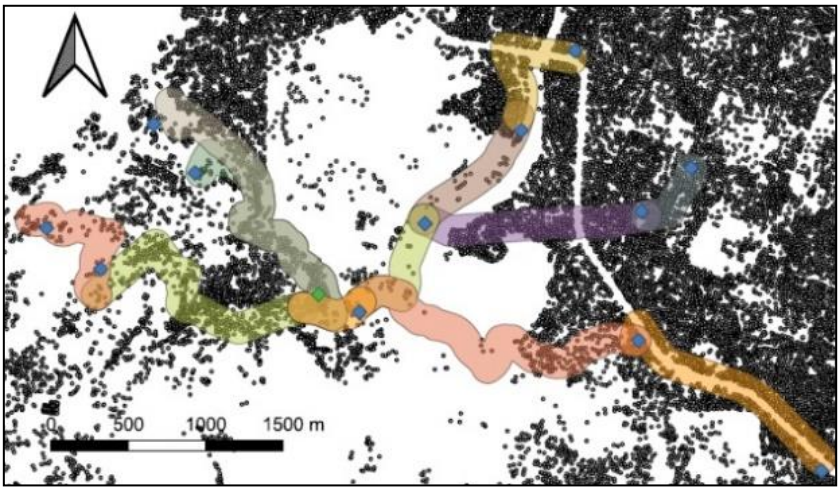


図 3-5 経路から作成した 100m バッファと建物情報

3.5.2. Dijkstra アルゴリズム

Dijkstra アルゴリズムはグラフ理論において説明される。本研究での道路網を使った疑似水道ネットワークの抽出では道路の交差点をノード、各ノードをつなぐ経路をエッジと定義する。また、各径路の長さをエッジが持つ重みをする。Dijkstra アルゴリズムでは、あるノードを始点とした際に、重みの和が最小となる他のノードへの到達方法を計算する。Javaid により(2013) Dijkstra アルゴリズムは分かりやすく説明されている。Dijkstra アルゴリズムは以下の操作で説明することができる。

1. ノードの集合を V とし、選択した始点からの最短経路が確定しているノードの集合を S と定義する。
2. 既に経路が判明しているノードのうち、始点からの距離が最小となるノードを S に追加する。
3. V-S に含まれるノードのうち、V と接するノードへの経路を既に求めた最短経路を元に計算し、最短距離が更新される場合は更新を行う。
4. V-S が空になるまで 2, 3 を繰り返す。

3.6. 標準化 (Standarization)

予測モデルを作成する前に、データの標準化を行う。標準化の役割は、すべてのデータの尺度を統一し、モデル作成時のバイアスを軽減することである。本研究が扱う説明変数でも日数や距離など複数の尺度が使用されており統一をする必要がある。本研究では、式(3.1)に示すように、Z-score standarization を使用した。

$$x_i^{'} = \frac{x_i - \mu}{\sigma} \tag{3.1}$$

3.7. 予測モデル

本研究では、予測モデルの手法として Linear Discriminant Analysis(LDA)、Logistic Regression (LR)、AdaBoost Random Forest(ARF)、SVM、RF、LightGBM を用いた。ステップ 1 での分析では LDA、LR、SVM、RF、ARF を、ステップ 2 の分析では SVM、RF、LightGBM を用いた。SVM は 1995 年に発表された手法で、超平面を最大化することにより回帰・分類を行う(CORINNA & VLADIMIR, 1995)。SVM は主に二値分類に用いられ、本研究の二値分類でも高い精度が得られると期待される。RF は 2001 年に発表された手法で、複数の決定木を組み合わせたものである(BREIMAN, 2001)。RF は高い精度が得られるが、データが小さいと過学習になることが多い。そこで、本研究で扱うデータに対して RF が過学習を起こさないかどうかを検討する。LightGBM は、2017 年に Microsoft が公開したアルゴリズムである(Ke et al., 2017)。

3.7.1. LDA

LDA は、データを線形変換によって直線上に縮約し、直線上でデータがどのクラスかを判別する手法である。以下、C1 と C2 の 2 つのクラスの二値分類を仮定し説明していく。線形縮約は式(3.2)のように求められる。

$$\boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{x} \tag{3.2}$$

y が縮約されたデータ、**x** が元データ、**w** が線形変換に相当する。クラスを最もよく分類する **w** を求めることが LDA の目的となる。最適な **w** を見つけるために以下の 2 つを検討する。

1. 射影先でのクラス毎の平均が離れるようにする。
2. 射影先でのクラス毎の分散が小さくなるようにする。

1. 射影先でのクラス毎の平均を計算し、なるべく離れるようにする。

C1、C2 の平均をそれぞれ m1、m2 と置く。この 2 点の線形変換後の距離は式(3.3)で表される。

$$\boldsymbol{w}^T m1 - \boldsymbol{w}^T m2 = \boldsymbol{w}^T(m1 - m2) \tag{3.3}$$

(3)の右辺を二次形式で表現すると式(3.4)になる。

$$\boldsymbol{w}^T(m1 - m2)(m1 - m2)^T\boldsymbol{w} \tag{3.4}$$

2. 射影先でのクラス毎の分散が小さくなるようにする。

C1 と C2 の射影先での分散s_1^2、s_2^2は式(3.5)、(3.6)で表せる。

$$s_1^2 = \sum_{n \in C1} (\boldsymbol{w}^T \mathrm{x1} - \boldsymbol{w}^T \mathrm{m1})^2 \tag{3.5}$$

$$s_2^2 = \sum_{n \in C2} (\boldsymbol{w}^T \mathrm{x2} - \boldsymbol{w}^T \mathrm{m2})^2 \tag{3.6}$$

s_1^2とs_2^2を足し合わせることでs^2を式(3.6)のように定義する。

$$s^2 = s_1^2 + s_2^2 \tag{3.7}$$

式(3.4)を式(3.7)で割った式を評価関数 J(**w**)と定義する。J(**w**)を最大化する **w** を求めることで、射影先のクラスが離れていて、かつ射影先のクラス毎のばらつきが小さい **w** を求める。

3.7.2. LR

図 3-8 は LR の流れを示している。**x** は説明変数、**w** は各変数に対する重みを表しており、初期はランダムに置かれることが多い。w_0 はバイアスである。これを踏まえると、y は以下の式(3.7)のように計算される。

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_m x_m \tag{3.8}$$

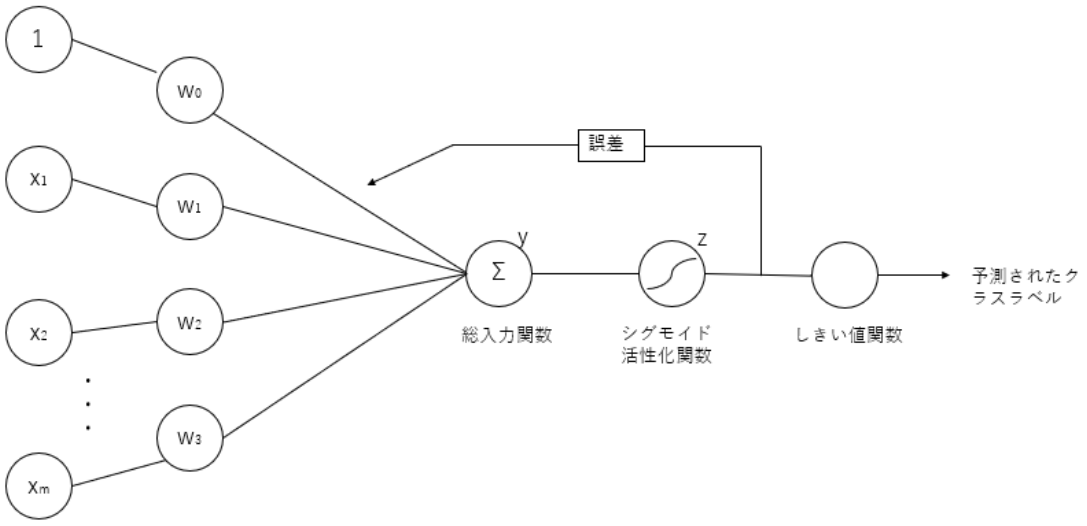


図 3-6 LR の流れ

y は各説明変数をそれぞれ重み付し、足し合わせた値である。次に、これをシグモイド関数に代入する。シグモイド関数は以下の式(3.8)で表され、図で現すと図 3-9 となる。

$$h(x) = \frac{1}{1 + \exp(-x)} \tag{3.9}$$

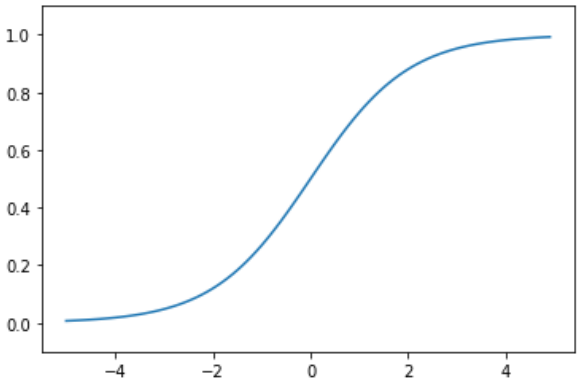


図 3-7 シグモイド関数

シグモイド関数は形状からもわかるように、代入された値を 0 から 1 の値に変換する。このシグモイド関数の出力は、サンプルがクラス 1 に属している確率と解釈することができる。ここでは 2 値分類の閾値を 0.5 と考える。すると、予測クラスラベル$\hat{L}$の出力は以下の式(3.10)となる。Z は y をシグモイドによって変換した値である。

$$\hat{L} = \begin{cases} 1 & z \geqq 0.5 \\ 0 & z < 0.5 \end{cases} \tag{3.10}$$

以上 LR によるクラスラベルの予測する方法を説明した。モデル構築では、予測された値と本来の値の誤差からコスト関数 J(**w**)を定義し、それを最小化するように重みを最適化していく。LR ではコスト関数 J(**w**)を尤度 j(w)の対数を取り、符号を負にすることによって定義している。尤度 j(**w**)、コスト関数 J(w)をそれぞれ式(3.9)、式(3.12)に示す。i はサンプル番号を示す。

$$\mathrm{j}(\mathbf{w}) = \mathrm{P}\,(\mathbf{L} \mid \mathbf{x};\ \mathbf{w}) = \prod_{i=1}^{n} \left(z^{(i)}\right)^{L^{(i)}} \left(1 - z^{(i)}\right)^{1-L^{(i)}} \tag{3.11}$$

$$J(\mathbf{w}) = \sum_{i=1}^{n} \left[-L^{(i)} log\left(z^{(i)}\right) - (1 - L^{(i)}) \log\,(1 - z^{(i)})\right] \tag{3.12}$$

3.7.3. ARF

ARF はブースティングの一種であるアダブーストのアルゴリズムを用いている。ブースティングとは弱学習機を繰り返し生成し、それぞれの結果を次の弱学習機の作成に反映さることで精度の高いモデルを作成する手法である。アダブーストは弱学習機によって正しく分類されたサンプルの重みを小さくし、誤って分類されたサンプルの重みを大きくすることで重みを更新していく。サンプル数を N とすると初期の重み wi は式(3.10)のように均等に与えられる。

$$w_i = \frac{1}{N} \tag{3.13}$$

弱学習機モデルT_jを式(3.11)のように生成する。**X** は説明変数、**y** は目的変数、**w** は重みである。

$$T_j = \text{train}\ (\mathbf{X}, \mathbf{y}, \mathbf{w}) \tag{3.14}$$

このモデルを元にクラスラベル$\hat{\boldsymbol{y}}$を予測する。

$$\hat{\boldsymbol{y}} = \text{predict}\ (T_j, \mathbf{X}) \tag{3.15}$$

y と$\hat{\boldsymbol{y}}$の整合性を見ることで、重み付された誤分類率 ε を計算する。

$$\varepsilon = \mathbf{w} \cdot (\mathbf{y} \neq \hat{\boldsymbol{y}}) \tag{3.16}$$

($\mathbf{y} \neq \hat{\boldsymbol{y}}$)は予測が正しい場合は 0、正しくない場合は 1 を出力する。εを使って重みの更新に使用する係数αjを以下式(3.17)のように定義する。

$$\alpha \text{j} = 0.5 \log \frac{1 - \varepsilon}{\varepsilon} \tag{3.17}$$

式(3.18)のように重みを更新し、式(3.19)により最後に重みを正規化して合計が 1 になるようにする。

$$\mathbf{w}: = \mathbf{w} \times \exp\left(-\alpha \text{j} \times \mathbf{y} \times \hat{\boldsymbol{y}}\right) \tag{3.18}$$

$$\mathbf{w}: = \frac{\boldsymbol{w}}{\sum_i w_i} \tag{3.19}$$

3.7.4. SVM

図 3-10 は SVM を図解したものである。

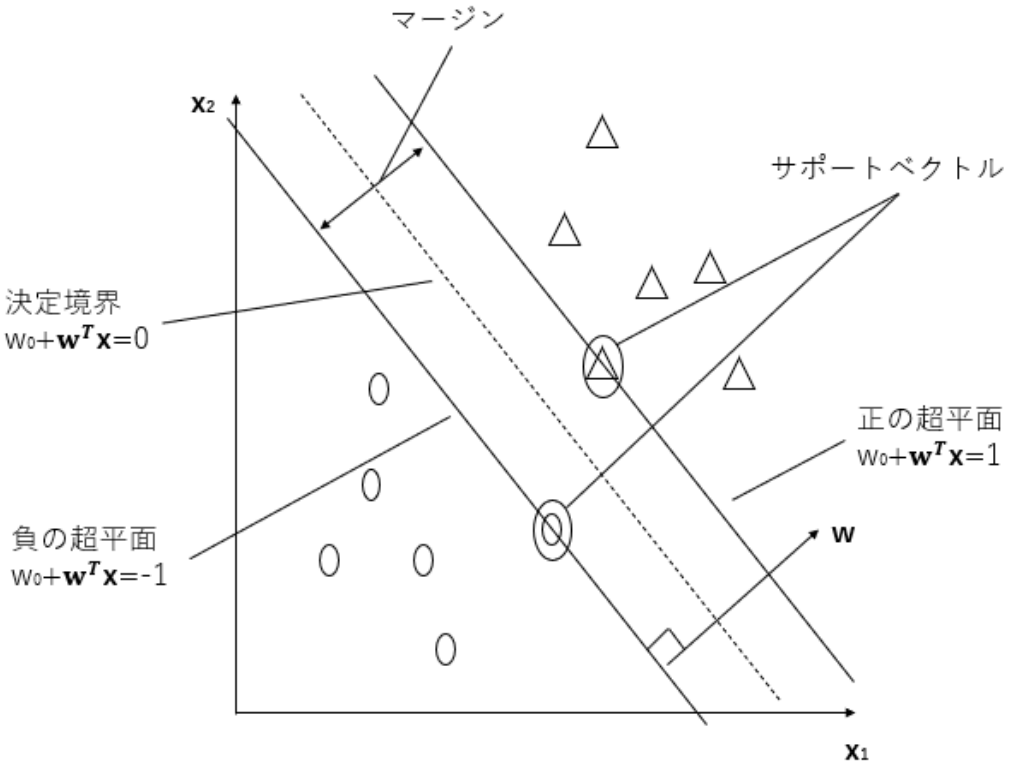


図 3-8 SVM 図解

ここでは、C1 と C2 の 2 つのクラス領域の分類を想定して説明する。SVM では，超平面と各クラス領域との最も近い距離、マージン M を最大化することで、2 つのクラス領域を分類する。超平面は式(3.20)のように求められる。

$$\boldsymbol{W}^T\boldsymbol{X}_i + w_0 = 0 \tag{3.20}$$

ここで、**W**(R ∈ n)は重みベクトルであり、超平面を定義するために最適化される。**X**i は各サンプル変数ベクトル(i ∈ {1, m})、w0 は超平面の定数値である。式(3.20)が満たすべき条件を、式(3.21)、(3.22)で定義する。

$$\mathrm{ti} = \begin{cases} 1 & \boldsymbol{X}\mathrm{i} \in K_1 \\ -1 & \boldsymbol{X}\mathrm{i} \in K_2 \end{cases} \tag{3.21}$$

$$\frac{\mathrm{ti}\,(\boldsymbol{W}^T\boldsymbol{X}\mathrm{i} + \mathrm{w0})}{\|\boldsymbol{W}\|} \geq M \tag{3.22}$$

式(3.22)を M で割ると式(3.23)になり、M̃ を用いてマージンを表すと式(3.24)のようになる。ここで、超平面は式(3.23)を満たしながら式(3.24)を最大化することで得られる。

$$t_i\left(\widetilde{\boldsymbol{W}}^T\boldsymbol{X}_i+\widetilde{w}_0\right)\geq 1 \tag{3.23}$$

where $\widetilde{\boldsymbol{W}}=\frac{\boldsymbol{W}}{M\|\boldsymbol{W}\|}$ $\widetilde{w}_0=\frac{w_0}{M\|\boldsymbol{W}\|}$

$$\widetilde{M}=\frac{t_i\left(\widetilde{\boldsymbol{W}}^T\boldsymbol{X}_i+\widetilde{w}_0\right)}{\|\widetilde{\boldsymbol{W}}\|}=\frac{1}{\|\widetilde{\boldsymbol{W}}\|} \tag{3.24}$$

計算を容易にするため、1/ ‖ $\widetilde{\boldsymbol{W}}$ ‖ の最大化は、式を変形することによって $\frac{1}{2}\|\widetilde{\boldsymbol{W}}\|^2$の最小化として計算される。式(3.25)の式(3.26)の ε とハイパーパラメータ C は、最小化計算における制約を緩和するために組み込まれる。

$$\frac{1}{2}\|\widetilde{\boldsymbol{W}}\|^2+C\sum_{i=1}^{n}\varepsilon_i \tag{3.25}$$

$$t_i\left(\widetilde{\boldsymbol{W}}^T\boldsymbol{X}_i+\widetilde{w}_0\right)\geq 1-\varepsilon_i \tag{3.26}$$

3.7.5. RF

RF とは、複数の決定木をブートストラップで取得したサンプルから作成し、それぞれの決定木を縮約して最適な分類器を作成する方法である。決定木は式(3.27)のように情報利得(IG)を最大化するように分岐を繰り返す.

$$\mathrm{IG}\left(D_p\right)=\mathrm{I}\left(D_p\right)-\sum_{j=1}^{m}\frac{N_j}{N_p}\mathrm{I}\left(D_j\right) \tag{3.27}$$

ここで、I は不純度、Dp と Dj はそれぞれ分岐前と分岐後のデータセット、Nj と Np はそれぞれ分岐前と分岐後のデータセットの数である。不純度の算出にはエントロピーとジニ係数を用い、それぞれ式(3.28)、(3.29)で示される。

$$\mathrm{IH}\ (\mathrm{t})=-\sum_{i=1}^{c}p(i\mid t)\ log_2 p(i\mid t) \tag{3.28}$$

$$\mathrm{IG}\ (\mathrm{t})\ = \sum_{i=1}^{c} p(i\ |t)\ (1 - p(i\ |t)) = \ 1\ - \sum_{i=1}^{c} p(i\ |t)\ 2 \tag{3.29}$$

ここで、p(i |t) はノード t においてクラス i に属するサンプルの割合である。

3.7.6. LightGBM

LightGBM は、gradient-based one-side sampling (GOSS) と exclusive feature bundling (EFB) を適用することで、GBDT (gradient boosting decision tree) の計算速度を向上させたアルゴリズムである(Ke et al., 2017)。GBDT は、実際の値と予測値の残差を、新しい木を繰り返し追加し学習しながら減らしていく。GBDT は高い予測精度を示すことで知られていたものの、学習速度が遅いという課題があった。GOSS、EFB は GBDT の高い予測精度を残しつつ、学習速度を改善するために導入されたアルゴリズムである。GOSS は、すべてのサンプルを使用するのではなく、学習すべき残差の大きいサンプルを抽出することで、GBDT の学習プロセスを高速化する。GOSS は残差の絶対値に従ってデータをソートし、上位 a/100(0 < a < 100)のデータを選択する。そして、残りのデータから b/100(0＜b＜100)のデータをランダムにサンプリングし、情報利得を計算する際にサンプリングしたデータを定数(1-a)/b で増幅させる。EFB は説明変数の候補を減らすことにより、学習速度を向上させる。いくつかの説明変数のデータがまばらな場合、これらの説明変数を 1 つの変数に統合する。

3.8. 評価指標

3.8.1. 層化 k 分割交差検証

本研究の評価指標としてまず使うのは正解率である。正解率は、本来のラベルに対して予測ラベルで正しく予測されたラベルの割合で求めることができる。しかし、トレーニングデータで正解率が高く出たとしても、そのモデルがよいと判断することはできない。サンプルの分け方や偏りから、偶然正解率が高くなることや、トレーニングデータに対してのみ過学習してしまっていることがあるからだ。ここで、未知のテストデータに対する識別能力、つまり汎化性能が高いことが求められる。本研究では、汎化性能を確かめる方法として層化 k 分割交差検証を用いる。層化 k 分割交差検証ではトレーニングデータをクラスの比率が均等になるように k 個に分割する。このうち k-1 個でモデルのトレーニングを行い、残りの一つで性能を評価する。これを k 回行い、k 個の性能評価を平均した値を出す。図 3-11 に層化 k 分割交差検証(k=10)を図式化した。

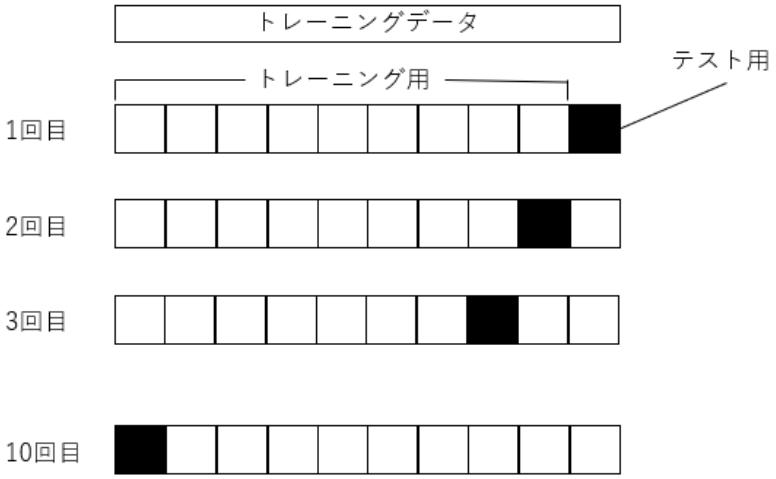


図 3-9 層化 k 分割交差検証(k=10)

n 回目の検証結果を E_n と置くと、層化 k 分割交差検証の結果 E は以下のように求められる。

$$\mathrm{E} = \frac{1}{10}\sum_{i=1}^{10} E_n \tag{3.30}$$

3.8.2. 混合行列

モデル評価として正解率が良く使われるが他にもモデルの評価指標がある。図 3-12 に混合行列を示す。

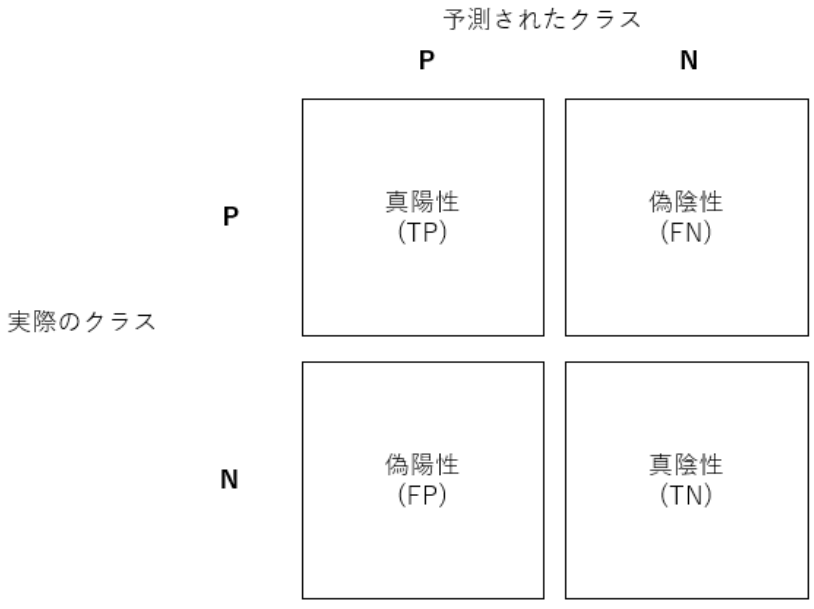


図 3-10 混合行列

混合行列は真陽性(TP)、真陰性(TN)、偽陽性(FP)、偽陰性(FN)からなる行列である。本実験の大腸菌の存在可否の分類を例に考えると以下のように表現できる。

TP…実際に大腸菌は存在し、予測でも存在すると判定されたサンプル。
TN…実際に大腸菌は存在せず、予測でも存在しないと判定されたサンプル。
FP…実際には大腸菌は存在しないが、予測では存在すると判定されたサンプル。
FN…実際には大腸菌は存在するが、予測では存在しないと判定されたサンプル。

さらに、TP, TN, FP, FN を使って、正解率(ACC)、誤分類率(ERR)、適合率(PRE)、真陰性率(TNR)、再現率・真陽性率(TPR)、偽陽性率(FPR)、偽陰性率(FNR)が定義される。

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.31}$$

$$ERR = \frac{FP + FN}{TP + FN + FP + TN} \tag{3.32}$$

$$PRE = \frac{TP}{TP + FP} \tag{3.33}$$

$$TNR = \frac{TN}{TN + FP} \tag{3.34}$$

$$TPR = \frac{TP}{FN + TP} \tag{3.35}$$

$$FPR = \frac{FP}{TN + FP} \tag{3.36}$$

$$FNR = \frac{FN}{FN + TP} \tag{3.37}$$

それぞれの解釈は以下のようになる。
ACC…大腸菌が存在するか否かが正しく分類された割合。

ECC…大腸菌が存在するか否かが正しく分類されなかった割合。
PRE…大腸菌が存在すると判定された中で実際に大腸菌が存在する割合。
TNR…本来大腸菌が存在しないサンプルで存在しないと判定された割合。
TPR…本来大腸菌が存在するサンプルで存在すると判定された割合。
FPR…本来大腸菌が存在しないサンプルで存在すると判定された割合。
FNR…本来大腸菌が存在するサンプルで存在しないと判定された割合。

3.8.3. ROC 曲線・AUC 値

ROC 曲線は閾値による TPR と FPR の変化を表現する。図 3-13 に ROC 曲線の例を示す。TPR は高いほうがよく、TPR を高くするためには少しでも大腸菌が入っていると疑われるサンプルは大腸菌が混入していると予測すること、つまり閾値を下げることになる。一方で FPR は低いほうがよく、FPR を低くするためには、大腸菌が混入していると確信できるサンプルのみ大腸菌が混入していると予測するので、閾値を上げることになる。TPR、FPR の両方を満たすモデルが良いモデルといえる。また TPR、FPR のどちらを優先するかは目的によって異なるため、目的に合わせて閾値を調整することが求められる。

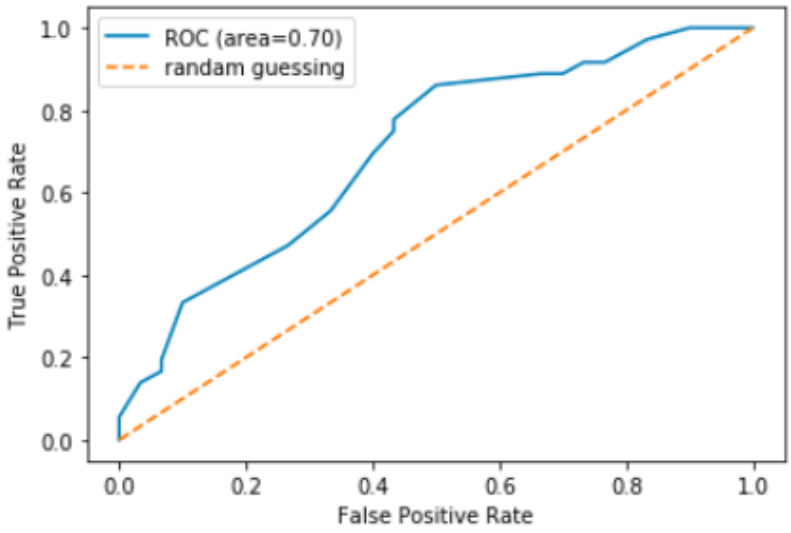


図 3-11 ROC 曲線

TPR が 1、FPR が 0 の状態、つまり図の左上が理想である。一方で点線は当て推量を実行した結果を ROC 曲線で表現したもので、モデルの精度がこれを下回った場合は使い物にならないことが分かる。ROC 曲線の曲線下面積(AUC)は分析モデルの性能を計るために使用される。

3.8.4. グリッドサーチ

グリッドサーチは、モデルを作成する際に人間が調整する必要があるハイパーパラメータを最適

化するための手法である。本研究では、ハイパーパラメータは SVM では C、RF と Light GBM では決定木の本数、各決定木の最大深さ、不純度の計算方法を設定した。指定されたハイパーパラメータのすべての組み合わせを網羅的に計算し、評価指標を最適化するハイパーパラメータを選択する。本研究では評価指標は正解率を使用する。

3.8.5. Shap

Shap はゲーム理論で使われるシャプレー値と同じで、機械学習や深層学習の説明力を高める手法として提案された(Lundberg & Lee, 2017)。この手法は、特定の説明変数を変化させたときの出力結果の変化を捉えることで、変数の重要度であるシャップ値を推定するものである。

3.8.6. RF による評価

RF による説明変数の評価を Feature importance と呼ぶ。Feature importance はある特徴量が分類にどれだけ寄与している指標ということができる。定義は以下の式で表すことができる。

$$\mathrm{I(j)} = \sum_{i=1}^{n \in F(j)} (N_{up}(i) \times G_{up}(i)) - (N_{left}(i) \times G_{left}(i) + N_{righit}(i) \times G_{righit}(i)) \qquad (3.38)$$

第4章　ステップ 1 での網羅的説明変数による分析結果

ステップ 1 での分析結果を示す。本章では予測モデルとして LDA、LR、ARF、SVM、RF を使用し、水道水質の予測における各予測モデルの有効性を検証する。また、本章ではテストデータを作成しておらず、サンプルをトレーニングデータとバリデーションデータに分割して使用する。218 のサンプルを 152 のトレーニングデータと 66 のバリデーションデータに分類した。バリデーションデータはトレーニングの過程を評価するために用いられるデータセットである。バリデーションデータでトレーニングデータによって作成されたモデルを評価することでパラメータを選択し、この結果得られたパラメータで層化 k 分割交差検証をトレーニングデータに対して適用することによってモデルの汎化性能を検証する。

4.1.　バリデーションデータでの正解率の最大化

予測モデルを評価する上で重要なのは未知のデータに対して正確に予測できるかどうかである。バリデーションデータの正解率が最大となるようにモデルの作成を行う。使用したモデルと、各モデルで指定したハイパーパラメータを表 4-1 に示す。

表 4-1 各モデルのハイパーパラメータ

モデル	ハイパーパラメータ
LDA	なし
LR	・C: 0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0
SVM	・C: 0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0
RF	・推定器の数: $1 \leqq x \leqq 50, x \in Z$
	・分割の基準: ジニ不純度またはエントロピー
	・各決定木の最大の深さ: $1 \leqq x \leqq 20, x \in Z$
ARF	・各決定木の最大の深さ: $1 \leqq x \leqq 20, x \in Z$
	・学習率: 0.5, 1.0, 1.5
	・推定器の数: $1 \leqq x \leqq 50, x \in Z$

4.1.1.　結果

バリデーションデータでの正解率で評価したところ、5 つのモデルから以下のような計算結果が得られた。結果を表 4-2 にまとめた。

表 4-2 各モデルの評価結果

LDA	LR	SVM	RF	ARF

P*	なし	C: 100	C: 100	・推定器の数: 23 ・不純度: エントロピー ・決定木の最大の深さ: 14	・決定木の最大の深さ: 4 ・学習率: 1.0 ・推定器の数: 5
Test	0.62	0.70	0.70	0.76	0.79
Validation	0.70	0.70	0.71	1	1

* パラメーター ** 層化 k 分割交差検証

4.1.2. 考察

バリデーションデータの正解率を比べるとARF>RF>SVM=LR>LDAとなる。これはRF、ARFは非線形モデルであることから、線形モデルより表現力が高いからだと考えられる。しかしRF, ARFではトレーニングデータの正解率が 100%であり、バリデーションデータの正解率との差が大きいことから過学習を起こしている可能性がある。RF、ARF はアンサンブル法を用いた比較的複雑なモデルであり、少量のデータに対して過度な適合をしていると考えられる。一方で LR、SVM についてはテストデータ、バリデーションデータの正解率が共に約 70%であり過学習が起こっていない。これは LR、SVM が線形判別を行う比較的単純なモデルだからだと考えられる。層化 k 分割交差検証の結果を見るとLDA、RF が最も高い 0.57 を示していることが分かる。

4.2. 過学習を考慮したモデル作成

RF、ARF で過学習が起こっていると考えられる。過学習が起こっているモデルは未知のデータに対応することができない。そこでパラメーターを調整することで過学習の抑制を試みる。指定したハイパーパラメータのすべての組み合わせでトレーニングデータでの学習及びバリデーションデータでの評価を繰り返し行い、バリデーションデータとトレーニングデータの正解率の差が 5%以上開くと過学習が起きているとし、過学習が起きていない中でバリデーションデータの正解率が最も高いモデルを選択する。

4.2.1. 結果

過学習を考慮して作成したモデルの評価をした結果を表 4-3 に示す。

表 4-3 LDA, LR, SVM の計算結果

	LDA	LR	SVM	RF	ARF
P*	なし	C: 100	C: 100	・決定木の数: 29 ・分割の基準: エントロピー ・決定木の最大の深さ: 3	・決定木の最大の深さ: 2 ・学習率: 1.5 ・決定木の数: 2
Test	0.62	0.70	0.70	0.71	0.67

Validation	0.70	0.70	0.71	0.7	0.70

* パラメーター　** 層化 k 分割交差検証

4.2.2.　考察

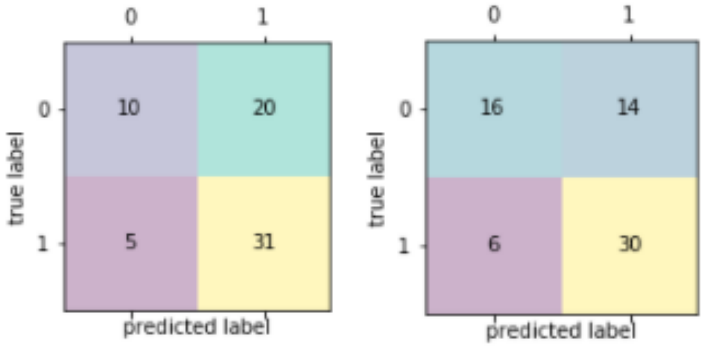
LDA, LR, SVM は 4.1 で得られたモデルを同じ結果になった。RF、ARF は過学習は抑制されたがバリデーションデータの正解率が下がった。RF、ARF 共にパラメータの決定木の最大の深さに着目すると、4.1 と比較して深さが浅くなっていることが分かる。これはモデルを単純化することでトレーニングデータへの過適合を抑制したと考えることができる。また RF と ARF は 4.1 と比較するとバリデーションデータの正解率は落ちているが、層化 k 分割交差検証で得られた結果は、RF では 0.57 から 0.59、ARF では 0.54 から 0.55 と大きくなっている。これは過学習が抑制されたことで、より汎化性能の高いモデルが作成されていると推察される。

4.3.　モデルの評価

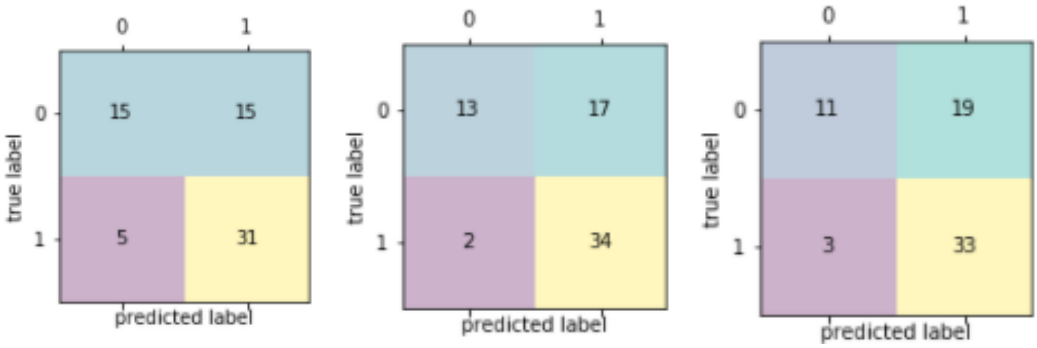
4.2 で作成したモデルのバリデーションデータに対する予測結果の、混合行列、ROC 曲線、AUC 値を求める。

4.3.1.　混合行列

図 4-1 に混合行列を可視化した図を示す。横軸が予測された結果、縦軸が実際の結果として表になっている。また表 4-4 に ACC、ERR、PRE 、TNR 、TPR 、FPR 、FNR をまとめた。



LDA　　LR



SVM　　RF　　ARF

図 4-1 混合行列

表 4-4 各評価指標まとめ

	LDA	LR	SVM	RF	ARF
ACC	0.6	0.70	0.70	0.71	0.67
ERR	0.38	0.30	0.30	0.29	0.33
PRE	0.61	0.68	0.67	0.67	0.63
TNR	0.33	0.53	0.50	0.43	0.37
TPR	0.86	0.83	0.86	0.94	0.92
FPR	0.57	0.39	0.43	0.53	0.58
FNR	0.14	0.17	0.14	0.06	0.08

4.3.2. ROC 曲線・AUC 値

図 4-2 に ROC 曲線を示す。また表 4-5 に AUC の値を示す。

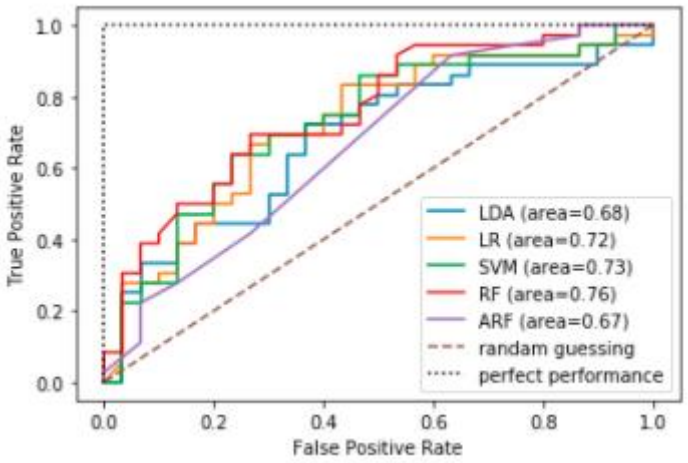


図 4-2 ROC 曲線

表 4-5 AUC の値

	LDA	LR	SVM	RF	ARF
AUC 値	0.68	0.72	0.72	0.75	0.67

4.3.3. 考察

表 4-4 に着目すると、RF の ACC が 0.71 と最も高くなっており大腸菌の存在確率を比較的正確に求められていることが分かる。PRE では LR が 0.68 と最も高くなっており、これは大腸菌が存在すると予測されたサンプルのうち、実際に大腸菌が含まれる確率が高いことを意味する。表 4-5 の AUC 値に注目すると、RF>SVM=LR>LDA>ARF の順に値が大きくなっており、AUC 値からは RF が最も評価されることが分かる。一方で ARF の AUC が最小になっていることが分かる。

4.4. RF による説明変数の評価

表 4-6 は 4.2.での RF による説明変数評価(上位 5 変数)の結果を表している。

表 4-6 RF による説明変数評価(上位 5 変数)

Variable	重要度
Days of no water	0.137
altitude	0.104
disto_syorizyo	0.090
disto_river	0.085
disto_stations	0.066

4.4.1. 考察

表 4-6 の結果から、RF では無給水日数、高度、浄水場への距離、河川への距離、最寄り駅への距離が予測結果に大きく寄与したことが分かる。無給水日数は浄水場から水道に水が供給されるまでに、水がどの程度の期間水道管内に留まっているかを表す変数として作用している。水は水道管に留まっている時間が長いほど外部から汚染を受ける可能性が高まり、大腸菌が混入する可能性も高まる。この結果から、水道管内にどの程度の期間、水が留まっているかは水道が大腸菌に汚染されているか否かを予測するのに重要な変数であると推察される。高度は表流水の水源を仮定した際に、その水源が上流、中流、下流のどこに位置するかを表していると考えられる。浄水場への距離は各水道別に浄水場への距離を求めた変数である。しかし、浄水場の位置情報は公式に発表された資料から得たものではなく Google map から検索することで取得したため、正しい位置を反映していない可能性がある。浄水場から水道への距離は水供給の間にどれだけ水が水道管内を通り大腸菌に汚染されるかのリスクを表していると考えられる。最寄り駅への距離は、その地点が他の地点と比べて相対的に都市または田舎かどうかを表す変数として作用していると推察される。

さらに、5 つの各変数についてウェルチの t 検定によって大腸菌の有無による平均値の違いを検定した。その結果の一部を表 4-7 に示す。p 値から浄水場への距離における大腸菌の有無による平均値の差は統計的に有意であるといえる。平均値を比べてみると、大腸菌が含まれない場合の浄水場への平均距離は 4317m、大腸菌が含まれる場合の浄水場への平均距離は 9421m であり、大腸菌が含まれる場合の方が 2 倍以上距離がある。これは、距離が長いほど水道管が破損している個所を通る可能性や、水道管に留まる時間が長くなることで汚染される可能性が高くなるからだと考えられる。有意水準 0.05 からは統計的に有意とは言えないが、有意水準を 0.1 と置いた際には無給水日数も大腸菌の有無によって平均値の違いがある。大腸菌が含まれない場合の日数は 20 日、大腸菌が含まれる場合の日数は 17 日であり、大腸菌が含まれない場合の方が日数が長い。これは日数が増えるほど水道管に水が留まる時間が長くなることを考えると、予想に反する結果であった。

表 4-7 p 値

Variable	p 値
Days of no water	0.082
altitude	0.192
disto_syorizyo	0.009
disto_river	0.054
disto_stations	0.220

4.5. 結論

4.5.1. 予測モデルの構築及び評価

発展途上国の少量のデータに対して SVM、RF、ARF といった機械学習による予測が、LDA、LR といった従来の予測手法による予測より優れているか否かを検証し、機械学習の有効性を検討した。バリデーションデータに対する予測精度最大化、過学習への適応、汎化性能の 3 つの観点で比較を行った。バリデーションデータの正解率では ARF>RF>SVM=LR>LDA の順に精度が良かった。過学習を抑えた状態での正解率ではその精度は RF>SVM=LR>ARF>LDA の順となった。交差検証による正解率では RF>ARF>SVM=LR>LDA の順に精度が良かった。バリデーションデータの正解率、交差検証による正解率から ARF, RF が他のモデルより相対的に優れた精度を示していることが分かる。一方で SVM は LR と常に同じ予測精度を示した。この結果から、ARF, RF モデルは従来の機械学習モデルより予測精度が高い可能性がある。しかし、バリデーションデータに対する予測精度最大化では過学習を示した。

4.5.2. 本章の課題

本章の予測モデルの構築の段階の課題として、説明変数の数が多いことがあげられる。本研究で扱う少量データに対して、多数の説明変数を組み込みモデルの作成を行うと、複雑な予測モデ

ルでは過学習を起こす可能性が高くなり、分析結果からも RF、ARF では過学習が見られた。さらに、変数の評価の際にも、変数間で多重共線性を持つ可能性が高まる。また、説明変数の重要度の解釈がしづらい点も課題としてあげられる。これは説明変数が水源から水道までの流れを捉えた変数となっていないことが原因の一つである。例えば、水道の位置する地点の高度や傾斜といった変数は、直感的には水道水内の大腸菌内の存在とは結び付かず、実際に RF, LR の変数評価でも重要な説明変数として扱われていない。変数選択の段階で目的変数と関連が高いと想定される変数を選択する必要がある。

次章では説明変数の改善、目的変数と関連の強い説明変数の絞り込みを行うことで、本章で得られた課題の解決を試みる。また本章の分析では層化 k 分割交差検証によりモデルの汎化性能を評価したが、次章ではテストデータを用意することでモデルの汎化性能を評価する。

第5章　ステップ 2 での説明変数改善後の分析結果

5.1.　説明変数の記述統計

選択した説明変数の記述統計量を表 5-1 に示す。カトマンズで行われている水質処理の平均が 5 つなのに対して、他の 25 都市では 0.49 である。都市別に行われている処理の数に差があることが分かる。またカトマンズの水源が含む大腸菌の平均数は 85.0 だが、他の 25 都市では 28.0 であった。カトマンズでは、生活排水などの要因で水源がより汚染されていることが分かる。無給水日数はカトマンズで 24.43 日、他の 25 都市で 13.0 日であり、カトマンズの方が無給水日数が多いことが分かる。カトマンズのような大都市で給水人口の増加、給水システムの煩雑化等から、継続的な給水が難しいと考えられる。建物の数、人口密度の比較ではカトマンズが値が大きくなっており、農地面積の比較では、その他の 25 都市の値が大きくなっている。

表 5-1 生データの記述統計

Variable	Kathmandu (n= 101)				25 cities except Kathmandu (n=101)			
	mean	std	min	max	mean	std	min	max
Number of processors	5	0	5	5	0.49	0.78	0	3
Amount of E.coli in water source	85.0	120	0	300	28.0	70.7	0	300
Length	1520	1210	0	6390	1360	1130	0	5621
Days of no water	24.3	0	24.3	24.3	13.0	13.4	0	56.8
Number of buildings	672	636	2	3070	367	445	0	2030
Population density	46500	37200	1570	120000	3190	2520	421	12400
Agrlcluture area	28200	73600	0	460000	66500	126000	0	562000

5.2.　各変数と水道水中の大腸菌数との相関関係

表 5-2 は、水道水中の大腸菌数と各変数の相関と p 値である。最も近い水源の種類は point-biserial correlation、その他の変数には pearson's corrlation を用いた。最も近い水源の種類、処理の数、人口密度の P 値は有意水準 0.05 未満であり、両者の相関は 0 とは異なると判断した。最も近い水源の種類の相関係数は 0.21 と正であるため、水源が表流水の場合、水道水中の大腸菌数は増加することが分かる。 ネパールの表流水の半分以上が大腸菌に汚染されているため (Government of Nepal, 2016)、ネパールの表流水はほぼすべて浄水処理が施されている。この結果から、既に汚染されている水源が十分に処理をなされないまま給水されている可能性があるこ

とが分かる。人口密度の相関係数は 0.28 と正であり、これは人口密度が高いほど水道水中の大腸菌数が多くなることが分かる。これは、人口密度が高いほど生活排水の排出量が多く、生活排水が水道管の破損部分から混入する等が発生する確率が高くなるためと考えられる。処理の数は相関値が 0.20 である。直感的には、処理数が多いほど大腸菌の汚染率は低くなると思われるが結果は逆であった。処理が多く行われている都市は大都市であり、水の供給量も多くなる。浄水処理を施す必要がある水の量が多く、処理が十分に行われずに供給が進んでしまっている等が推察される。このように、いくつかの変数は水道水中の大腸菌数と相関があることがわかったが、相関係数が小さいため、個々の変数から大腸菌汚染を予測することは困難である。また、カトマンズを除く 25 都市についても、水道水中の大腸菌数と各変数の相関と p 値を算出した。その結果、浄水場または上流水道からの距離は相関は 0 と異なるが、他の変数では相関がないことがわかった。全都市での相関の結果とカトマンズを除いた都市での相関の結果が大きく異なることから、カトマンズと他の都市で説明変数の傾向が異なることが分かる。しかし、本研究で使用するサンプル数が少ない(n=202)ことから、カトマンズとそれ以外の都市で別々に予測モデルを作成することは難しい。

表 5-2 各変数と水道水内の大腸菌数との関係

Variable	All citiy (26 cities) (n=202)		25 cities except Kathmandu (n=101)	
	Correlation	p-value	Correlation	p-value
Type of the water source	0.21	0.0022	0.11	0.29
Number of processing	0.20	0.0039	-0.0028	0.98
Population density	0.28	0.73e-05	0.047	0.64
Number of E.coli in water source	0.10	0.15	0.0049	0.96
Length	0.048	0.50	0.24	0.017
No water days	0.11	0.13	0.0025	0.98
Number of buildings	0.11	0.13	0.18	0.068
Area of agriculture	-0.054	0.45	0.090	0.37

5.3. 疑似水道ネットワークの妥当性

変数データを取得する際、作成した擬似水道ネットワークを用いた。推定した擬似水道ネットワークの妥当性を評価するために、取得した経路の上流と下流の大腸菌数を比較した。擬似水道ネットワークが適切に取得されていれば、下流側の経路の大腸菌数が上流側の経路の大腸菌数より少なくなることはないと考えられる。各水源から各水道までの最短経路を Dijkstra で計算すると、上流側の大腸菌数を含む経路が 59 本得られた。表 5-3 は、各経路の上流と下流の大腸菌の関係の割合を示したものである。カトマンズを除く 25 都市では、上流と下流に大腸菌を含む経路の割合が 0%であった。これは、上流で大腸菌が検出された経路には必ず下流でも大腸菌が検出され

ていることを意味し、現実的な結果であると考えられる。一方カトマンズでは、上流で大腸菌が検出され、下流で大腸菌が検出されなかった経路の割合が 26%であった。これは直感に反した結果であると言える。カトマンズにおける分析結果がうまくいっていない理由として、カトマンズの水道サンプル数が他の都市より多く密集していたことがあげられる。本研究では分析において各水源から各水道までの長さのみを考慮しているが、距離に関わらず区域によって給水範囲が決められることがある。したがってカトマンズのようにサンプルが密集している場合は、区域の境目に位置する水道サンプルの数が多くなり、距離のみでの分析では誤った水源と水道の紐づけが起こっている可能性がある。

表 5-3 上流と下流の大腸菌の関係

	Kathmandu	25 cities except Kathmandu
上流でも下流でも大腸菌を含む割合	24%	53%
上流で大腸菌を含む下流で大腸菌を含まない割合	26%	0%
上流で大腸菌を含まず下流で大腸菌を含む割合	24%	15%
上流でも下流でも大腸菌を含まない割合	26%	31%

5.4. 層化 k 分割交差検証での正解率

本研究では 202 の水道サンプルを用いて分析を行った。恣意的なデータの偏りが生まれないよう、excel 上では各都市の名前順でサンプルを扱い、分析の前には常にサンプルの並び順をランダムに変更した(常に同じランダムを使用)。学習では、全 202 サンプルの 73%にあたる 148 サンプルをトレーニングデータとして、残りの 54 サンプルをテストデータとして使用した。サンプルの分割にあたっては、トレーニングデータとテストデータにおける都市の割合ができるだけ同じになるようにした。グリッドサーチと層化 k 分割交差検証を用いたハイパーパラメータの選択では、分割後のデータが小さくなり過ぎないようにするため、分割数を 5 とした。グリッドサーチでハイパーパラメータを網羅的に設定した後、層化 k 分割交差検証で最も評価の高いハイパーパラメータを選択した。各モデルで設定したハイパーパラメータと選択されたハイパーパラメータを表 5-4 に示す。表 5-5 は、最も評価の高いパラメータを用いた各分割における予測値の正解率を示している。RF の平均値が 0.61 と最も大きく、汎化性能が比較的高いことが分かる。SVM の K-fold-3、RF の K-fold-5、LightGBM の K-fold-1 の精度が低くなっていることが分かる。これはサンプル数が少ないため、トレーニングデータ、テストデータ、またはその両方に偏りがあり、正しい学習ができなかったためであると考えられる。

表 5-4 ハイパーパラメータの設定と選択

モデル	段階	ハイパーパラメータ
SVM	設定	C: [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0]
	選択	C: 100
RF	設定	決定木の本数: [i for i in range(1, 101)], 不純度: ["gini", "entropy"], 各決定木の最大の深さ:[i for i in range(1, 21)],
	選択	決定木の本数: 5 不純度: entropy 各決定木の最大の深さ: 2
LightGBM	設定	決定木の本数: [i for i in range(1, 101)], 不純度: ["gini", "entropy"], 各決定木の最大の深さ:[i for i in range(1, 21)],
	選択	決定木の本数: 96 不純度: gini 各決定木の最大の深さ: 1

表 5-5 層化 k 分割交差検証と予測結果

Prediction models	**K-fold cross validation (148 samples)**						**Evaluation**			
	K-1	K-2	K-3	K-4	K-5	mean	Training (148 samples)	Test (54 samples)		
								All (54 samples)	Kathmandu (30 samples)	25 cities except Kathmandu (24 samples)
SVM	0.50	0.53	0.37	0.62	0.69	0.54	0.66	0.70	0.70	0.71
RF	0.70	0.53	0.77	0.58	0.48	0.61	0.66	0.61	0.67	0.54
LightGBM	0.43	0.50	0.70	0.55	0.69	0.55	0.70	0.57	0.43	0.75

5.5. 精度評価

残りの 54 サンプルをテストデータとして使用し、チューニングしたモデルの精度を評価した。テストデータは、大腸菌を含まない 24 の水道と大腸菌を含む 30 の水道で構成される。また、カトマンズにおける、大腸菌を含む 13 の水道、大腸菌を含まない 17 の水道、カトマンズ以外の 25 都市における大腸菌を含む 11 の水道、大腸菌を含まない 13 の水道のサンプルでの精度の確認を行った。表 5-5 にトレーニングデータとテストデータでの精度を示す。SVM の精度は 0.70 であり、54 サンプルにおいて最も優れていた。また、LightGBM は 25 都市では最も精度が良かったが、カトマンズでは最も悪かった。これは、モデルが 25 都市のデータに過学習し、その結果、カトマンズの

テストサンプルではうまく予測できなかったことを示唆している。SVM はカトマンズ、それ以外の 25 都市において同様に 70%前後の正解率を示した。

5.6. 結果の可視化

SVM モデルの probalibity の出力結果(予測結果を 0 から 1 までの値で返す。本研究の場合、大腸菌が含む確率と解釈できる。)と、実測結果を図 5-1 に示す。左図が実測値を表し、右図は予測値を表している。予測値では 50%を閾値に、大腸菌が混入している確率が 50%以上の水道には赤色を、50%以下の確率で汚染されている水道には青色を使用した。

まず全体的に実測地と予測値を比較すると、予測値では大腸菌を含む可能性が 50%以上であると予測された水道(赤色)が多くなっていることが分かる。実測図に示した範囲 a では大腸菌なしの水道が見られる。右図の予測結果の北東部を見てみると、大腸菌を含む確率が 50%以内と予測されている水道が 2 地点存在することが分かる。範囲 b については実測値で大腸菌が存在しており、予測地でも同様に大腸菌が混入している確率が高くなっていることが分かる。範囲 a, b では本件研究で作成した予測モデルで正しく予測ができている可能性がある。一方で範囲 c では、予測値ではすべての水道で大腸菌の混入確率が 50%以上になっているのに対して、実測値を見ると大腸菌が存在しない水道が複数みられる。このことからこの地点周辺の情報を予測モデルが正確に捉えることができていないことが分かる。範囲 d を見ると大腸菌が高い確率で混入していると予測された水道が多数存在する。同範囲の実測値には大腸菌なしの水道が存在するため、この地点についても正確に予測モデルが学習しきれていないことが分かる。このように機械学習の出力では、分類を行うのみではなく確率を出力をすることができ、割合を考慮に入れた考察が可能になる。

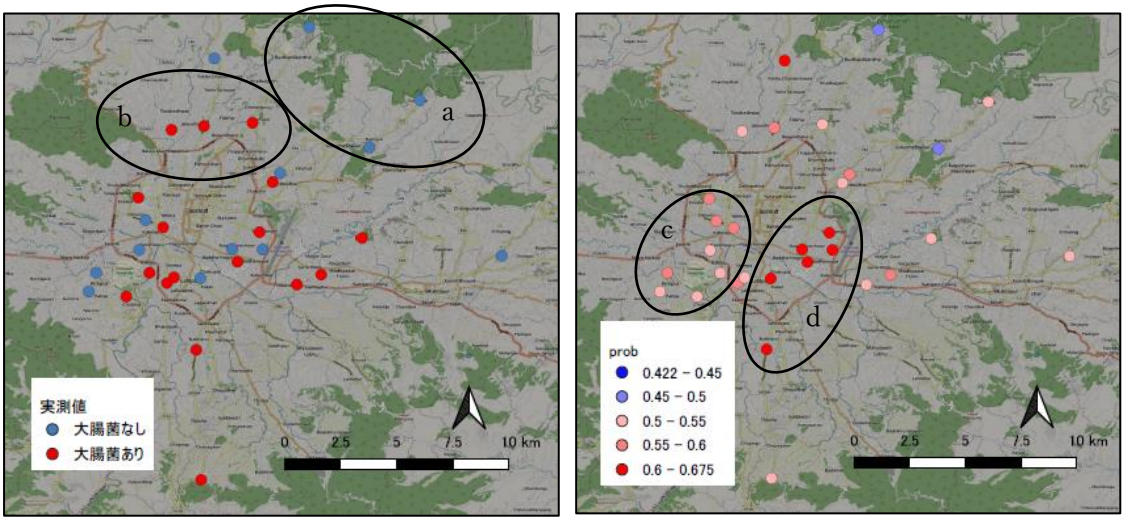


図 5-1 カトマンズでの結果の可視化(SVM)

5.7. 説明変数の評価

図 5-2 は、テストデータの 54 サンプルについて予測精度が最も高いモデルの SVM の説明変数を shap により評価したものである。縦軸は、各変数を重要度の高い順に上から並べたものである。横軸は shap の値を表している。各点は各サンプルを表し、精度評価にはテストデータを使用したため、各変数の点は 54 点となる。色は、各変数の値の高さと低さを表す。最も近い水源の種類、水源または上流の水道からの長さ、人口密度、最も近い水源の大腸菌数の値が大きいほど、水道に大腸菌が存在する確率が高くなることが分かった。このことから、水源が表流水であること、水源または上流の水道からの距離が長いこと、人口密度が高いこと、水源の汚染は、大腸菌が存在する可能性を高めることがわかる。また、無給水日数、建物の数、農地面積、処理数の値が大きいほど、水道から大腸菌が検出される可能性が低くなることがわかった。無給水日数については、水道管の圧力が低いと外部からの汚染が起こりやすいため、無給水日数が少ないほど良質な水を供給できることが多い。しかし、結果はその逆であった。これは、水道管の漏水はそれほど大きくなく、無給水日数の少ない都市ほど、断水回避を優先して不十分な処理水を提供する傾向にある可能性を示唆している。建物数については、建物数が少ない地域では、水道管や浄水場が適切に整備されていない可能性がある。農業の分野では、堆肥の使用が大腸菌汚染につながると想定されたが、分析結果からはそのようなことはよみとれなかった。図 5-3 は RF での shap による説明変数の評価を示している。無水給水日、水源または上流水道からの長さ、農地面積については SVM での shap による評価と同様な傾向を示している。一方で、浄水場での処理数については、値が大きくなるほど shap 値が大きくなっており、処理数が多くなるほど水道内の大腸菌数が増えると解釈することができる。また建物の数が多くなるほど、shap 値は小さくなっており、建物の数が増えると水道内の大腸菌数が減っていることが分かる。このようにモデルによって shap の評価が異なるのはモデルの構造の影響を受けるためである。その為、説明変数と目的変数の関係性の分析を目的におく場合は、機械学習による評価は参考程度に留め、統計による分析を行うことが必要である。

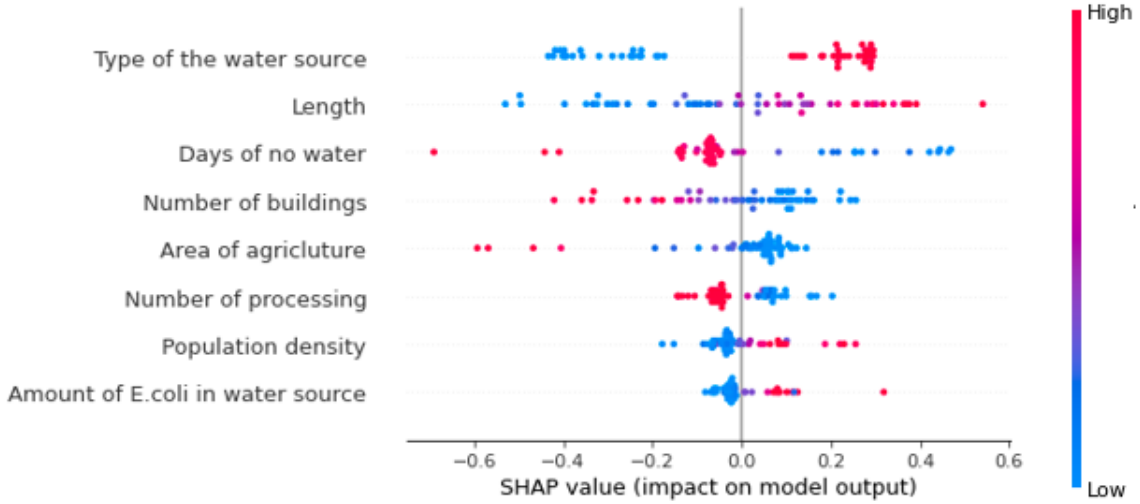


図 5-2 shap による評価(SVM)

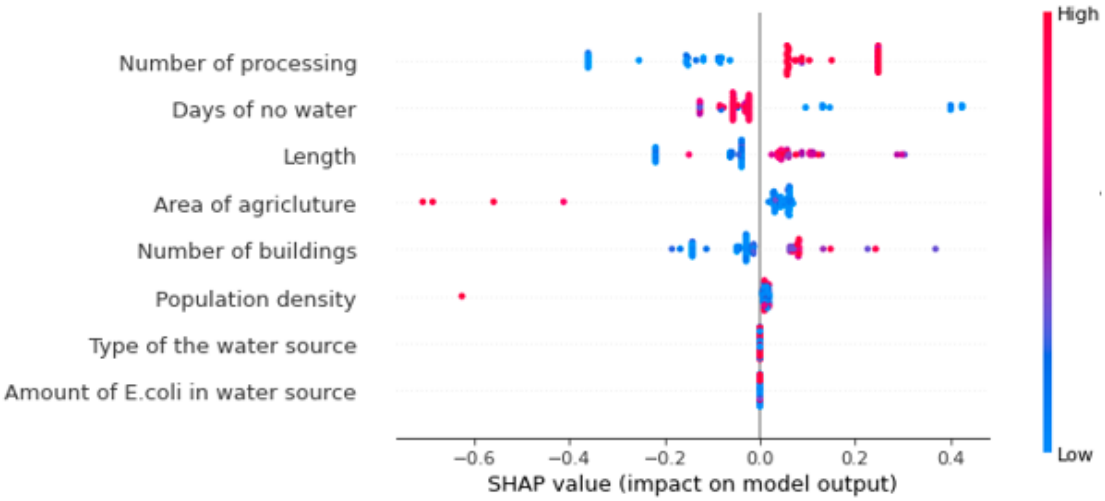


図 5-3 shap による評価(RF)

6.1.1.　本研究のまとめ

各変数と水道水中の大腸菌数の相関を分析した結果、最も近い水源の種類、処理の数、人口密度が水道水中の大腸菌数と相関する可能性があることが分かった。しかし、それらの相関値は小さく、個々の変数から大腸菌の汚染を予測することは困難である。機械学習による大腸菌の存在予測では、SVM モデルが水道水中の大腸菌の存在予測に対して 70%の精度を持つことが確認された。また、SVM を用いた shap による説明変数の評価から、最も近い水源の種類、水源または上流の水道からの長さ、人口密度、最も近い水源の大腸菌数の値が大きいほど大腸菌の存在確率が高くなることが分かった。疑似水道ネットワークについて、上流と下流の水道水中の大腸菌を比較し、妥当性を検証した。カトマンズを除く 25 都市の結果は、上流と下流の関係を正確に捉えていることから、水道ネットワークの分析が適切に行われている可能性を示した。一方、カトマンズの結果から、この分析が必ずしも実態を表していないことがわかった。カトマンズで分析がうまくいかなかった理由として、カトマンズから取得された水道サンプルは地理的に密集しており、区域等を考慮しない距離による分析では限界があったからだと考えられる。

6.1.2.　提言

本研究では得られた 70%の精度は、一般的な機械学習(ILSVRC など)が 90%以上の精度を達成するのに比べ、低い精度であった。しかし、データが取得できない状況を考慮すると、70%の精度の水道水水質予測は、発展途上国においてそれほど悪いものではないと考えられる。例えば、水道水を塩素で消毒する際に、大腸菌の混入が予測される地域から実施することで、より効率的に消毒を行うことができるなど、70%の精度のモデルを実用的に活用することができる。また本研究で示した技術は、TAIGER 社を例にあげたハイブリッド型 AI のように、現地で水道水質を管轄する方々の知見を組み合わせることで機能させることができる可能性がある。

6.1.3.　本研究の限界と今後の展望

実用化のためにはさらなる予測精度の向上が不可欠である。精度の低さの主な原因は、サンプル数が少ないことである。サンプルは、202 の水道水情報から成るが、通常機械学習が大量のデータを使用して学習を行うことを念頭に置くと、十分なサンプル数を確保することができていない。とはいえ、水道サンプルの取得は時間と費用がかかり、広範囲での水道サンプルの取得は容易でない。予測モデル精度向上を目的とした際に効率的にサンプル取得を行う方法として、本研究で示した予測と既存の知見を照らし合わせ、大きく結果が異なる地点からサンプル取得を行う方法があげられる。既存の知見と大きく結果が異なる場合は、予測モデルがその地域の情報を適切に学習ができていない可能性があり、その情報を新たに追加し学習をさせることで、予測モデルを効率よく学習させることができるからである。さらに、サンプル数が十分に集まった後は、カトマンズとそれ以外の都市で別々に予測モデルを作成することも効果的であると考えられる。また、説明変数の

絞り混みについても本研究では水質から水道までの流れから主観的に変数選択を行ったが、説明変数の評価等を照らし合わせることで客観的な変数選択が可能である。また、各水道水と水源との関係、各浄水場データ、水道ネットワークを入手できなかったことも限界の一つである。各水源と浄水場の結びつき、及び各浄水場データについては、時間がかかる水質検査等の調査が必要なわけではないので、比較的容易にデータを収集することができると考えられる。水道ネットワークについては、ネパールでそのようなデータが整備され公開されていないので、本研究で検討した疑似水道ネットワークの作成は有効な代替手段と考えられる。本研究では疑似水道ネットワークを各水源と水道の距離から作成したが、さらに傾斜、地域情報を追加することで、より正確な水道ネットワークを捉えられるだろう

GIS 分析に用いた主要コード

本研究での GIS 分析は主に python を使用した。Python では geopandas というライブラリを使用することでGISデータの分析を行うことができる。以下では geopandas における基本的な関数の確認と、疑似的水道ネットワークの分析で使用したコードを示す。

geopandas でできる基本的な処理

```
#まずは geopandas ライブラリを python にインストールをする。それ以外でも分析によ#く使用するライブラリをインストールする。
pip install geopandas
pip install Shapely

#python で geopandas を呼び出す。geopandas は名前が長いので私は gpd として扱う。
import geopandas as gpd

#gpd による shp ファイル gpkg ファイルの読み込みは read_file を使用する。
df = gpd.read_file(“ファイルの path”)

#座標の定義、座標の変換
df.crs = {'init': '定義したい座標'}
df = df.to_crs('変更したい座標')

#分割されているメッシュデータ等を結合して一つのデータとしたいときは shaply の
#union_cascaded を使う。
from shapely.ops import cascaded_union
# dfs: 複数のメッシュを持つデータ
# df: 結合後のデータ
df = gpd.GeoSeries(cascaded_union(dfs.geometry)

#バッファを作成したいときは buffer を使用することができる。
#df: geodataFrame 形式
df = df.buffer(作成したいバッファの大きさ)

#GIS データの重ね合わせを行いたいときは sjoin を使用する。
#df_a, df_b を重ね合わせたいとき。
df = gpd.sjoin(df_a, df_b, how=”inner”)
```

#ある GIS データの範囲で別の GIS データを切り抜きたいときは clip を使用する。
df = gpd.clip(df_a, df_b)

#以上に示した以外でも、最近棒の計算(点、線形)やメッシュ内の面積を求める等、簡単に実施することができる。QGIS でも同様な作業はできるが、同じ作業を繰り返し行う際は実行コードを書き一連の作業を自動で実行すると効率的である。

疑似水道ネットワーク分析で使用した最短経路分析

local_list: 対象地域名を含むリスト
district: 対象地域を含む map データ
tap: 水道の位置情報
source: 水源の位置情報
road: 道路の位置情報

```
for local in local_list: #地域ごとに処理を行う
    # 対象地域データの切り抜き
    map = district[district['LOCAL'] == local]  ＃対象地域の map の取り出し
    tap_clipped = gpd.clip(tap, df_dis)  ＃対象地域の水道位置情報の切り抜き
    source_clipped = gpd.clip(source, df_dis)  ＃対象地域の水源位置情報の切り抜き
    road_clipped = gpd.clip(road, df_dis)  ＃対象地域の道路位置情報の切り抜き

   ＃各切り抜きデータの保存
    gpd.GeoDataFrame.to_file(tap_clipped, …)
    gpd.GeoDataFrame.to_file(source_clipped, …)
    gpd.GeoDataFrame.to_file(road_clipped, …)

    # 道路を形成するポイントデータの始点と終点を抜き出す
    road_id_list = []  ＃道路の id を追加
    raodpoint_id = 0  ＃振り分ける id
    roadpoint_id_list = []  ＃取り出した始点と終点を追加
    roadpoint_list = []  ＃始点と終点の位置情報

    with fiona.open(“保存した road_clipped”) as lines:
        for line in lines:
        road_id_list.append(line['properties']['index'])  ＃road_id の追加
```

```
            roadpoint_list.append(str(Point(line['geometry']['coordinates'][0])))  ＃始点の追加
            roadpoint_id_list.append(raodpoint_id)  ＃始点の id をリストに追加
            raodpoint_id += 1

        road_id_list.append(line['properties']['index'])  ＃road_id の追加
            roadpoint_list.append(str(Point(line['geometry']['coordinates'][-1])))  ＃終点の追加
            roadpoint_id_list.append(raodpoint_id)  ＃終点の id をリストに追加
            raodpoint_id += 1

＃取得したリストから GeoDataFrame の作成
df = pd.DataFrame({'id': id_list, 'index': index_list, 'geometry': geopoint_list})
df = gpd.GeoDataFrame(df, geometry=[loads(wkt) for wkt in df['geometry']])

# 位置情報(X, Y)が同じポイントデータのリストを作成する
df['X'] = df['geometry'].x
df['Y'] = df['geometry'].y
dup_df = pd.DataFrame(df.groupby(['X', 'Y']).agg(id=('id', lambda x: list(x.value_counts().index))))
dup_list = dup_df['id'].values

# 同じ位置のポイントデータのリストから一つ目のみを取り出す
unique_point_list = []
for point_list in dup_list:
    unique_list.append(point_list[0])
unique_df = df.set_index('id').iloc[unique_list, :].reset_index()[['id', 'index', 'geometry']]

# 2 点と距離の情報
start_id_list = []  #始点の id を追加
end_id_list = []  ＃終点の id を追加
length_list = []  ＃始点と終点の間の距離
road_id_list = []  #道路の id を追加
raodpoint_id = 0  ＃振り分ける id

with fiona.open(“保存した road_clipped”) as lines:
    for line in lines:
        start_id_list.append(raodpoint_id)  ＃始点の id
        raodpoint_id += 1
```

```
            end_id_list.append(raodpoint_id)  ＃終点の id
            raodpoint_id += 1
            length_list.append(line['properties']['LENGTH'])  ＃始点と終点の間の距離
            road_id_list.append(line['properties']['index'])  ＃道路の id
    distance = pd.DataFrame({'start_id': start_id_list, 'end_id': end_id_list, 'length': length_list, 'index': index_list})

    # 同じ位置にあるポイントの id をリストの一番初めにあるポイントの id に統一する
    #重なるポイント id→一番初めのポイント id  の辞書を作る
    dic = {}
    for point_list in dup_list:
        for point in point_list:
            dic[point] = point_list[0]
    distance['start_id'] = distance['start_id'].apply(lambda x: dic[x])
    distance['end_id'] = distance['end_id'].apply(lambda x: dic[x])

    #隣接行列作成
    node_num = len(road_id_list)
    INF = 0
    W = [[INF] * node_num for _ in range(node_num)]
    for i in range(distance.shape[0]):
        W[distance.loc[i, 'start_id']][distance.loc[i, 'end_id']] = distance.loc[i, 'length']
        W[distance.loc[i, 'end_id']][distance.loc[i, 'start_id']] = distance.loc[i, 'length']

    data = np.array(W)
    G=nx.from_numpy_matrix(data)

    # 水源から一番近い道路ポイントの id リストを取り出す
    source_nearest_list = ckdnearest(source_clipped, unique_df)[['Code', 'id']]
    # 水道から一番近い道路ポイントの id リストを取り出す
    tap_nearest_list = ckdnearest(tap_clipped, unique_df)[['Rowid', 'id']]

    # 以下の関数で最短経路及び距離を計算することができる
    nx.shortest_path(G, source="始点 id", target="終点 id", weight='weight')
    nx.shortest_path_length(G, source="始点 id", target="終点 id", weight='weight')
```

追加分析

Gaussian naive Bayes (GNB)

アルゴリズム

追加分析として、Gaussian naive Bayes (GNB)での予測を行った。追加分析を行った理由として、複数の先行研究で予測モデルとして使用されていたこと、また一般的に少量データに対しても効果的であるといわれているからである。GNB はベイズによる分類器の一つであり、説明変数がガウス分布に従くことを仮定している。よってある変数 x = v が与えられたとき、クラス C に属する確率は以下のように表すことができる。

$$P(x = v|\, c) = \frac{1}{\sqrt{2 \times \pi \times \sigma_c^2}} \times e^{-\frac{(v-\mu_c)^2}{2 \times \sigma_c^2}} \tag{1}$$

GNB では、与えられたサンプル、つまり test data がどのクラスに属するかを式(2)で表すことができる。

$$P(y|\, x_1, \ldots, x_n) = \frac{P(y) \times \prod_{i=1}^{n}(x_i|y)}{P(x_1,\ \ldots,\ x_n)} \tag{2}$$

ここで、分母の$P(x_1,\ \ldots,\ x_n)$は固定値なので式(3)のように変形することができる。

$$P(y|\, x_1, \ldots, x_n) \ \propto\ P(y) \times \prod_{i=1}^{n}(x_i|y) \tag{3}$$

x_i (i ∈ {1, n})は説明変数を表し、$P(y|\, x_1, \ldots, x_n)$はあるサンプルの説明変数を過程したときに、そのサンプルがどのクラスに属するかの確率を表している。P(y)は事前分布であり、変数が与えられる以前のどのサンプルに属するかの確率である。通常は P(y)はサンプルの割合から計算され、例えばサンプル数 100 の C1 とサンプル数 200 の C2 の 2 値分類を行う際、C1 に属する確率は 1/3、C2 に属する確率は 2/3 となる。$(x_i|y)$はとトレーニングデータから得られるある変数の確率分布における、テストデータの変数の尤度を意味する。よって右辺は、事前部分布に尤度を掛け合わせていると解釈できる。以上の計算によって各クラスに分類される確率(事後確率)を計算することができる。最後に、最も高い確率だと予測されたクラスを予測として出力する。

$$\hat{y} = argmax_y\, P(y) \times \prod_{i=1}^{n}(x_i|y) \tag{4}$$

学習結果

テストデータとトレイニングデータは同様なものを用いた。テストデータに対する予測精度は 55%とり、SVM, RF の予測精度と比較すると低い予測精度となった。

参考文献

Abyaneh, H. Z. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*, *12*(1), 1–8. https://doi.org/10.1186/2052-336X-12-40

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Environmental Chemistry of Water Quality Monitoring*, *11*(11), 1–14. https://doi.org/10.3390/w11112210

BREIMAN, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1007/978-3-030-62008-0_35

Budhathoki, C. B. (2019). Water Supply, Sanitation and Hygiene Situation in Nepal: A Review. *Journal of Health Promotion*, *7*(June), 65–76. https://doi.org/10.3126/jhp.v7i0.25513

Capacity Assessment of Water Supply System Managed by NWSC , WSMB and KUKL. (2016). *December*.

CORINNA, C., & VLADIMIR, V. (1995). Support-Vector Networks. *IEEE Expert-Intelligent Systems and Their Applications*, *7*(5), 63–72. https://doi.org/10.1109/64.163674

Corporation, H. A. (2015). *Project Commissioned by the Ministry of Health , Labour and Welfare Project to Provide Planning Guidance for the Water Supply Project Bhaktapur Water Supply Improvement Project in the Federal Democratic Republic of Nepal Study Report*. *March*.

Health, W. (2005). *Ministry of Physical Planning and Works Singhadarbar kathmandu National Drinking Water Quality Standards , 2005 Implementation Directives for National Drinking Water Quality Standards , 2005 Government of Nepal Notice issued by Ministry of Physical Planni*.

Javaid, M. A. (2013). Understanding Dijkstra Algorithm. *SSRN Electronic Journal*, *January 2013*. https://doi.org/10.2139/ssrn.2340905

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Conference on Neural Information Processing Systems (NIPS 2017)*. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Koju, N. K., Prasai, T., Shrestha, S. M., & Raut, P. (2015). Drinking Water Quality of Kathmandu Valley. *Nepal Journal of Science and Technology*, *15*(1), 115–120. https://doi.org/10.3126/njst.v15i1.12027

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NIPS 2017*.

Memon, N. A., Unar, M. A., & Ansari, A. K. (2016). *pH Prediction by Artificial Neural Networks for the Drinking Water of the Distribution System of Hyderabad City*. *31*(1), 137–146. http://arxiv.org/abs/1604.00552

Ministry of Health, L. and W. (2016). *Capacity Assessment and Benchmarking*. *72*. https://www.jica.go.jp/nepal/english/office/others/c8h0vm0000bjww96-att/publications_06.pdf

Ministry of Health Ramshah Path, Kathmandu Nepal, K. N. (2016). *NEPAL DEMOGRAPHIC AND HEALTH SURVEY*. *636*. https://doi.org/10.1080/19485565.1967.9987700

Nepal, W. (n.d.). *Wateraid in Nepal(2011) Report-People ' s perception on sanitation : Findings from Nepal*. http://www.wateraid.org/~/media/Publications/perception-sanitation-nepal.ashx

Ogata, R., Khatri, N., & Sakamoto, M. (2019). Illuminating utility benchmarking data with analysis and consumer feedback – Insights from Nepal. *Journal of Water Sanitation and Hygiene for Development*, *9*(2), 356–362. https://doi.org/10.2166/washdev.2019.159

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Singha, S., Pasupuleti, S., Singha, S. S., Singh, R., & Kumar, S. (2021). Prediction of groundwater quality using efficient machine learning technique. *Chemosphere*, *276*, 130265. https://doi.org/10.1016/j.chemosphere.2021.130265

Uddin, K., Shrestha, H. L., Murthy, M. S. R., Bajracharya, B., Shrestha, B., Gilani, H., Pradhan, S., & Dangol, B. (2015). Development of 2010 national land cover database for the Nepal. *Journal of Environmental Management*, *148*, 82–90. https://doi.org/10.1016/j.jenvman.2014.07.047

Warner, N. R., Levy, J., Harpp, K., & Farruggia, F. (2008). Drinking water quality in Nepal's Kathmandu Valley: A survey and assessment of selected controlling site characteristics. *Hydrogeology Journal*, *16*(2), 321–334. https://doi.org/10.1007/s10040-007-0238-1

WHO/UNICEF. (2019). Progress on household drinking water, sanitation and hygiene 2000-2017. *Unicef/Who*, 140. https://washdata.org/sites/default/files/documents/reports/2019-07/jmp-2019-wash-households.pdf

WHO, & Unicef. (2000). Global Water Supply and Sanitation Assessment 2000 Report. *Water Supply*, 87. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Global+Water+Supply+a

nd+Sanitation+Assessment+2000+Report#0

日本語文献

JICA. (2020). ネパール国ポカラ上水道改善計画準備調査.

日本経済新聞. (2021) シンガポールの新興タイガー、「データ×知能」AI で革新　機械学習の弱点克服　安価で正確

日本経済新聞. (2021) フラクタ、AI で水処理のコスト削減　栗田工業と

謝辞

本研究を進めるにあたり、多くの方々ご指導とご支援を賜りました。ここに感謝の意を述べさせていただきます。

本研究の指導教員である、東京大学大学院新領域創成科学研究科国際協力学専攻 准教授 坂本麻衣子先生には、研究の方針をはじめとして、技術的なご指導、数多くのご助言を賜りました。特に論文の執筆指導において、執筆経験の少ない筆者の原稿を、何度も見直していただきました。多大な時間を要する作業だったかと存じます。また、紹介いただいた企業様では、ご縁もあり長期インターンを行うこととなり、貴重な体験をさせていただくことができました。心から感謝を申し上げます。

研究室の先輩であり、現在 JICA に勤められている緒方隆二さんには、本研究で使用したサンプルデータを提供いただき、さらにネパール現地での経験から的確なアドバイスをいただきました。深く感謝いたします。

副査を務めていただいた東京大学大学院新領域創成科学研究科国際協力学専攻 教授 本田利器先生、同じく副査を務めていただいた東京大学大学院新領域創成科学研究科国際協力学専攻 准教授 吉田 貢士先生には、ご多忙中にもかかわらず貴重なお時間を割いていただきました。誠にありがとうございます。

また研究室の先輩や同輩、後輩のみなさんとは、研究活動において互いに意見を交換し、また励まし合い、切磋琢磨しながら研究を進めることができました。大学院生活を有意義なものにしてくださった研究室の皆様はじめ専攻の皆様にも心より感謝致します。

そしてこの研究をきっかけに本学において様々な方と議論を交わし自分の知見をより深め、様々な物事の見方を学びました。そのきっかけを与えたくれた本学で出会った全ての方へ感謝を申し上げます。