平成24年度 卒業論文

ICカード打刻データと修学データを用いた
学生の将来の学習レベル予測と特徴分析

指導教員
舟橋　健司　准教授
伊藤　宏隆　助教

名古屋工業大学　工学部　情報工学科
平成21年度入学　21115011番

名前　伊藤 雄真

目 次

第1章　はじめに

名古屋工業大学では, 早期の修学指導を行うための双方向型支援システム構築を目的とし, コースマネジメントシステムと IC カード出欠管理システムを導入している [1]. コースマネジメントシステムは情報技術やインターネットを使った e-learning を支援するシステムであり, 教材の作成支援や課題の提出管理, 小テストの実施, 学生の受講管理を行う機能を持っている. IC カード出欠管理システムは, IC カード化された学生証を出席の際に講義室に設置された IC カードリーダにかざすことにより, 打刻情報（ID m, 打刻時間, 打刻 IC カードリーダ番号）がリアルタイムで出欠管理サーバに送信され, 蓄積される.

学習指導において, 教員 1 名あたりが受け持つ学生数の多い大学では教員の負担が多くなる. また社会の変化に伴い, 学生の目的意識や興味が多様化している近年は同一の入学試験を経てきたとしても, 成績不振学生や授業に満足できない吹きこぼれ学生などが目立つようになる. 従来このような学生の指導は成績が出た後で行っていたが, 授業への出席率が極端に悪い学生の場合, 手遅れであることは明白である. そこで名古屋工業大学では, コースマネジメントシステムや IC カード出欠管理システムで得られた情報を用いている. これにより出席率が極端に下がった学生を早期に発見し, 担当部局に連絡することや, データマイニングの手法により早期にリスク予知し対応することで, 成績不振学生を見つけるができる. また成績優秀な学生にはより高度な課題を与え学習意欲の向上を図ることができる [2].

過去の関連研究では, 1 つの授業における学生の出欠状況や課題提出状況が成績に影響を与えることが証明されている. それらを用いることで成績予測が可能であること [3][4] や, 集められたデータをより有効に使うために学生が Web 上で予測結果を確認できる環境を期待し, Web でも実装可能なニューラルネットワークに着目した研究 [5] などか行われている. またニューラルネットワークでは計算過程を知ることができないことから説得力に欠けてしまうことを危惧し, 出力結果の直感的な分かりやすさと計算過程を参照できるベイジアンネットワークに着目した研究 [6] が行われている.

従来の出欠データは出席, 早退, 遅刻, 欠席という情報しか使っていなかったが, 本来 IC カード出席管理システムでは打刻情報が蓄積されており, 誰がいつ何時何分何秒に打刻したかまで知ることができる. これを用いて成績予測することにより成績優秀者の行動パターンを知ることや成績不振者の早期発見に繋がることを期待し, 打刻データに着目した. 打刻データには仮の学生番号と受講教室, 打刻日, 打刻時間の情報が含まれている. 打刻時間だけから成績予測するにはデータが限られており, それだけでは成績予測は厳しいと考えられるので, 打刻データを拡張して利用する. 例えば, ある日に打刻時間が記録されていない場合は, その日は出席していないと予想することで欠席したという結果が得られる. これを利用し欠席回数を求めることができる. ただし, この例において出席していないと考える以外にもその日に授業が無いということも考えられるので厳密な欠席回数とはならないことに注意する必要がある.

本研究では, 学習データとしてある年度入学に入学した学生の 1 年次と 2 年次, 2 年分の成績データと打刻データを用いている.1 年終了時点と 2 年終了時点において打刻データがどの程度成績予測に有効かを決定木とベイジアンネットワークを用いることで検証する. 決定木はデータマイニングの中でも最もメジャーな手法の一つであり, ある要素の値がある一定値を超えるか超えないかで 2 つに分類していく. その結果が木の根構造のように見えることから分岐点の各要素に打刻データが使われているかを確認することができ, これによりどの要素が成績予測に関わっているかを確認することができる. また, ベイジアンネットワークは専門的な知識をもとに各変数の因果関係を矢印に見立て有効グラフを構築することから変数間の関係性を確認することができる. どちらの手法も計算過程を参照できるという利点が挙げられ, これらを用いることで打刻データが有用かどうか検証する. 検証した結果, 成績予測精度が高くならない可能性がある. これは単純に用いるデータが予測するのに有効ではなかった可能性とデータ量が多すぎて不必要なデータが予測精度を下げている可能性がある. 不必要なデータがある場合はできるだけそれを除いて予測したいため, データ削減方法として本研究では情報利得を採用している. また特性をうまく抽出し, 予測するのに必要な抽出部分だけを使うことにより次元削減を行う主成分分析も採用した. 打刻データには打刻した場所と時間が含まれているが何の講義を受講していたかまでは打刻データから確認できないので, より詳細な成績データを使うことができない. よって成績データは各学生の GPA を用いることにする. 成績データには総合的な GPA だけでなく分野ごとの GPA も含まれている. 例えば, 名古屋工業大学では理系科目や専門科目, 選択科目などに分けられる. 本研究では総合的な GPA だけでなく各分野ごとの GPA も用いることにする. 成績予測について, 1 年終了時点では 1 年後期の成績あるいは 1 年通年の成績が予測できることを期待し, 1 年前期までのデータを利用する. 同様に 2 年終了時点では 2 年後期あるいは 1, 2 年通年の成績予測を期待し 2 年前期までのデータを利用する. また成績優秀者や成績不振者の行動パターンがどうなっているかを変数間の関係や特徴を比較的容易に見つけることができる主成分分析を用いて検証する. 予測精度を調べる際は leave one out 法を用いて評価し, 評価モデルの信頼度がある程度あるものとした.

本論文では, 第 2 章において本研究で用いる手法の理論を述べ, 第 3 章において打刻データにおける成績予測の有効性を示す. また第 4 章においてベイジアンネットワークと決定木による成績予測の結果を述べ, 第 5 章に成績による学生の打刻傾向についての分析した結果を述べる. 第 6 章にて本研究の結論と今後の課題を述べる. ちなみに, 本研究に用いる学習データには学生を特定できる情報は一切含まれておらず仮の番号により管理されているため, プライバシーが侵害されることはないことをここに付記する.

第2章　本研究で用いる手法の理論

本研究では成績予測の手法としてベイジアンネットワークと決定木を提案している. 本章ではその二つの手法の概要について記す.

2.1　決定木

決定木は主に分類や予測を行う際に広く用いられる手法である. 利点として, 多くの計算を必要とせずに分類できることや学習過程が知識として理解しやすく、どの項目が最も重要かを明確に示すことできるという点が挙げられる. 目的変数がカテゴリー型である分類木と, 数値型である回帰木に分けられるがデータマイニングで用いられる決定木は分類木を指すことが多い. 所与データから決定木を構築する際に, 根ルートから葉ルートまでのパス長をできるだけ最小にすることにより予測精度を高く保つことができるが, パス長が最小の決定木を作るという問題を解くのは NP 困難とされている. このように, サイズを最小とする決定木を構築することは困難な問題であるが, ある程度の実用的なサイズの決定木を構築することは以下の再帰的アルゴリズムにより構築できる.

1. 根ノードに置く属性を決定し, その属性値に応じて分岐を作成
2. その分岐にそって子ノードを作成し, 与えられたデータ集合を部分集合へと分割
3. 1, 2 のプロセスを再帰的に繰り返し, 決定木を成長させる
4. 子ノードのすべての事例が同一クラスに属していれば決定木の成長を止める

ただし, 決定木が必要以上に大きくなると過学習の問題が発生することがある. ある程度の運用可能な分類制度が期待できるならば不必要な部分木は切り落とすほうが視覚的にもわかりやすくなる. この切り落とす行為を枝刈りと呼び, これによりある程度実用的なサイズの決定木のなかでも最良のものを作成することができる.

決定木の主な生成アルゴリズムについて説明する.[7]

2.1.1　ID3

ID3 はエントロピー (情報量) を用いて木を分岐させる条件を決定する. あるデータ集合の事象 X における「あいまいさ」をエントロピー $Info$ で表すと

$$Info(X) = -\sum_{J=1}^{k} p(j|t) log_k p(j|t) \tag{2.1}$$

となる.$p(j|t)$ はノード t 内の k 種類あるクラス $a_1, \cdots, a_k$ のうちクラス $a_j (1 \leq j \leq k)$ の出現確率を表している. ただし $a_i \cap a_j = 0 (i \neq j)$.

分岐を決定する際, 式 (2.1) より分岐前のエントロピー $Info_p$ を求め, 次にある属性がある値であるかどうかで分岐させたと仮定したときのエントロピー $Info_l$ を求める. そして $Info_p - Info_l$ が最大となるような条件のものを選択する. このときの $Info_p - Info_l$ は Gain(相互情報量) と呼ばれ, 親ノードのエントロピーと各子ノードの合計したエントロピーとの差を表している.

2.1.2　C4.5

C4.5 は ID3 を元に改良されたプログラムであり式 (2.1) の log の底を 2 で計算し、$SplitInfo(X)$ でその Gain(X) を割り規格化を行う. また規格化したものを Gain に対して Gain 比と呼ぶ.

$$SplitInfo(X) = \sum_{j}^{C} \frac{N(t_j)}{N(t)} \times log_2 \frac{1}{\frac{N(t_i)}{N(t)}} \tag{2.2}$$

$$Gainratio(X) = \frac{Gain(X)}{SplitInfo(X)} \tag{2.3}$$

なお $N(t)$ はノード t 内のデータ数、$N(t_j)$ はノード t 内のクラス j をもつデータ数、C はクラスのカテゴリ数を表しており, Gain 比が最大になるものを分岐の条件として選択する.

2.1.3　CART

CART は ID3 と同様に 2 分木を生成するアルゴリズムであり, GiniIndex という不純度を指す指標を用いて分岐を行う.

$$GiniIndex(t) = 1 - \sum_{J=1}^{k} p^2(j|t) \tag{2.4}$$

$p(j|t)$ はノード t 内のクラス j の割合を表している.CART は ID3 のように属性のとりうる値を 1 つずつ調べて行くだけでなく, とりうる値の組合せも考慮していく. 例えばある属性のとりうる値が A, B, C, D だとすると (A|B,C,D),(B|A,C,D),(C|A,B,D),(D|A,B,C),(A,B|C,D),(A,C|B,D),(A,D|B,C) の 7 通りの GiniIndex を求めることになる. 分岐の手順は ID3 とほぼ同様で GiniIndex を求めた後, 親ノードと子ノードの GiniIndex の差をが大きいものを分岐の条件として選択する.

今回データマイニングの手法として決定木を利用するためにフリーソフトの Weka[8] を用いている.Weka で決定木を行う際 J4.8 というアルゴリズムを利用する. これは C4.5 という決定木, データ更新, 枝刈りをまとめて実行するアルゴリズムである.C4.5 は各ノードで枝分かれの数が異なる木を作る. デフォルトでは変数がとりうる各値に対して, その数だけ枝分かれを行うため, 例えば色の変数が赤, オレンジ, 黄, 緑, 青, 紫, 白の値を含んでいると次のレベルでは 7 つのノードができている. Weka では出力された木構造をビジュアル化できるため本研究ではこれを利用し木構造に打刻データが使われているかを確認する.

2.2　ベイジアンネットワーク

ベイジアンネットワークは事象同士の依存関係を矢印に見立てた有向グラフで表現し, 事象の起こりやすさを条件付き確率によって確率的に表現する確率モデルである. そのため目的変数と説明変数を区別しない手法として注目されている. 確率モデルの一種として不確実性を含む事象の予測や合理的な意思決定, 観測結果から原因を探る障害診断などに利用されている.[9]

ベイジアンネットワークは確率変数, その間の依存関係を表すグラフ構造, 条件付き確率で定義される.

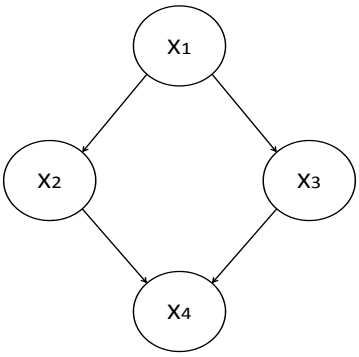

図 2.1: ベイジアンネットワークの例

図 2.1 はベイジアンネットワークの例である. 確率変数 X_1 と X_2 の関係に着目すると, 条件付き依存性を $X_1 \rightarrow X_2$ と表しており, リンク先であるノード (X_2) は子ノード、リンク元であるノード (X_1) は親ノードとして扱われる. 親ノードが複数ある場合, 子ノード X_2 の親ノードの集合を $P_a(X_2)$ と書くと、X_2 と $P_a(X_2)$ の間の依存関係は $P(X_2|P_a(X_2))$ という条件付き確率により定量的に表せる. 図 2.1 全体より 4 個の確率変数 $X_1, \cdots, X_4$ のそれぞれを子ノードとして同様に考えた場合, すべての確率変数の同時分布 $P(X_1, \cdots, X_4)$ は

$$\begin{aligned} P(X_1, \cdots, X_4) &= P(X_1|P_a(X_1))P(X_2|P_a(X_2))\cdots P(X_4|P_a(X_4)) \\ &= P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2, X_3) \end{aligned} \tag{2.5}$$

と表せる. すべての変数の確率分布は, 同時分布を計算することによって得られるのでベイジアンネットワークはこれを用いることで求めることができる. ベイジアンネットワークによる確率推論は以下の手順で行われる.

1. 親ノードも観測値も持たないノードに事前確率分布を与える
2. 観測された変数の値（エビデンス）e をノードにセットする
3. 知りたい対象の変数 X の事後確率 $P(X|e)$ を得る

この計算により観測情報 $e = X_2 = 1, X_3 = 1$ から事後確率 $P(X_4|e)$ を得ることができる.

事後確率を求めるために, 変数間の局所計算を繰り返しながら確率をネットワーク中に伝搬することにより各変数の確率分布を更新していく計算法を確率伝搬法という. 確率伝搬法の説明のために今回はある複雑な構造を持つモデルの一部分として次の単純なモデルを取り出しその部分における確率計算を考える.

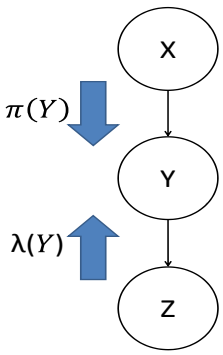

図 2.2: 単純な構造における確率伝搬

図 2.2 のように $X \to Y \to Z$ の間に依存関係があり条件付き確率が与えられているものとする. 計算しようとしているノードを Y とし, 親ノードに与えられる観測情報を, 子ノードに与えられる観測情報をとする. また e^+ と e^- は Y を固定したときに条件付き独立になるため、$\alpha = \frac{1}{P(e^+|e^-)}$ を Y の値によらない正規化された値とすれば求めたい事後確率 $P(Y|e)$ はベイズの定理により、

$$P(Y|e) = \alpha P(e^-|Y)P(Y|e^+) \tag{2.6}$$

となる. ここで図 2.2 のように親ノードからの寄与確率値を $P(Y|e^+) = \pi(Y)$ とおくと

$$\pi(Y) = \sum_X P(Y|X)P(X|e^+) \tag{2.7}$$

のように変形できる. このときノード X に親ノードがいない場合は予め用意された事前確率を与え, 観測情報が与えられている場合, その値は決定できる. ノード X に入力がなく、かつノード X に親ノードが存在するとき式 (2.7) を再帰的に適用することによりその値を求めることができる. 同様にノード Z についても考える. 子ノードからの寄与確率値を $P(e^-|Y) = \lambda(Y)$ と置くと

$$\lambda\ (Y) = \sum_Z P(e^-|Y, Z)P(Z|Y) \tag{2.8}$$

となる. 観測情報 e^- は Y の値に関係なく独立であることから

$$\lambda\ (Y) = \sum_Z P(e^-|Z)P(Z|Y) \tag{2.9}$$

ここで $P(Z|Y)$ は条件付き確率表として与えられいることから $P(e^-|Z) = \lambda\ (Z)$ は観測情報が与えられているとき値が決定できる. また, 観測情報がなくそのノードが子ノードを持たない下端のノードの場合は、無情報であることから一様分布確率として Z のいかなる状態について等しい値とする. また, ノードが子ノードを持つ場合, 式 (2.9) を再帰的に適用することで最終的に下端のノードの値を求めることができる. これらを利用することでノード Y の事後確率を求めることができる. よってグラフ構造内のすべてのパスがループを持たないとき任意のノードの事後確率を局所的に求めることができる.

ベイジアンネットワークのモデルは有向グラフで表されている. よってベイジアンネットワークで予測を行った場合, 有向グラフが予測結果に大きく関係してくる. 一般的に, 有向グラフでは専門的な知識や経験則によりモデルを構築して行くがそういった特別な知識を知らなくてもグラフ構造を決定することができる手法がいくつかある. その代表的な構造について簡単な説明を述べる.

2.2.1　Naive Bayes

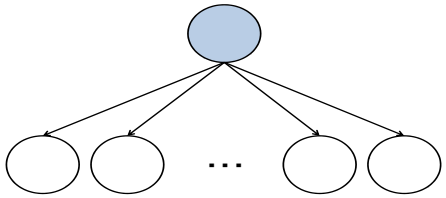

図 2.3: Naive Bayes の例

NaiveBayes はあるノードの親ノードが 1 つしかないベイジアンネットワークのことである. そのため目的変数が木構造における根の部分となり, 他の変数は根ノードの葉の部分となる.NaiveBayes は最もシンプルのグラフ構造をしているため条件付き確率を繰り返し適用することで求めることができる. ただし, 葉となるノードを無闇に増やしても結果が良くなるとは限らず, 悪くなる可能性もあるので適切な変数を選ばなければならない.

2.2.2　TAN(Tree Augmented Network)

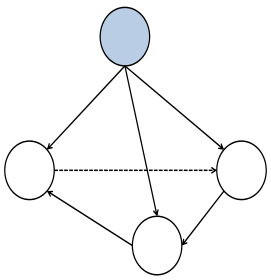

図 2.4: TAN の例

TAN は Naive Bayes と似た基本的構造を有している.Naivebayes と違う点は目的変数以外の変数がもう一つの変数を親ノードとして持っている点である. 条件付き相互情報量により親となるノードが決められる. 目的変数を C, その他の 2 つの変数を X,Y としたとき目的変数 C が与えられたという条件付き相互情報量は

$$I(X;Y|C) = \sum_{x,y,c} P(x,y,c) log \frac{P(x,y|c)}{P(x|c)P(y|c)} \tag{2.10}$$

で求めることができる. このときの $I(X;Y|C)$ が最大となる X,Y をもとめることにより TAN のグラフを決定することができる.

2.3　データマイニング

データマイニングとは, 大量のデータから有用な知識や情報を見つけることである.[10] データマイニングは主に以下の過程により実行される.

1. データの収集
2. データの選択
3. パターン発見
4. 見つけ出したパターンの解釈

本研究では, 打刻データ及び成績データを収集しベイジアンネットワークと決定木によりモデルを構築, そしてそのモデルについて評価を行っている. また最適なモデルを構築するためにデータの選択について, ある手法を採用している. 以下より, 採用した手法についての概要を述べる.

2.3.1　情報利得

情報利得とは「ある素性が出現したか否か」という情報がクラスに関するあいまいさをどれくらい減少させるかを表したものであり, 確率変数を C とするとエントロピー $H(C)$ は

$$H(C) = -\sum P(c)logP(c) \tag{2.11}$$

で表される.「ある素性が出現した」ことが分かっている場合の条件付きエントロピーは

$$H(C|X_w = 1) = -\sum_c P(c|X_w = 1)logP(c|X_w = 1) \tag{2.12}$$

で表される. また「ある素性が出現しなかった」ことが分かっている場合の条件付き確率は

$$H(C|X_w = 0) = -\sum_c P(c|X_w = 0)logP(c|X_w = 0) \tag{2.13}$$

で表される. 以上より素性 w の情報利得 $IG(w)$ は次のように定義される.

$$IG(w) = H(C) - (P(X_w = 1)H(C|X_w = 0) + P(X_w = 0)H(C|X_w = 0)) \tag{2.14}$$

2.3.2　主成分分析

主成分分析は多変量解析の 1 つであり, データの傾向を調べたり分類するのに用いられる. 主成分分析は主に以下の過程により導出される.

1. 平均値, 分散, 共分散を求めて相関行列を求める
2. 固有値, 寄与率, 累積寄与率の算出

各説明変数 $x_1, x_2, \cdots, x_n$ の単位が揃っている場合において新たな変数 z_1 を導入し, 不偏分散が最大になるように係数 $a_1, a_2, \cdots, a_n$ を定めると

$$z_1 = \sum_{i=1}^{n} a_i x_i \tag{2.15}$$

となる. ただしベクトル a の大きさが 1 になるようにする.

$$\sum_{i=1}^{n} a_i^2 = 1 \tag{2.16}$$

このように定めた z_1 を第 1 主成分と言う. 次に第 1 主成分の時と同様に $z_1 = \sum_{i=1}^{n} b_i x_i$ で定義される新たな変数 z_2 を導入する. このときベクトル b の大きさが 1 かつベクトル b とベクトル a が垂直になるように制限を付ける.

$$\sum_{i=1}^{n} b_i^2 = 1 \qquad\qquad \sum_{i=1}^{n} a_i b_i = 0 \tag{2.17}$$

このように主成分を 1 つずつ増やして行くにはベクトルの大きさが 1 であることとそれまでに定義されたベクトルと垂直となるよう制限することで定義することができる. 元の変数 $x_1, x_2, \cdots, x_n$ の不偏分散の和を

$$S = \sum_{i=1}^{n} S(x_i) \tag{2.18}$$

とおくとき, 寄与率を C とおくと

$$C = \frac{S(Z)}{S} \tag{2.19}$$

また累積寄与率 P は

$$P = \sum_{i=1}^{n} C_i \tag{2.20}$$

で表される. 一般的に累積寄与率は 70 %〜80 %あたりになるまで主成分を求める. 本研究では 80 %以上になった時点までの主成分を用いて検証する.

　各主成分における説明変数の係数の大きさを見て, どの変数を重視しているかによりそのときの主成分が何を表しているかを定義する.

第3章　成績レベル予測に用いるデータについて

本研究では成績予測の手法としてベイジアンネットワークと決定木を採用している. また予測に用いるデータについては打刻データに着目している. ベイジアンネットワーク及び決定木による予測を行う際, 打刻データの質によっては最適なグラフを構築することが難しくなる. そこで, 本章では打刻データについての概要及び予測に用いる成績データについて解説していく.

3.1　打刻データ及び成績データの概要

打刻データは学生番号, 教室, 打刻日, 打刻時間が 1 レコードとなっており大学 1 年次は約 11 万レコード , 大学 2 年次は約 10 万レコードのデータが記録されている. また学生は 171 人分の打刻情報が記録されている. ただし , 学生番号は個人が特定されないようあらかじめ暗号化されている. また学生 171 人は学科は同じだがクラスが特定されていないため授業構成により打刻時間帯もさまざまであることを言及しておく. 教室についてはすべての打刻に対応しているわけではなく , 記載されていない場合もある.

授業に関係の無い打刻 (休日の打刻など) を使用しないためにまず休日に打刻しているデータは無視する. また前期は 4 月〜7 月, 後期は 10 月〜1 月のデータ部分を用いる. 記録された打刻データには間違えて打刻した場合や打刻したか不安になり安全のために打刻し直す場合,IC カードリーダの反応が悪く何回も打刻をやり直す状況が考えられる. 全部を特定することは不可能なので本研究では打刻してから 1 分以内に再び打刻されたデータについては無視することにする. これらにより大学 1 年次は約 2 万レコード , 大学 2 年次は約 5 千レコード削減された.

成績データは打刻データに記録されている学生 171 人分の成績が記録されており学生番号, 講義名,GP 数値, 開講学期が 1 レコードとなっている. しかし打刻データからはどの講義の打刻であるかが分からないため各分野の講義成績データを予測することは難しい. そのため予測する成績は各学期または学年終了時の GPA とする. ここで前期分については打刻データで成績を予測するための補助するデータとして用いるため, 後期分および学年終了時の GPA が予測の対象となる.

決定木や特にベイジアンネットワークからモデルを構築するにはある程度の量を有するデータから学習しなければならない. 打刻データについてそのまま用いると情報量に乏しいため満足できるモデルを作成することができない. そこでデータの拡張を行うことで情報量を増やすことにする.

3.2　データの拡張

打刻データについて今のままでは学生番号, 教室, 打刻日, 打刻時間の 4 変数しかないのでこのままで用いても最適なモデルを作成することは難しい. そこで打刻データの 4 変数から新たに変数を作成する.

まず打刻日から前期 4 月〜7 月, 後期 10 月〜1 月の打刻を特定することができる. 打刻時間が記録されている数より打刻回数を求めることができるので, これより前期後期分の月毎打刻回数を変数として用いる. またある日に打刻時間が記録されていない場合, その日は打刻するのを忘れているもしくは休んでいると考えることができる. これより厳密ではないが欠席回数を求めることができるため打刻回数のときと同様に月毎の欠席回数を変数として用いる. また学生の性格面を踏まえ考えてみると, 祝日前後や長期休暇直後は欠席しがちということが期待できる. よって祝日前後欠席回数を変数として用いる.

打刻日により曜日を特定することができる. 学生の打刻傾向が成績に影響するかどうか調べるため打刻回数から各曜日の平均打刻回数と分散を調べる. これにより例えば, 分散が高ければ安定した打刻がされていないことを示すことができる. また分散が低く安定した打刻がされているとしても打刻していない (打刻回数 0) の可能性があるので, 平均より打刻をしているかどうか確認することができる. 以下の表 3.1 にここで述べた変数を列挙した. これらはすべて説明変数として用いる.

表 3.1: 説明変数として用いる打刻データ

番号	変数名	意味
1	1 年 4 月打刻回数	1 年 4 月に打刻した回数
2	1 年 5 月打刻回数	1 年 5 月に打刻した回数
3	1 年 6 月打刻回数	1 年 6 月に打刻した回数
4	1 年 7 月打刻回数	1 年 7 月に打刻した回数
5	1 年 4 月欠席回数	1 年 4 月に欠席または打刻していない日の回数
6	1 年 5 月欠席回数	1 年 5 月に欠席または打刻していない日の回数
7	1 年 6 月欠席回数	1 年 6 月に欠席または打刻していない日の回数
8	1 年 7 月欠席回数	1 年 7 月に欠席または打刻していない日の回数
9	1 年前期祝日前後欠席回数	1 年前期における祝日前後あるいは長期休暇直後に欠席または打刻していない日の回数
10	1 年前期月曜打刻平均	1 年前期月曜における打刻回数の平均
11	1 年前期火曜打刻平均	1 年前期火曜における打刻回数の平均
12	1 年前期水曜打刻平均	1 年前期水曜における打刻回数の平均
13	1 年前期木曜打刻平均	1 年前期木曜における打刻回数の平均
14	1 年前期金曜打刻平均	1 年前期金曜における打刻回数の平均
15	1 年前期月曜打刻分散	1 年前期月曜における打刻回数の分散
16	1 年前期火曜打刻分散	1 年前期火曜における打刻回数の分散
17	1 年前期水曜打刻分散	1 年前期水曜における打刻回数の分散
18	1 年前期木曜打刻分散	1 年前期木曜における打刻回数の分散
19	1 年前期金曜打刻分散	1 年前期金曜における打刻回数の分散
20	1 年 10 月打刻回数	1 年 10 月に打刻した回数
21	1 年 11 月打刻回数	1 年 11 月に打刻した回数

22	1 年 12 月打刻回数	1 年 12 月に打刻した回数
23	1 年 1 月打刻回数	1 年 1 月に打刻した回数
24	1 年 10 月欠席回数	1 年 10 月に欠席または打刻していない日の回数
25	1 年 11 月欠席回数	1 年 11 月に欠席または打刻していない日の回数
26	1 年 12 月欠席回数	1 年 12 月に欠席または打刻していない日の回数
27	1 年 1 月欠席回数	1 年 1 月に欠席または打刻していない日の回数
28	1 年祝日前後欠席回数	1 年次における祝日前後あるいは長期休暇直後に欠席または打刻していない日の回数
29	1 年後期月曜打刻平均	1 年後期月曜における打刻回数の平均
30	1 年後期火曜打刻平均	1 年後期火曜における打刻回数の平均
31	1 年後期水曜打刻平均	1 年後期水曜における打刻回数の平均
32	1 年後期木曜打刻平均	1 年後期木曜における打刻回数の平均
33	1 年後期金曜打刻平均	1 年後期金曜における打刻回数の平均
34	1 年後期月曜打刻分散	1 年後期月曜における打刻回数の分散
35	1 年後期火曜打刻分散	1 年後期火曜における打刻回数の分散
36	1 年後期水曜打刻分散	1 年後期水曜における打刻回数の分散
37	1 年後期木曜打刻分散	1 年後期木曜における打刻回数の分散
38	1 年後期金曜打刻分散	1 年後期金曜における打刻回数の分散
39	2 年 4 月打刻回数	2 年 4 月に打刻した回数
40	2 年 5 月打刻回数	2 年 5 月に打刻した回数
41	2 年 6 月打刻回数	2 年 6 月に打刻した回数
42	2 年 7 月打刻回数	2 年 7 月に打刻した回数
43	2 年 4 月欠席回数	2 年 4 月に欠席または打刻していない日の回数
44	2 年 5 月欠席回数	2 年 5 月に欠席または打刻していない日の回数
45	2 年 6 月欠席回数	2 年 6 月に欠席または打刻していない日の回数
46	2 年 7 月欠席回数	2 年 7 月に欠席または打刻していない日の回数
47	2 年前期祝日前後欠席回数	2 年前期における祝日前後あるいは長期休暇直後に欠席または打刻していない日の回数
48	2 年前期月曜打刻平均	2 年前期月曜における打刻回数の平均
49	2 年前期火曜打刻平均	2 年前期火曜における打刻回数の平均
50	2 年前期水曜打刻平均	2 年前期水曜における打刻回数の平均
51	2 年前期木曜打刻平均	2 年前期木曜における打刻回数の平均
52	2 年前期金曜打刻平均	2 年前期金曜における打刻回数の平均
53	2 年前期月曜打刻分散	2 年前期月曜における打刻回数の分散
54	2 年前期火曜打刻分散	2 年前期火曜における打刻回数の分散
55	2 年前期水曜打刻分散	2 年前期水曜における打刻回数の分散
56	2 年前期木曜打刻分散	2 年前期木曜における打刻回数の分散
57	2 年前期金曜打刻分散	2 年前期金曜における打刻回数の分散

また成績データの前期分は打刻データによる予測の補助として用いる. よって表 3.2 に成績データより用いた説明変数を列挙する.

表 3.2: 説明変数として用いる成績データ

番号	変数名	意味
58	1 年前期	1 年次の前期に受講した講義の GPA 値
59	1 年前期英語	1 年次の前期に受講した英語教科の GPA 値
60	1 年前期人文	1 年次の前期に受講した「人間文化」に分類される講義の GPA 値
61	1 年前期体育	1 年次の前期に受講した体育教科の GPA 値
62	1 年前期理系	1 年次の前期に受講した理系基礎 (数学, 理科系) の講義の GPA 値
63	1 年前期専門	1 年次の前期に受講した専門科目の講義の GPA 値
64	1 年前期理科	1 年次の前期に受講した理科系の講義の GPA 値
65	1 年前期数学	1 年次の前期に受講した数学系の講義の GPA 値
66	1 年後期	1 年次の後期に受講した講義の GPA 値
67	1 年後期英語	1 年次の後期に受講した英語教科の GPA 値
68	1 年後期人文	1 年次の後期に受講した「人間文化」に分類される講義の GPA 値
69	1 年後期体育	1 年次の後期に受講した体育教科の GPA 値
70	1 年後期理系	1 年次の後期に受講した理系基礎 (数学, 理科系) の講義の GPA 値
71	1 年後期専門	1 年次の後期に受講した専門科目の講義の GPA 値
72	1 年後期理科	1 年次の後期に受講した理科系の講義の GPA 値
73	1 年後期数学	1 年次の後期に受講した数学系の講義の GPA 値
74	2 年前期	2 年次の前期に受講した講義の GPA 値
75	2 年前期英語	2 年次の前期に受講した英語教科の GPA 値
76	2 年前期人文	2 年次の前期に受講した「人間文化」に分類される講義の GPA 値
77	2 年前期体育	2 年次の前期に受講した体育教科の GPA 値
78	2 年前期理系	2 年次の前期に受講した理系基礎 (数学, 理科系) の講義の GPA 値
79	2 年前期専門	2 年次の前期に受講した専門科目の講義の GPA 値
80	2 年前期理科	2 年次の前期に受講した理科系の講義の GPA 値
81	2 年前期数学	2 年次の前期に受講した数学系の講義の GPA 値

目的変数には 1 年後期の GPA, 1 年通年の GPA, 2 年後期の GPA, 1,2 年複合した GPA を用いた.1 年後期および 1 年通年の GPA は 1 年前期のデータ, つまり表 3.1 の番号 1〜19, 表 3.2 の番号 58〜65 をも用いて予測している. また 2 年後期および 1,2 年複合した GPA は 2 年前期までのデータ, つまり表 3.1, 表 3.2 で書かれている変数全てを用い予測を行っている.

ベイジアンネットワークや決定木はモデル構築の際, 数値型である目的変数を名詞型に書き換えなければならない. そのため目的変数を離散化することにより書き換える. 簡易な離散化の方法として

1. 等頻度による離散化
2. 等間隔による離散化

がある. 名古屋工業大学では成績は「秀, 優, 良, 可, 不可」の 5 段階で表されている. よって本研究では目的変数に用いた各 GPA を等間隔による離散化を行い 5 段階 (S, A, B, C, D) に分けた.S は成績が優秀であることを表しており A, B と進むにつれて成績が悪くなっていることを表している. 表 3.3 に目的変数を 5 段階に分けた詳細について列挙する.

表 3.3: 目的変数として用いる成績データ

番号	予測する学期	離散化	GPA 数値範囲
82	1 年後期	S	2.914286 以上
		A	2.185714〜2.914286
		B	1.457143〜2.185714
		C	0.728571〜1.457143
		D	0.728571 以下
83	1 年通年	S	2.918519 以上
		A	2.225926〜2.918519
		B	1.533333〜2.225926
		C	0.840741〜1.533333
		D	0.840741 以下
84	2 年後期	S	3.066667 以上
		A	2.3〜3.066667
		B	1.533333〜2.3
		C	0.766667〜1.533333
		D	0.766667 以下
85	1,2 年複合	S	2.915451 以上
		A	2.213255〜2.91545
		B	1.511059〜2.213255
		C	0.808863〜1.511059
		D	0.808863 以下

第4章　ICカード打刻データと修学データを用いた学生の将来の成績レベル予測

打刻データより特徴を抽出するためにベイジアンネットワークおよび決定木を成績予測手法として用いる. それによりモデルを構築しその構築過程において打刻データが確認できれば, 成績に係わる特徴として言うことができる. しかし, 予測が正しくなければモデルから打刻データの特徴を見付けることができたとしてもモデル自体が正しくないために見付けた特徴が有用であるとは言い難い. そこでまず成績を予測するにあたり予測の精度がどの程度のものか検証した.

4.1　前期までのデータによる予測の検証

本研究で予測する成績と予測するにあたり用いるデータの組合せは

- 1 年前期までのデータから 1 年後期の成績を予測
- 1 年前期までのデータから 1 年通年の成績を予測
- 2 年前期までのデータから 2 年後期の成績を予測
- 2 年前期までのデータから 1,2 年複合した成績の予測

となっている. それぞれの予測対象を第 3 章で述べたデータ部分を用いて予測した結果を示す. 評価モデルについては leave one out 法により信頼度がある程度あるものとし, ベイジアンネットワークの構築モデルについては予測精度よりモデル精度を重視し TAN を用いて予測を行っている. また予測結果を示した表 4.1〜表 4.8 は行が離散化で分けられた成績を持つ学生数, 列が予測された成績を持つ学生数を表している.

1 年前期までのデータから 1 年後期の成績を予測

表 4.1: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	0	22	2	0	0
A	4	31	21	1	0
B	0	20	25	11	1
C	0	0	18	5	2
D	0	0	5	0	3

表 4.2: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	7	11	6	0	0
A	10	20	24	2	1
B	3	24	22	6	2
C	0	5	10	7	3
D	0	0	3	5	0

表 4.1 よりベイジアンネットワークによる予測的中精度は 37.4267 %, 表 4.2 より決定木による予測的中精度は 32.7485 %になった. 1 ランク違いを含めた的中の割合に着目するとベイジアンネットワークによる予測的中率は 94.7368 %, 決定木による予測的中率は 87.1345 %となり, これにより成績を完全に的中させることは難しいが, 大体の成績を的中させることは可能ということが言える.

1 年前期までのデータから 1 年通年の成績を予測

表 4.3: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	15	9	0	0	0
A	1	47	19	1	0
B	0	6	48	3	0
C	0	0	5	13	0
D	0	0	0	4	0

表 4.4: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	18	6	0	0	0
A	8	49	10	1	0
B	1	8	44	3	1
C	0	0	9	8	1
D	0	0	0	4	0

表 4.3 よりベイジアンネットワークによる予測的中精度は 71.9298 %, 表 4.4 より決定木による予測的中精度は 69.5906 %となり, 成績を完全に的中することはある程度可能であると言える. また 1 ランク違いを含めた的中の割合に着目するとベイジアンネットワークによる予測的中率は 99.4152 %, 決定木による予測的中率は 98.2456 %となり, ほぼ完全に的中することができる.

2 年前期までのデータから 2 年後期の成績を予測

表 4.5: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	0	15	0	0	0
A	1	45	6	4	0
B	0	14	39	6	3
C	0	0	6	8	7
D	0	0	0	7	10

表 4.6: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	0	14	1	0	0
A	0	45	9	1	1
B	0	16	38	2	6
C	0	2	7	5	7
D	0	0	7	3	7

表 4.5 よりベイジアンネットワークによる予測的中精度は 59.6491 %, 表 4.6 より決定木による予測的中精度は 55.5556 %となり的中率はあまり高くない. 1 ランク違いを含めた的中の割合に着目するとベイジアンネットワークによる予測的中率は 95.9064 %, 決定木による予測的中率は 90.0584 %となり, 大体の成績は予測可能であると言える.

2 年前期までのデータから 1,2 年複合した成績の予測

表 4.7: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	17	8	0	0	0
A	5	54	2	0	0
B	0	5	46	6	2
C	0	0	7	11	3
D	0	0	0	4	1

表 4.8: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	15	10	0	0	0
A	0	55	6	0	0
B	0	5	42	12	0
C	0	0	13	7	1
D	0	0	1	3	1

表 4.7 よりベイジアンネットワークによる予測的中精度は 75.4386 %, 表 4.8 より決定木による予測的中精度は 70.1754 %となり, 予測的中率自体は高い値では無いがそこそこの的中精度は保っている. 1 ランク違いを含めた的中の割合に着目するとベイジアンネットワークによる予測的中率は 98.8304 %, 決定木による予測的中率は 99.4152 %となっており, ほぼ的中することができる.

4.2　後期半学期までのデータによる予測の検証

次に後期の打刻のデータが半学期分わかっている, つまり後期 10 月〜1 月の内 10 月と 11 月が終った時点で予測をする場合予測的中率がどうなるか検証する. この場合における予測する成績と予測するにあたり用いるデータの組合せは

- 1 年後期半学期までのデータから 1 年後期の成績を予測
- 1 年後期半学期までのデータから 1 年通年の成績を予測
- 2 年後期半学期までのデータから 2 年後期の成績を予測
- 2 年後期半学期までのデータから 1,2 年複合した成績の予測

となる.4.1 で述べた内容と同様,leave one out 法を用い, ベイジアンネットワークの構築モデルについては TAN を用いて予測を行っている. 予測結果を表 4.9〜表 4.16 に示す.

1 年後期半学期までのデータから 1 年後期の成績を予測

表 4.9: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	0	22	2	0	0
A	5	29	19	4	0
B	0	20	24	12	1
C	0	0	15	7	3
D	0	0	2	2	4

表 4.10: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	9	10	4	1	0
A	11	23	20	2	1
B	1	31	17	7	1
C	0	1	11	11	2
D	0	1	0	3	4

表 4.9 よりベイジアンネットワークによる予測的中精度は 37.4269 %, 表 4.10 より決定木による予測的中精度は 32.4269 %になった.

1 年後期半学期までのデータから 1 年通年の成績を予測

表 4.11: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	15	9	0	0	0
A	2	58	8	0	0
B	0	10	45	2	0
C	0	0	4	12	2
D	0	0	0	4	0

表 4.12: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	17	7	0	0	0
A	9	48	11	0	0
B	0	17	37	3	0
C	0	0	9	9	0
D	0	0	0	2	2

表 4.11 よりベイジアンネットワークによる予測的中精度は 76.0234 %, 表 4.12 より決定木による予測的中精度は 66.0819 %になった.

2 年後期半学期までのデータから 2 年後期の成績を予測

表 4.13: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	0	15	0	0	0
A	1	45	6	4	0
B	0	14	39	8	1
C	0	0	6	10	5
D	0	0	0	8	9

表 4.14: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	0	14	1	0	0
A	0	48	7	0	1
B	1	17	38	5	1
C	0	2	7	9	3
D	0	0	2	2	13

表 4.13 よりベイジアンネットワークによる予測的中精度は 60.2339 %, 表 4.14 より決定木による予測的中精度は 63.1579 %になった.

2 年後期半学期までのデータから 1,2 年複合した成績の予測

表 4.15: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	17	8	0	0	0
A	5	54	2	0	0
B	0	5	47	5	2
C	0	0	8	11	2
D	0	0	0	4	1

表 4.16: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	15	10	0	0	0
A	0	55	6	0	0
B	0	5	42	12	0
C	0	0	11	8	2
D	0	0	2	2	1

表 4.15 よりベイジアンネットワークによる予測的中精度は 76.0234 %, 表 4.16 より決定木による予測的中精度は 70.7602 %になった.

前期までのデータから成績予測する場合と後期半学期分までのデータから成績予測する場合の結果を比較すると表 4.17 になった.

表 4.17: 前期までのデータと後期半学期までのデータによる成績予測結果の比較

目的変数	予測手法	1 年前期までのデータによる予測的中率	1 年後期半学期までのデータによる予測的中率
1 年後期	ベイジアンネットワーク	37.4267 %	37.4269 %
	決定木	32.7485 %	37.4269 %
1 年通年	ベイジアンネットワーク	71.9298 %	76.0234 %
	決定木	69.5906 %	66.0819 %

目的変数	予測手法	2 年前期までのデータによる予測的中率	2 年後期半学期までのデータによる予測的中率
2 年後期	ベイジアンネットワーク	59.6491 %	60.2339 %
	決定木	55.5556 %	63.1579 %
1,2 年複合	ベイジアンネットワーク	75.4386 %	76.0234 %
	決定木	70.1754 %	70.7602 %

表 4.17 の比較結果から後期半学期までのデータにより成績予測したほうが予測的中率が高くなる場合が多いが, 前期までのデータにより成績予測しても予測的中率に大幅な違いは出ていない. できるだけ早い段階で予測できることが理想なので本研究では表 4.17 より前期までのデータによる成績予測でも有用であると判断し後期半学期分のデータは用いないことにした. よって前期の結果, 表 4.1〜4.8 に着目すると 1 年後期など予測的中率の高くない部分も存在するが 1 ランク違いを含めると大体の成績を予測することは可能であることが検証された.

4.3　データの削減による予測的中率の向上

1 ランク違いを含めると大体の成績を予測することは可能だが, できるだけ 1 ランク違いを含めなくても予測的中率が高くなることが望ましい. 前期までのデータの中には予測的中率を下げる要因となっている変数が含まれている可能性がある. また構築されたモデルが複雑になり特徴が掴みにくくなる可能性もある. そのような事態を防ぐためデータの削減を行う. 削減方法は

- 情報利得
- 主成分分析

の 2 種類を用いて行う. 情報利得は変数が多すぎて予測に不必要な変数がノイズとなるのを避けるため, 予測に必要な変数を取捨選択することによりデータ量を削減する方法である. 主成分分析は多次元のデータを次元削減することにより必要になってくるデータを可視化する方法である. この 2 通りの手法を用いてそれぞれベイジアンネットワーク, 決定木により予測した結果がそのまま予測する場合と比べて予測的中率がどうなるかを検証した.

情報利得による削減

1 年後期,1 年通年を予測するために用いる 1 年前期までのデータについて情報利得をもとに削減した. なお取捨選択は有用性が高いものから変数 5 個区切りを基本にして行っている. 情報利得で取捨選択したデータについてベイジアンネットワークと決定木を用いて予測した結果を表 4.18, 表 4.19 に示す.

表 4.18: 情報利得による 1 年後期の成績予測的中率

取捨選択数	予測方法	1 年後期
5 個	ベイジアンネットワーク	37.4269 %
	決定木	44.4444 %
9 個	ベイジアンネットワーク	37.4269 %
	決定木	35.0877 %

表 4.19: 情報利得による 1 年通年の成績予測的中率

取捨選択数	予測方法	1 年通年
5 個	ベイジアンネットワーク	75.4386 %
	決定木	70.1754 %
10 個	ベイジアンネットワーク	74.8538 %
	決定木	67.8363 %
12 個	ベイジアンネットワーク	72.5146 %
	決定木	67.2515 %

情報利得をもとに削減した結果,1 年後期の予測に関しては 9 個の変数が有用であることが分かった. 表 4.18 より予測的中率はベイジアンネットワークに関しては取捨選択数 5 個のときと取捨選択数 9 個のとき, 共に 37.4269 %を示した. また決定木に関しては取捨選択数 5 個のときに最も高い予測的中率 44.4444 %を示した. 1 年通年の予測に関しては 12 個の変数が予測に有用であることが分かった. ベイジアンネットワークに関しては取捨選択数 5 個のときが最も高い予測的中率 75.4386 % を示した. また決定木も取捨選択数 5 個の時に最も高い予測的中率 70.1754 %を示した. 最も予測的中率が高いときの予測結果を表 4.20〜表 4.23 に示す.

1 年前期までのデータから 1 年後期の成績を予測

表 4.20: ベイジアンネットワークによる予測結果 (取捨選択数 5 個,9 個)

実際＼予測	S	A	B	C	D
S	0	23	1	0	0
A	4	31	21	1	0
B	0	20	25	11	1
C	0	0	18	5	2
D	0	0	5	0	3

表 4.21: 決定木による予測結果 (取捨選択数 5 個)

実際＼予測	S	A	B	C	D
S	7	12	5	0	0
A	12	22	18	4	1
B	4	15	30	7	1
C	0	4	6	14	1
D	0	0	1	4	3

1 年前期までのデータから 1 年通年の成績を予測

表 4.22: ベイジアンネットワークによる予測結果 (取捨選択数 5 個)

実際＼予測	S	A	B	C	D
S	15	9	0	0	0
A	1	48	19	0	0
B	0	1	56	0	0
C	0	0	8	10	0
D	0	0	0	4	0

表 4.23: 決定木による予測結果 (取捨選択数 5 個)

実際＼予測	S	A	B	C	D
S	15	9	0	0	0
A	6	47	15	0	0
B	0	3	49	5	0
C	0	0	10	7	1
D	0	0	0	2	2

次に 2 年後期,1,2 年複合を予測するために用いる 2 年前期までのデータについて情報利得をもとに削減した. そのときの結果を表 4.24, 表 4.25 に示す.

表 4.24: 情報利得による 2 年後期の成績予測的中率

取捨選択数	予測方法	2 年後期
5 個	ベイジアンネットワーク	57.3099 %
	決定木	49.7076 %
10 個	ベイジアンネットワーク	56.7251 %
	決定木	45.614 %
15 個	ベイジアンネットワーク	57.3099 %
	決定木	52.6316 %
20 個	ベイジアンネットワーク	56.7251 %
	決定木	50.8772 %
25 個	ベイジアンネットワーク	57.3099 %
	決定木	58.4795 %
30 個	ベイジアンネットワーク	60.2339 %
	決定木	59.0643 %
35 個	ベイジアンネットワーク	59.6491 %
	決定木	60.2339 %
40 個	ベイジアンネットワーク	59.6491 %
	決定木	57.8947 %

表 4.25: 情報利得による 1,2 年複合の成績予測的中率

取捨選択数	予測方法	1,2 年複合
5 個	ベイジアンネットワーク	69.5906 %
	決定木	69.5906 %
10 個	ベイジアンネットワーク	78.9474 %
	決定木	72.5146 %
15 個	ベイジアンネットワーク	78.3626 %
	決定木	74.269 %
20 個	ベイジアンネットワーク	79.5322 %
	決定木	68.4211 %
25 個	ベイジアンネットワーク	78.9474 %
	決定木	67.2515 %
30 個	ベイジアンネットワーク	78.3626 %
	決定木	67.8363 %
35 個	ベイジアンネットワーク	80.117 %
	決定木	71.345 %
40 個	ベイジアンネットワーク	77.7778 %
	決定木	70.7602 %

45 個	ベイジアンネットワーク	77.7778 %
	決定木	73.6842 %
50 個	ベイジアンネットワーク	78.3626 %
	決定木	70.7602 %
55 個	ベイジアンネットワーク	75.4386 %
	決定木	71.345 %

情報利得をもとに削減した結果,2 年後期の予測に関しては 40 個の変数が有用であることが分かった. 表 4.24 より予測的中率はベイジアンネットワークに関しては取捨選択数 30 個のとき, 60.2339 %を示した. また決定木に関しては取捨選択数 35 個のときに最も高い予測的中率 60.2339 %を示した. 1,2 年複合の予測に関しては 55 個の変数が予測に有用であることが分かった. ベイジアンネットワークに関しては取捨選択数 35 個のときが最も高い予測的中率 80.117 % を示した. また決定木に関しては取捨選択数 15 個の時に最も高い予測的中率 74.269 %を示した. 最も予測的中率が高いときの予測結果を表 4.26〜表 4.29 に示す.

2 年前期までのデータから 2 年後期の成績を予測

表 4.26: ベイジアンネットワークによる予測結果 (取捨選択数 30 個)

実際＼予測	S	A	B	C	D
S	0	15	0	0	0
A	0	45	10	0	1
B	0	16	42	2	2
C	0	1	9	9	2
D	0	0	9	1	7

表 4.27: 決定木による予測結果 (取捨選択数 35 個)

実際＼予測	S	A	B	C	D
S	0	15	0	0	0
A	1	46	5	4	0
B	0	14	39	6	3
C	0	0	6	8	7
D	0	0	0	7	10

2 年前期までのデータから 1.2 年複合の成績を予測

表 4.28: ベイジアンネットワークによる予測結果 (取捨選択数 35 個)

実際＼予測	S	A	B	C	D
S	15	10	0	0	0
A	1	55	5	0	0
B	0	5	49	4	1
C	0	0	4	17	0
D	0	0	0	4	1

表 4.29: 決定木による予測結果 (取捨選択数 15 個)

実際＼予測	S	A	B	C	D
S	19	6	0	0	0
A	6	52	3	0	0
B	0	6	44	9	0
C	0	0	10	11	0
D	0	0	0	4	1

主成分分析による削減

情報利得と同様に目的変数である 1 年後期, 1 年通年, 2 年後期, 1,2 年通年のデータを主成分分析を用いて次元削減を行った. 説明変数である 1 年前期までのデータに関しては第 5 主成分まで削減し,2 年前期までのデータに関しては第 11 主成分まで削減した. 削減したデータについてベイジアンネットワークと決定木を用いて予測した結果を表 4.30 に示す. またそのときの予測結果を表 4.31〜表 4.38 に示す.

表 4.30: 主成分分析による成績予測的中率

予測 GPA	ベイジアンネットワーク	決定木
1 年後期	36.8421 %	37,4269 %
1 年通年	73.6842 %	64.3275 %
2 年後期	49.7076 %	50.2924 %
1,2 年複合	76.6082 %	74.8538 %

1 年前期までのデータから 1 年後期の成績を予測

表 4.31: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	0	23	1	0	0
A	4	31	18	4	0
B	0	20	25	12	0
C	0	0	18	7	0
D	0	0	4	4	0

表 4.32: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	8	13	3	0	0
A	14	19	21	3	0
B	4	20	26	6	1
C	0	3	7	10	5
D	0	1	2	4	1

1 年前期までのデータから 1 年通年の成績を予測

表 4.33: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	14	10	0	0	0
A	1	59	7	1	0
B	0	10	44	3	0
C	1	0	8	9	0
D	0	0	0	4	0

表 4.34: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	16	8	0	0	0
A	8	51	9	0	0
B	0	14	36	6	1
C	0	0	10	7	1
D	0	0	1	3	0

2 年前期までのデータから 2 年後期の成績を予測

表 4.35: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	0	15	0	0	0
A	6	41	5	3	1
B	1	16	35	8	2
C	0	1	12	3	5
D	0	0	4	7	6

表 4.36: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	0	15	0	0	0
A	0	45	9	1	1
B	1	15	30	15	1
C	0	1	12	3	5
D	0	0	7	2	8

2 年前期までのデータから 1,2 年複合した成績の予測

表 4.37: ベイジアンネットワークによる予測結果

実際＼予測	S	A	B	C	D
S	17	8	0	0	0
A	5	51	5	0	0
B	0	3	48	5	3
C	0	0	8	12	1
D	0	0	0	2	3

表 4.38: 決定木による予測結果

実際＼予測	S	A	B	C	D
S	15	10	0	0	0
A	0	55	6	0	0
B	0	6	47	5	1
C	0	0	12	8	1
D	0	0	1	1	3

表 4.1〜表 4.8, 表 4.18, 表 4.19, 表 4.24, 表 4.25, 表 4.30 より · 削減方法を用いない · 情報利得を用いる · 主成分分析を用いる の 3 通りのうち予測的中率が最も高くなるときの削減方法についてまとめたものを表 4.39 に示す.

表 4.39: 最も高い予測的中率を示す削減方法

予測 GPA	削減方法	予測方法	予測的中率
1 年後期	情報利得 (5 個)	ベイジアンネットワーク	37.4269 %
	情報利得 (5 個)	決定木	44.4444 %
1 年通年	情報利得 (5 個)	ベイジアンネットワーク	75.4386 %
	情報利得 (5 個)	決定木	70.1754 %
2 年後期	情報利得 (30 個)	ベイジアンネットワーク	60.2339 %
	情報利得 (35 個)	決定木	60.2339 %
1,2 年複合	情報利得 (35 個)	ベイジアンネットワーク	80.117 %
	主成分分析	決定木	74.8538 %

表 4.39 により目的変数である 1 年後期の成績についてはベイジアンネットワークと決定木, 共に情報利得で有用性の高い変数から 5 個を取捨選択したデータが最も予測的中率が高くなった.1 年通年の成績に付いてもベイジアンネットワークと決定木, 共に情報利得により有用な変数から 5 個取捨選択したデータが最も高い予測的中率を示した. これらは説明変数として 1 年前期までのデータによって予測が行われているが 1 年後期, 1 年通年予測するにあたりデータ量が少ないため, 有用成の高い変数による予測に重みが置かれていると考えられる.2 年後期の予測に関してはベイジアンネットワークに関しては情報利得で有用性の高い変数から 30 個を取捨選択したデータが最も予測的中率が高くなり, 決定木に関しては情報利得で有用性の高い変数から 35 個を取捨選択したときが最も高い予測的中率を示した. これより有用性が低い変数も予測に必要であることが分かる.1,2 年複合の予測に関してベイジアンネットワークは情報利得により有用性の高い変数から 35 個を取捨選択したデータが最も予測的中率が高くなり, 決定木に関しては主成分分析により次元削減したデータが最も高い予測的中率を示した. 情報利得に関しては 5 個区切りで行っているため取捨選択数 30 個〜取捨選択数 35 個の間に予測に関わる変数があったと考えられる. 主成分分析に関してはモデル構造をみないと詳細はわからない.

第5章　成績レベル予測モデルからの特徴分析

表 4.39 は本研究で予測するにあたり一番高い予測的中率を示した表である. よって表 4.39 に書かれている削減方法を用いて成績を予測した結果, モデルがどう構築されたかを検証する.

5.1　1年後期における構築モデル

1 年後期は 1 年前期までのデータを用いて予測を行っている. 表 4.39 より削減方法としてはベイジアンネットワーク, 決定木共に情報利得で有効性の高い変数から 5 個取捨選択している. このとき選択された変数を表 5.1 に示す.

表 5.1: 情報利得により取捨選択された変数

番号	変数名
3	1年6月打刻回数
7	1年6月欠席回数
8	1年7月欠席回数
58	1年前期
62	1年前期理系

取捨選択された 5 つの変数と目的変数である 1 年後期を離散化したデータをベイジアンネットワークと決定木で予測し, それにより構築されたモデルを図 5.1, 図 5.2 に示す.

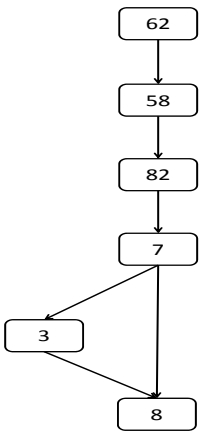

図 5.1: ベイジアンネットワークによって構築されたモデル

図 5.1 より 1 年 6 月欠席回数が 1 年後期と関係があることがわかる. 条件付き確率を見てみると表 5.2 になった.

表 5.2: 1 年 6 月欠席回数における条件付き確率

1 年後期	4.5 以下	4.5〜7	7〜12.5	12.5 以上
S	94.2 %	1.9 %	1.9 %	1.9 %
A	95.8 %	2.5 %	0.8 %	0.8 %
B	95.8 %	0.8 %	0.8 %	2.5 %
C	64.8 %	31.5 %	1.9 %	1.9 %
D	55 %	5 %	35 %	5 %

表 5.2 より 1 年次の 6 月に欠席した回数より成績が左右されることが分かり, 成績が中間から上の人は 1 年次の 6 月にあまり欠席せず, 下の人は欠席回数が多くなることが分かった.

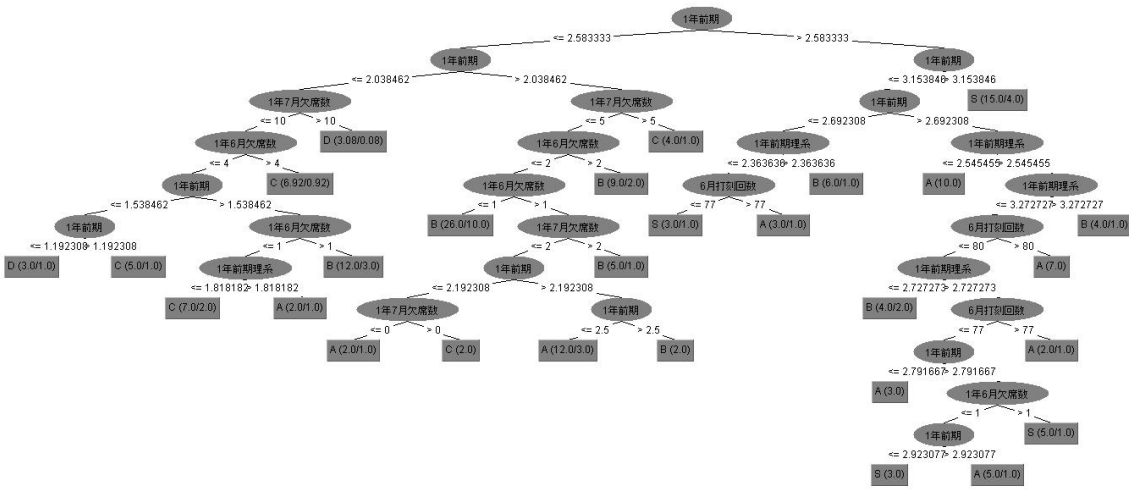

図 5.2: 決定木によって構築されたモデル

図 5.2 より 1 年後期の成績はある程度成績 1 年前期の成績により成績上位と成績下位に分けられるが,6 月の欠席回数と 7 月の欠席回数が多いと成績が悪くなる傾向にあることが分かる. 以上より 1 年後期を予測するには 6 月の欠席回数が関わっているように思える. これは大学に慣れてきはじめ 6 月あたりから授業を欠席する人がいるからではないかと考えられる.

5.2　1 年通年における構築モデル

1 年通年は 1 年前期までのデータを用いて予測を行っている. 表 4.39 より削減方法としてはベイジアンネットワーク, 決定木共に情報利得で有効性の高い変数から 5 個取捨選択している. このとき選択された変数を表 5.3 に示す.

表 5.3: 情報利得により取捨選択された変数

番号	変数名
3	1 年 6 月打刻回数
58	1 年前期
62	1 年前期理系
63	1 年前期専門
65	1 年前期数学

取捨選択された 5 つの変数と目的変数である 1 年通年を離散化したデータをベイジアンネットワークと決定木で予測した. ベイジアンネットワークによって構築されたモデルをを図 5.3 に示す.

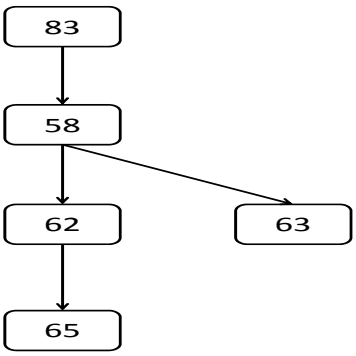

図 5.3: ベイジアンネットワークによって構築されたモデル

図 5.3 によりベイジアンネットワークにおいて打刻データが使われないほうが予測的中率は高くなることが分かった. また決定木によって構築されたモデルを図 5.4 に示す.

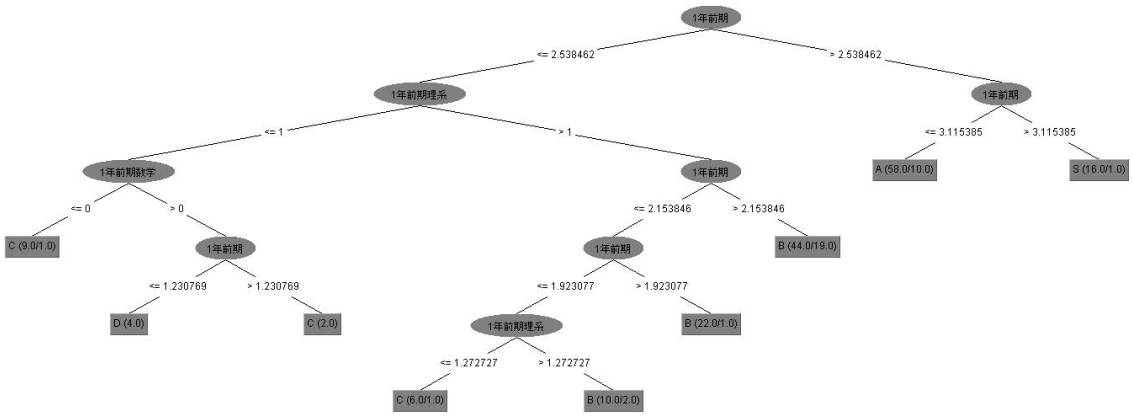

図 5.4: 決定木によって構築されたモデル

図 5.4 より決定木についても打刻データが使われない方が予測的中率は高くなることが分かった. これより 1 年通年を予測するのに 1 年前期までの打刻データでは情報が少なく成績データだけで予測する方が的中しやすいのではないかと考えられる.

5.3　2 年後期における構築モデル

2 年通年は 2 年前期までのデータを用いて予測を行っている. 表 4.39 よりベイジアンネットワーク, 決定木共に情報利得を用いて削減することにより予測的中率が高くなっている. ベイジアンネットワークでは情報利得で有用性の高い変数から 30 個を取捨選択したデータが最も予測的中率が高く, 決定木では情報利得で有用性の高い変数から 35 個を取捨選択したときが最も高い予測的中率を示した. このとき選択された変数を表 5.4 に示す.

表 5.4: 情報利得により取捨選択された変数

番号	変数名
74	2 年前期
42	2 年 7 月打刻回数
82	1 年後期
50	2 年前期水曜打刻平均
47	2 年前期祝日前後欠席回数
29	1 年後期月曜打刻平均
44	2 年 5 月欠席回数
26	1 年 12 月欠席回数
45	2 年 6 月欠席回数
62	1 年前期理系
81	2 年前期数学
46	2 年 7 月欠席回数
28	1 年祝日前後欠席回数
79	2 年前期専門
58	1 年前期
23	1 月打刻回数
65	1 年前期数学
40	2 年 5 月打刻回数
51	2 年前期木曜打刻回数
63	1 年前期専門
25	1 年 11 月欠席回数
27	1 年 1 月欠席回数
41	2 年 6 月打刻回数
30	1 年後期火曜打刻平均
70	1 年後期理系
21	1 年 11 月打刻回数
57	2 年前期金曜打刻分散
72	1 年後期理科

52	2 年前期金曜打刻平均
31	1 年後期水曜打刻平均
75	2 年前期英語
78	2 年前期理系
71	1 年後期専門
80	2 年前期理科
64	1 年前期理科

構築されたモデルを図 5.5, 図 5.6 に示す.

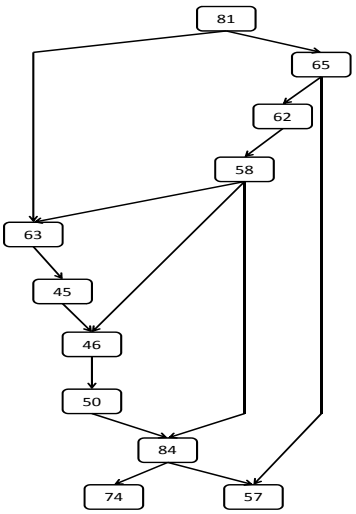

図 5.5: ベイジアンネットワークによって構築されたモデル

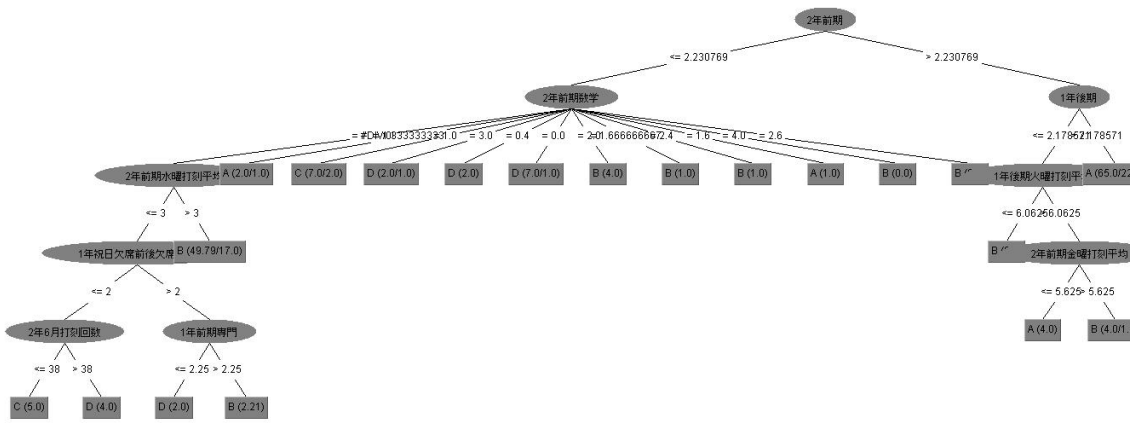

図 5.6: 決定木によって構築されたモデル

図 5.5 により 2 年後期の予測は 2 年前期の水曜に打刻する平均回数と 2 年前期金曜の打刻回数の分散に関わっていることが分かった. そのときの条件付き確率を表 5.5, 表 5.6 に示す.

表 5.5: 2 年前期水曜打刻平均による条件付き確率

2 年前期水曜打刻平均	1 年前期	S	A	B	C	D
3.38 以下	1.98 以下	2.2 %	2.2 %	20 %	24.4 %	51.1 %
3.38 以下	1.98 より上	1.8 %	5.5 %	49.1 %	30.9 %	12.7 %
3.38 より上	1.98 以下	3 %	21.2 %	21.2 %	33.3 %	21.2 %
3.38 より上	1.98 より上	13.5 %	45.9 %	37.1 %	3.1 %	0.4 %

表 5.6: 2 年前期金曜打刻分散における条件付き確率

1 年前期数学	2 年後期	4.53 以下	4.53 より上
0.9 以下	S	50 %	50 %
0.9 以下	A	12.5 %	87.5 %
0.9 以下	B	13.6 %	86.4 %
0.9 以下	C	7.1 %	92.9 %
0.9 以下	D	34.6 %	65.4 %
0.9 より上	S	96.9 %	3.1 %
0.9 より上	A	75 %	25 %
0.9 より上	B	85.8 %	14.2 %
0.9 より上	C	34.4 %	65.6 %
0.9 より上	D	8.3 %	91.7 %

表 5.5 より 2 年前期の打刻平均が 3.38 より高いと成績は良くなる傾向にあるが 1 年前期の成績によっては成績が下位の方になる. 表 5.6 より 1 年前期数学の成績が 0.9 より上ならば 2 年前期の打刻回数の分散が 4.53 より低いと 2 年後期の成績は良い傾向が見られる.

また図 5.6 より 2 年後期の予測はある程度, 成績データにより成績上位か下位かに分けられるがさらに詳しく成績予測するために打刻データが用いられていることがわかる.

5.4　1,2 年複合における構築モデル

1 年通年は 1 年前期までのデータを用いて予測を行っている. 表 4.39 より削減方法としてベイジアンネットワークは情報利得で有効性の高い変数から 20 個取捨選択しており, 決定木は主成分分析により次元削減している. このとき情報利得により取捨選択された 20 個の変数を表 5.7 に示し, 主成分分析した結果を図 5.7 に示す.

表 5.7: 情報利得により取捨選択された変数

番号	変数名
52	2 年前期金曜打刻平均
26	1 年 12 月欠席回数
43	2 年 4 月欠席回数
74	2 年前期
66	1 年後期
58	1 年前期
3	1 年 6 月打刻回数
29	1 年後期月曜打刻平均
22	1 年 12 月打刻回数
9	1 年前期祝日前後欠席回数
4	1 年 7 月打刻回数
49	2 年前期火曜打刻平均
42	2 年 7 月打刻回数
11	1 年前期火曜打刻平均
47	2 年前期祝日前後欠席回数
50	2 年前期水曜打刻平均
8	1 年 7 月欠席回数
62	1 年前期理系
25	1 年 11 月欠席回数
30	1 年後期火曜打刻平均
44	2 年 5 月欠席回数
81	2 年前期数学
63	1 年前期専門
23	1 年 1 月打刻回数
27	1 年 1 月欠席回数
24	1 年 10 月欠席回数
71	1 年後期専門
21	1 年 11 月打刻回数
65	1 年前期数学
41	2 年 6 月打刻回数
79	2 年前期専門
59	1 年前期英語
31	1 年後期水曜打刻平均
20	1 年 10 月打刻回数
5	1 年 4 月欠席回数

主成分負荷量

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分	第6主成分	第7主成分	第8主成分	第9主成分	第10主成分	第11主成分
1年祝日前後欠席回数	0.721825315	-0.410700631	-0.128400584	0.139700012	-0.248586552	-0.013513524	-0.055784497	0.001594366	0.048595623	0.043155061	0.092219195
1年4月打刻回数	-0.403780112	-0.077673377	-0.372280692	-0.177998189	0.03769513	0.381140028	0.117099061	-0.060636415	0.398152721	-0.033104878	0.062566582
1年5月打刻回数	-0.628627823	0.012783404	-0.382303594	-0.118260986	-0.221199715	0.388986742	0.133801596	0.182158319	0.142666611	0.000749247	-0.109631528
1年6月打刻回数	-0.800669974	-0.208395609	-0.252504174	-0.192424718	-0.245475033	0.127906398	-0.05408721	0.144527137	0.016948638	0.117825927	-0.027147628
1年7月打刻回数	-0.805745928	-0.150637998	-0.20099732	-0.251751027	-0.168280444	0.174889128	-0.034087645	0.183290785	-0.047894136	0.067237863	-0.001509646
1年10月打刻回数	-0.788081417	0.0067064	-0.125691603	-0.185429114	0.267999782	-0.135764865	0.250499932	0.002510478	-0.040806286	-0.156583472	-0.254331412
1年11月打刻回数	-0.886715816	0.012116335	-0.132988388	-0.11971395	0.101670116	-0.165614964	0.102708146	-0.124338175	0.045411125	-0.059260447	-0.013466705
1年12月打刻回数	-0.850504543	0.070316564	-0.300980317	-0.128092025	0.071675176	-0.202259838	0.056930509	-0.187511358	0.00174545	-0.042892626	0.046244023
1年1月打刻回数	-0.839501152	-0.069300266	-0.27926661	-0.097832603	0.008311092	-0.270582749	0.043241534	-0.152852953	-0.041944485	-0.019426949	0.071306382
1年4月欠席回数	0.396927772	0.166030053	-0.122040349	0.428133637	-0.013888612	-0.412968818	0.232795831	0.502932334	0.227511807	0.167894302	-0.076131944
1年5月欠席回数	0.398221379	0.145556755	-0.108244166	0.43326795	-0.055399683	-0.426885252	0.202098338	0.522135211	0.203281702	0.142668311	-0.094339373
1年6月欠席回数	0.666769886	-0.350070149	-0.021756244	0.175076038	0.042835306	-0.313111914	0.202211296	-0.159418174	0.027505405	-0.318753482	0.026084352
1年7月欠席回数	0.716337074	0.194504018	-0.237437952	0.338953209	-0.116256475	-0.07584495	0.177740931	-0.057390053	-0.003068643	-0.002399849	0.062779824
1年10月欠席回数	0.583418977	-0.2940087	0.031470621	0.243210318	-0.115582065	0.105885008	0.017849794	-0.162361318	0.06134269	0.067316522	0.480980954
1年11月欠席回数	0.75261079	-0.224527181	-0.122188806	0.207887033	-0.061305895	0.196308698	0.220442867	0.108650723	-0.056888118	-0.120664516	0.053125209
1年12月欠席回数	0.59361077	-0.328035589	0.061916418	0.21193679	-0.06066694	0.496320511	0.181480732	0.002980211	-0.043906677	-0.116631281	-0.11277385
1年1月欠席回数	0.632367091	-0.226703124	0.013849581	0.223149164	0.027458947	0.486666613	0.199356554	-0.045804996	-0.089512864	-0.013282473	-0.224949363
1年前期月曜平均	-0.501419729	0.062906418	-0.583171142	-0.058174368	-0.169179492	0.292839377	0.08017022	0.093583757	0.321785042	-0.125937354	0.143065531
1年前期月曜分散	0.167973809	-0.095674019	-0.675511206	0.29855618	0.005910044	-0.038916991	-0.058254686	-0.141339635	0.066457167	-0.114247383	-0.07993456
1年前期火曜平均	-0.794951342	-0.184819841	-0.130046201	-0.24338193	-0.13460593	0.198402611	0.026093919	0.172404701	-0.056846722	0.106921961	-0.140638127
1年前期火曜分散	0.481702478	-0.230769007	-0.229432237	0.302904079	0.186092778	-0.093318648	0.05975638	-0.389226035	0.155978733	0.063510836	-0.036337305
1年前期水曜平均	-0.459739925	-0.745270041	-0.18865557	-0.102204966	-0.315370684	-0.064134547	-0.05861801	0.070853034	-0.046568541	0.092922058	-0.043299026
1年前期水曜分散	0.078351359	-0.835093908	0.274044738	-0.091809501	-0.042075131	-0.194338511	0.080068735	-0.035256189	0.082660689	-0.086566219	0.015609017
1年前期木曜平均	-0.071397996	0.905598922	-0.182294935	0.087171126	0.102963381	0.206616554	-0.05985933	-0.01141492	-0.070271407	0.055942296	-0.057211677
1年前期木曜分散	0.365680177	0.130822712	-0.629808329	0.366340211	-0.06437227	-0.101968092	0.02114895	-0.032329155	0.001559913	-0.038362434	-0.031674466
1年前期金曜平均	-0.267095001	-0.662096012	0.490199505	-0.348343213	0.013765309	-0.003847639	0.107023806	0.132152362	0.042298599	0.025349673	0.019657681
1年前期金曜分散	0.012199443	-0.802366493	0.318852998	-0.181015388	-0.041676303	-0.033852436	0.031943863	-0.006151129	0.006746128	0.072107227	0.165029742
1年後期月曜平均	-0.689811416	0.074936744	-0.187540853	-0.230075095	0.158393007	-0.046897152	0.202914993	0.004469329	0.31663367	-0.252362188	0.206575914
1年後期月曜分散	0.032060899	-0.290057547	-0.136664324	-0.152101762	0.455608133	0.103719753	0.169185503	0.066821628	0.366075963	0.120109068	0.226771224
1年後期火曜平均	-0.845830407	0.042926318	-0.17541226	-0.094229657	0.153203594	-0.253189136	0.089377893	-0.150984084	-0.073504103	0.007476059	-0.071809008
1年後期火曜分散	0.068979922	-0.243462012	-0.363730844	0.188172518	0.279870656	-0.01191517	0.174954633	-0.263130542	-0.154346665	0.53308296	-0.004923535
1年後期水曜平均	-0.821782189	0.112260117	-0.190472816	-0.114783238	0.109513832	-0.176974711	0.170578997	-0.162230698	-0.128196381	-0.085479898	-0.129143743
1年後期水曜分散	0.307081087	-0.502829571	-0.103722061	0.023946746	0.389818452	0.171863361	0.340671259	-0.092687258	-0.16512366	0.150061204	-0.114688198
1年後期木曜平均	-0.613762482	-0.528144553	-0.374851707	-0.046598457	-0.22188873	-0.257684321	-0.035059445	-0.061749964	-0.061125348	0.026518499	-0.066903166
1年後期木曜分散	-0.203521856	-0.73178517	-0.388315735	0.127787956	-0.174051473	-0.033462201	-0.059024319	0.014161108	-0.129403564	0.093633481	-0.16532301
1年後期金曜平均	-0.224494329	0.718426217	0.305833192	-0.073246497	0.460525271	0.103532908	0.126041474	-0.058422822	-0.010882143	-0.050650732	-0.02578741
1年後期金曜分散	0.007681545	0.107498393	0.043020754	-0.043039	0.654336666	0.134099047	0.366225938	0.057868643	-0.019051101	0.174383606	-0.01993971
2年4月打刻回数	-0.69991789	-0.170715145	0.067437078	0.452725609	0.220901823	0.04206861	-0.176137574	-0.0141049	0.071006856	-0.06671818	0.036953564
2年5月打刻回数	-0.801797995	-0.120103945	0.198923861	0.442902219	0.125504852	0.031049589	-0.000694484	0.097300213	-0.051765048	-0.011107364	0.121912721
2年6月打刻回数	-0.778586224	-0.106980999	0.290990128	0.431182226	0.069758534	0.037476191	0.077904988	0.044817097	-0.023217091	-0.075272098	0.02954617
2年7月打刻回数	-0.792285487	-0.041698929	0.041731211	0.445052821	0.003637084	0.034482648	0.063645482	0.069270029	-0.166163826	-0.113218184	0.084073876
2年4月欠席回数	0.587761022	-0.046393843	-0.267779312	-0.129953736	-0.012652297	-0.05991876	0.330354556	0.024892579	-0.311977434	-0.187241547	0.051065447
2年5月欠席回数	0.674970201	-0.044636403	-0.289400139	-0.26192783	0.176742442	-0.063424275	0.009901294	0.115289226	-0.142064862	-0.311877398	0.053703113
2年6月欠席欠席回数	0.511030275	0.026992591	-0.395184876	-0.283480954	0.36895229	-0.262200726	-0.222637047	0.220010388	-0.080663296	0.002087401	0.151152709
2年7月欠席回数	0.377278676	-0.068526392	-0.167713607	-0.418741092	0.521554674	-0.000568052	-0.307013254	0.10721496	0.015864758	0.185064352	-0.048013145
2年前記祝日 前後欠席回数	0.726256283	-0.065222545	-0.156711793	-0.275218566	-0.148116515	-0.046416307	0.029816209	-0.024481341	0.023736068	-0.199622453	-0.128671313
2年前期月曜平均	-0.48817628	-0.178443846	0.41574927	0.455706422	-0.075962996	0.024472954	-0.06959783	-0.116296852	0.257272232	-0.117750302	-0.22471077
2年前期月曜分散	-0.022738762	-0.191066899	0.089173104	0.444647526	0.237340291	-0.056998305	-0.287276007	-0.208271129	0.369780553	-0.007251284	-0.364590944
2年前期火曜平均	-0.601303771	-0.230017957	0.228916753	0.245484596	0.31700471	-0.029735534	0.034283914	0.130021949	0.103735519	-0.065031546	0.138423657
2年前期火曜分散	0.284547032	-0.306998414	-0.188422093	-0.009528532	0.547246443	0.074881554	-0.248077647	-0.02080339	0.130397058	-0.051671209	0.045518529
2年前期水曜平均	-0.835475121	0.008974007	0.113807645	0.339192344	0.071799172	-0.039548444	0.102595976	0.013158924	-0.128990779	0.086914585	0.101188351
2年前期水曜分散	0.019222081	-0.522087354	0.18691656	-0.140389937	0.521494112	-0.042254098	-0.090614129	0.109429637	-0.112543785	-0.129317911	0.003808872
2年前期木曜平均	-0.690748528	0.089486958	-0.126387108	0.445956292	-0.081242909	0.07372251	0.000708895	-0.053588367	-0.210552268	-0.011397831	0.236954068
2年前期木曜分散	-0.126053861	0.003414107	-0.490337529	0.416290087	0.182397184	0.155989768	-0.352216879	0.026382355	-0.147466209	0.001000296	0.182005391
2年前期金曜平均	-0.676956785	-0.098458322	0.049609577	0.355275945	0.146839627	0.098083556	-0.007730077	0.241820698	-0.258529767	-0.12621346	0.028576088
2年前期金曜分散	0.147240756	-0.412049497	-0.301299206	0.18768788	0.424114095	0.154415295	-0.252630888	0.291691674	-0.140270793	-0.216743453	-0.168637654
固有値	18.59599962	6.533317909	4.271100276	4.008219649	3.257142832	2.120860011	1.50688658	1.461383278	1.352852243	1.077820795	1.062736909
寄与率	33.20714218	11.66663912	7.626964779	7.157535087	5.816326486	3.787250019	2.690868893	2.609612996	2.415807577	1.924679991	1.89774448
累積寄与率	33.20714218	44.87378131	52.50074608	59.65828117	65.47460766	69.26185768	71.95272657	74.56233957	76.97814714	78.90282713	80.80057161

図 5.7: 主成分分析の結果

またそのときベイジアンネットワークと決定木により構築されたモデルを図 5.8, 図 5.9 に示す. なお主成分分析は累積寄与率が 80 %を越える第 11 主成分までを用いている.

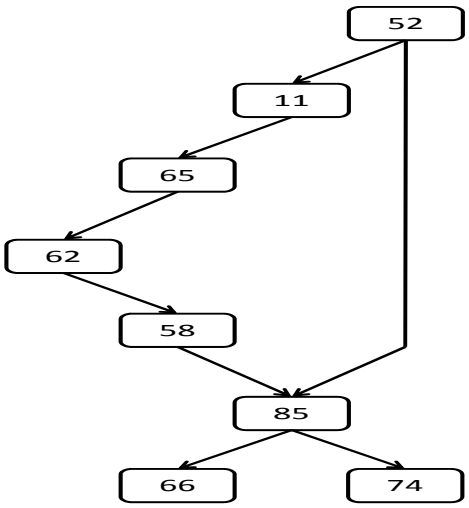

図 5.8: ベイジアンネットワークによって構築されたモデル

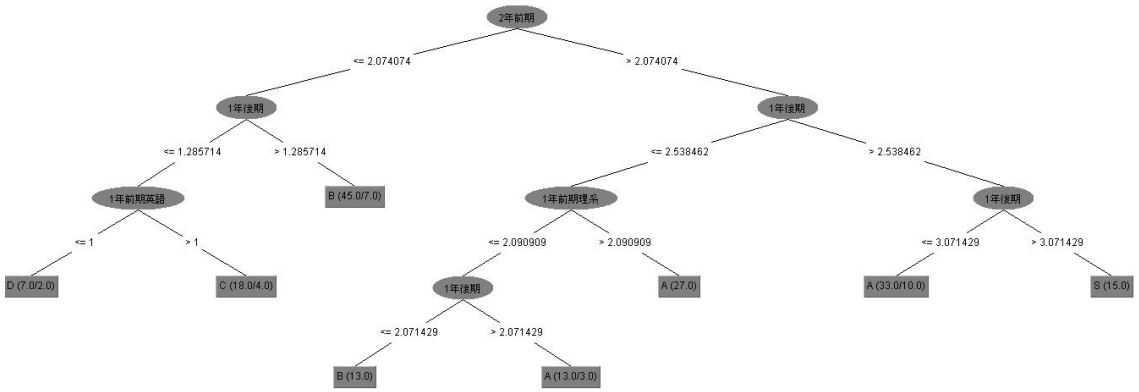

図 5.9: 決定木によって構築されたモデル

図 5.8 より 1,2 年複合と関係がありそうな打刻データとして 2 年前期金曜打刻回数の平均が挙げられる.2 年前期金曜打刻平均における条件付き確率は表 5.8 のようになった.

表 5.8: 2 年前期金曜打刻平均による条件付き確率

1 年前期	2 年前期金曜打刻平均	S	A	B	C	D
1.98 以下	2.31 以下	7.7 %	7.7 %	7.7 %	7.7 %	69.2 %
1.98 以下	2.31 より上	1.5 %	4.6 %	38.5 %	50.8 %	4.6 %
1.98〜2.59	2.31 以下	20 %	20 %	20 %	20 %	20 %
1.98〜2.59	2.31 より上	2.2 %	28.1 %	61.2 %	7.9 %	0.7 %
2.59〜3.13	2.31 以下	20 %	20 %	20 %	20 %	20 %
2.59〜3.13	2.31 より上	18.6 %	69.9 %	9.7 %	0.9 %	0.9 %
3.13 より上	2.31 以下	20 %	20 %	20 %	20 %	20 %
3.13 より上	2.31 より上	78.4 %	13.5 %	2.7 %	2.7 %	2.7 %

表 5.8 より 1 年前期の成績が 1.98 以下だと 1,2 年複合の成績は中間より下位になる傾向が見られる. また 1 年前期の成績が 3.13 より上だと 1,2 年複合の成績は中間より上位になる傾向が見られる. また全体的に 2 年前期金曜における打刻回数の平均が 2.31 より上だと成績が良くなる傾向が見られる.

図 5.9 は打刻データ部分について主成分分析を行い, 次元削減したデータと成績データを決定木により構築したモデルである. このモデルより主成分データは使われず成績データのみで構築されていることがわかる. これは細かい部分の分類について打刻データのとある変数が使われている可能性があるが, 主成分分析により複数の変数が統合されてしまったため細かい分類ができなくなったことが原因として考えられる.

ここである曜日における打刻回数の平均について考えてみる. 平均の値が低いときに考えられるパターンとしては

- その曜日に受ける授業が少ない
- その曜日はよく休む
- 遅れてくることが多い, または打刻忘れが多い

の 3 パターンが考えられる. 平均から特徴を求めるにあたり考えられるパターンが 3 つあるためどれが正確であるか特定できない. そこで平均の代わる変数として曜日毎の欠席回数と曜日毎の打刻回数の最頻値を用いる. 新たな変数を用いるにあたり平均を求めた変数は除く. またベイジアンネットワークでは変数同士が独立であることが望ましいので曜日毎の欠席回数を用いる代わりに月毎の欠席回数と祝日全後の欠席回数は除く. 新たに置き換えた変数を表 5.9 に示す.

表 5.9: 説明変数として新たに加えた打刻データ

番号	変数名	意味
86	1 年前期月曜欠席回数	1 年前期の月曜日に欠席した回数
87	1 年前期火曜欠席回数	1 年前期の火曜日に欠席した回数
88	1 年前期水曜欠席回数	1 年前期の水曜日に欠席した回数
89	1 年前期木曜欠席回数	1 年前期の木曜日に欠席した回数
90	1 年前期金曜欠席回数	1 年前期の金曜日に欠席した回数
91	1 年後期月曜欠席回数	1 年後期の月曜日に欠席した回数
92	1 年後期火曜欠席回数	1 年後期の火曜日に欠席した回数
93	1 年後期水曜欠席回数	1 年後期の水曜日に欠席した回数
94	1 年後期木曜欠席回数	1 年後期の木曜日に欠席した回数
95	1 年後期金曜欠席回数	1 年後期の金曜日に欠席した回数
96	2 年前期月曜欠席回数	2 年前期の月曜日に欠席した回数
97	2 年前期火曜欠席回数	2 年前期の火曜日に欠席した回数
98	2 年前期水曜欠席回数	2 年前期の水曜日に欠席した回数
99	2 年前期木曜欠席回数	2 年前期の木曜日に欠席した回数
100	2 年前期金曜欠席回数	2 年前期の金曜日に欠席した回数
101	1 年前期月曜打刻最頻値	1 年前期月曜日打刻回数において頻繁に出てくる値
102	1 年前期火曜打刻最頻値	1 年前期火曜日打刻回数において頻繁に出てくる値
103	1 年前期水曜打刻最頻値	1 年前期水曜日打刻回数において頻繁に出てくる値
104	1 年前期木曜打刻最頻値	1 年前期木曜日打刻回数において頻繁に出てくる値
105	1 年前期金曜打刻最頻値	1 年前期金曜日打刻回数において頻繁に出てくる値
106	1 年後期月曜打刻最頻値	1 年後期月曜日打刻回数において頻繁に出てくる値
107	1 年後期火曜打刻最頻値	1 年後期火曜日打刻回数において頻繁に出てくる値
108	1 年後期水曜打刻最頻値	1 年後期水曜日打刻回数において頻繁に出てくる値
109	1 年後期木曜打刻最頻値	1 年後期木曜日打刻回数において頻繁に出てくる値
110	1 年後期金曜打刻最頻値	1 年後期金曜日打刻回数において頻繁に出てくる値
111	2 年前期月曜打刻最頻値	2 年前期月曜日打刻回数において頻繁に出てくる値
112	2 年前期火曜打刻最頻値	2 年前期火曜日打刻回数において頻繁に出てくる値
113	2 年前期水曜打刻最頻値	2 年前期水曜日打刻回数において頻繁に出てくる値
114	2 年前期木曜打刻最頻値	2 年前期木曜日打刻回数において頻繁に出てくる値
115	2 年前期金曜打刻最頻値	2 年前期金曜日打刻回数において頻繁に出てくる値

ベイジアンネットワークで構築された図 5.8 のモデルを見てみると平均が打刻と関係ありそうなので平均の代わりに表 5.9 に書かれた変数を用いてもう一度情報利得をもとにモデルを再構築してみた. 情報利得した結果を表 5.10 に示す.

表 5.10: 情報利得による成績予測的中率

取捨選択数	1,2 年複合
5 個	69.5906 %
10 個	77.193 %
15 個	80.7018 %
20 個	80.7018 %

表 4.25 と表 5.10 を見比べると, 情報利得によって取捨選択した数が 15 個のとき 80.7018 %と変数を置き換える前に比べて. 高い予測的中率を示した. これより打刻データが成績データに関係している部分が少なくともあることがわかった. またそのときのベイジアンネットワークモデルを図 5.10 に示す.

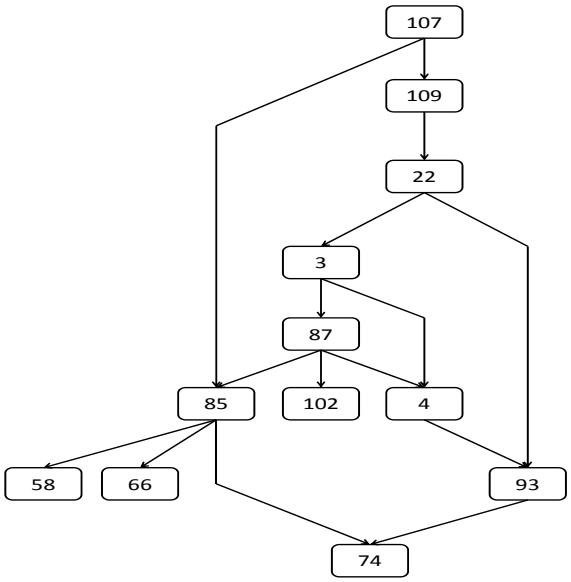

図 5.10: 再構築されたモデル

図 5.10 より 1 年前期火曜日の欠席回数と 1 年後期火曜日の打刻回数の最頻値が 1,2 年複合に関係あることがわかる. そのときの条件付き確率を表 5.11 に示す.

表 5.11: 2 年前期金曜打刻分散における条件付き確率

1 年前期火曜欠席	1 年後期火曜最頻値	S	A	B	C	D
1.5 以下	2.5 以下	7.7 %	7.7 %	23.1 %	38.5 %	23.1 %
1.5 以下	2.5 より上	16.4 %	42.9 %	33.8 %	6.6 %	0.3 %
1.5〜4.5	2.5 以下	20 %	20 %	20 %	20 %	20 %
1.5〜4.5	2.5 より上	10.6 %	2.1 %	44.7 %	40.4 %	2.1 %
4.5 より上	2.5 以下	9.1 %	9.1 %	9.1 %	9.1 %	63.6 %
4.5 より上	2.5 より上	11.1 %	11.1 %	11.1 %	33.3 %	33.3 %

表 5.11 より 1 年後期火曜の最頻値が 2.5 以下だと成績が下位になりやすい傾向があることが分かった. また 1 年前期火曜の欠席回数が多いと成績が悪くなる傾向があることがわかった.

第6章　むすび

本研究では, ベイジアンネットワークと決定木による予測技術を用いて構築されたモデルから打刻データの有用性について特徴分析した.2 章では本研究で用いた理論について説明した.3 章では元となるデータとその拡張について述べ, また予測する GPA の離散化について解説した.4 章では目的変数となる GPA が最も高い予測的中率を示した手法を述べた.5 章では 4 章で最も高い予測的中率を示した手法により構築されるモデルが打刻データと目的変数の間に関係があるのかどうかについて検証した.

目的変数を予測するため前期までのデータからの予測と後期半学期までのデータからの予測を比べ, 後期半学期までのデータから予測した方が予測的中率が高くなったが前期までのデータからの予測でも後期半学期の予測に近い予測的中率を示した. 本研究ではできるだけ早い段階で予測したいため前期までのデータから予測することにした. できるだけ予測的中率を高くしたかったため変数の削減を情報利得と主成分分析で行い、その結果, 全ての目的変数に関して予測的中率を上げることができた。データの削減を有用なものには行い, ベイジアンネットワーク, 決定木を leave one out 法により分類しモデルを構築した結果,1 年前期までのデータから 1 年後期, 1 年通年の成績を予測するには打刻データの情報が少なくモデルに打刻データによる特徴があまり見られなかった.2 年前期までのデータから 2 年後期, 1,2 年複合の成績を予測することに関しては, ある曜日による打刻回数の平均が成績予測と関係がありそうだということがわかった. ここで平均についてもう少し詳しく検証してみるために曜日毎の欠席回数と曜日毎の打刻回数の最頻値を変数として 新たに置き換えもう一度予測することで予測的中率をあげることができた. これより打刻データが成績データに関係していることがわかった.

ただし, 今回の研究で用いたデータは名古屋工業大学のある年度におけるデータなので, 違う年度のカリキュラムが変わっていたり他の大学の打刻データを用いて本研究で行った操作を同様に検証すると違った特徴が出てくることが考えられる. そのため今後は大学の人数による相対値や, カリキュラムより必須科目や選択科目の単位数を取り入れることで世間一般的に打刻が成績とどう関係しているか検証することが必要であると言える.

謝辞

本研究を進めるにあたって，日頃から多大な御尽力を頂き，ご指導を賜りました名古屋工業大学，舟橋健司　准教授，伊藤宏隆　助教，山本大介　准教授　に心から感謝致します.

また, 本研究の実験のためのデータの提供元である, 出欠システム及びコースマネージメントシステムの開発に尽力されました, 名古屋工業大学情報基盤センター長　松尾啓志　教授，内匠逸　教授，情報基板センター教職員の皆様に心から感謝いたします.

そして，本研究に対して御討論頂きました本学　中村研究室の皆様ならびに中部大学　岩堀研究室の皆様に深く感謝致します.

最後に，本研究に多大な御協力頂きました舟橋研究室諸氏に心から感謝致します.

参考文献

[1] 伊藤宏隆, 舟橋健司, 中野智文, 内匠逸, 松尾啓志, 大貫徹,“名古屋工業大学における Moodle の構築と運用”, メディア教育研究,4 巻,2 号,15-212008

[2] 松尾啓志,“情報基板システムが支えるケータイ世代の学びの場とは?”, サイエンティフィック・システム研究会,2009 年度教育環境分科会, 第 1 回会合

[3] 堀江匠, “データマイニングによる学生の修学傾向分析とその修学指導への適用有効性の検証”, 平成 20 年度名古屋工業大学卒業研究論文, 2008

[4] 伊藤宏隆, 舟橋健司, 内匠逸, 松尾啓志, “IC カード出欠データと CMS 学習データを用いたデータマイニング”, メディア教育研究,4,2, pp.15-21, 2008

[5] 伊藤暁人, “ニューラルネットワークによる学生の成績予測とその学習指導への適用可能性の検討”, 平成 22 年度名古屋工業大学卒業研究論文, 2010

[6] 伊藤圭祐,“ベイジアンネットワークを用いた学生の修学傾向予測とその有用性の検証” 平成 23 年度名古屋工業大学卒業研究論文,2011

[7] 鈴木崇広,“決定木によるデータマイニングの比較”

[8] Weka
http://www.cs.waikato.ac.nz/ml/weka/

[9] 本村陽一, “ベイジアンネットワーク:入門からヒューマンモデリングへの応用まで”, 行動計量学会セミナー資料,2004

[10] 石井一夫, “図解よくわかるデータマイニング”, 日刊工業新聞社 (2004)