研究雑まとめ

MBTIデータセットにおける特徴分析

田中直哉

1. 研究背景および目的

近年SNS上で16性格診断と題した高精度な性格診断としてMBTIが存在する。

　しかしKaggle上に存在するMBTIのデータセットは本来の質問数である93個の特徴量に対して8個しか存在しない。

* 1. データセットについて

Kaggleより引用(Google翻訳を使用)

説明

この合成データセットは、人口統計学的要因、興味分野、および性格スコアの組み合わせに基づいて、マイヤーズ・ブリッグス・タイプ指標（MBTI）の性格タイプを探索および予測するために設計されています。10万以上のサンプルが含まれており、それぞれがMBTIタイプを決定する様々な特徴を持つ個人を表しています。このデータセットは、異なる性格特性と、年齢、性別、教育、興味などの外的要因との相関関係を研究するために使用できます。

データセットの説明

Age:個人の年齢を表す連続変数。

Gender:個人の性別を示すカテゴリ変数。可能な値は「男性」と「女性」です。

Education:バイナリ変数。値が 1 の場合、個人は少なくとも大学院レベル (またはそれ以上) の教育を受けていることを示し、値が 0 の場合、大学、高校レベル、または無学であることを示します。

Interest:個人の主な興味分野を表すカテゴリ変数。

Introversion Score： 0から10までの連続変数で、個人の内向性または外向性の傾向を表します。スコアが高いほど、外向性の傾向が強いことを示します。

Sensing Score： 0から10までの連続変数で、感覚と直観のどちらを重視するかを表します。スコアが高いほど、感覚が重視されることを示します。

Thinking Score： 0から10までの連続変数で、思考と感情のどちらを優先するかを示します。スコアが高いほど、思考を優先する傾向があります。

Judging Score： 0から10までの連続変数で、知覚よりも判断を好む傾向を表します。スコアが高いほど、判断を好む傾向を示します。

Personality:人物パーソナリティタイプを含むターゲット

引用終わり

表1-1　データセット最初の5行

グラフィカル ユーザー インターフェイス, アプリケーション

AI によって生成されたコンテンツは間違っている可能性があります。

* 1. MBTIの特徴量

先述した通りMBTIは93個の質問によって精度が100%となっているが本データセットは8個の特徴量であらわされているため、その特徴量の妥当性を検証する。また、質問と関係ない特徴量に対してMBTIへの影響度を測定することに加えデータ構造を分析することで各変数の特徴を分析する。

1. 研究手法
   1. 精度の測定

　クロスバリデーションで500分割して精度の検証を行う。

　目的としては特徴量の妥当性および2.2項における分類寄与率の測定において精度を担保する上で測定を行う。

* 1. 分類寄与率の測定

　GBDTは決定木系アルゴリズムであるため一つの変数でどれだけ分類できるか測定することができる。

そこで、各変数がMBTIを分類するうえでの影響度を測定する。

* 1. 間違った傾向についての分析

　Hold-out方式で訓練データ70%でテストデータ30%とし、2.1項において精度の代表値に近い精度を出したときにClassification\_Report(各目的変数の値毎に)と混合行列を出力し、誤った傾向について考察する。

* 1. MBTI別特徴量のヒストグラム

　MBTI別に特徴量をヒストグラムで画像を保存する。この画像をMBTI別に比較することでMBTI別の特徴量の傾向を分析する。

1. 結果
   1. 精度の測定

　クロスバリデーションで500回精度の検証を行った結果得られたようやく統計量を表3-1に記す。また、ヒストグラムを図3-1に示す。

表3-1　500回クロスバリデーションした結果得られたようやく統計量

テーブル

AI によって生成されたコンテンツは間違っている可能性があります。

表3-1の結果から平均値および中央値ともに約89%という事が分かる。

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図3-1　500回クロスバリデーションを行った際に得られた精度のヒストグラム

　図3-1の結果から500回クロスバリデーションを行い最頻値は90%程度であると考えられる。

* 1. 分類寄与率の測定

　GBDTでMBTIにおける各特徴量について89%の信頼度で分類寄与率を測定した結果を表3-2に記す。

表3-2　GBDTを用いた分類寄与率(信頼度90%)

グラフィカル ユーザー インターフェイス, テキスト, チャットまたはテキスト メッセージ

AI によって生成されたコンテンツは間違っている可能性があります。

この結果から各種スコアと年齢がMBTIに関係していることが考えられる。

一方で教育や性別、興味などはMBTIの分類結果に寄与しにくいことも考えられる。

* 1. 間違った傾向についての分析

　精度が代表値だった時のClassification\_Reportを図3-2に、混合行列を表3-3に記す。

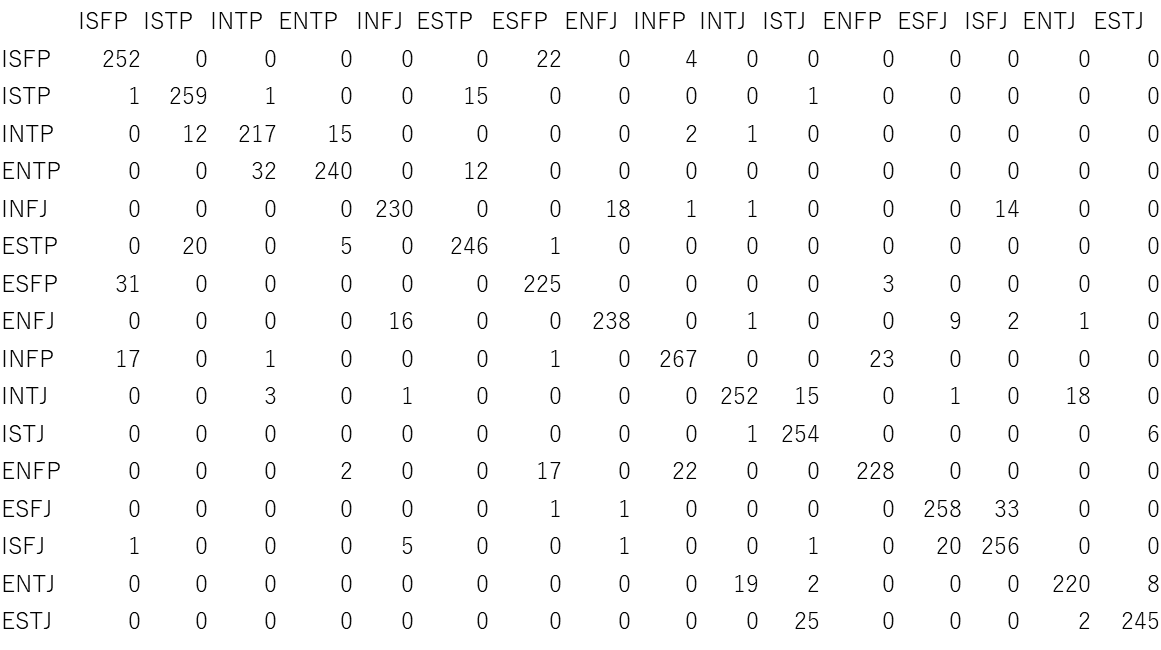
テーブル

AI によって生成されたコンテンツは間違っている可能性があります。

図3-2　MBTIを特徴量で分類予測した時のClassification\_Report

f1スコアに着目すると、どのMBTIも正解率の代表値程度(誤差±3程度)に間違えていることが分かる。

表3-3　MBTIを分類予測した際の混合行列



　間違えた傾向としては、まずIとE(内向的か外向的か)で間違えており、その次にN(直感型)とS(感覚型)で間違えている。

間違ている項目を見てみるとIとEは多くて正解に対して10％以上間違えており、SとNは間違えた場合は多くて5％程度間違えている。

* 1. MBTI別特徴量のヒストグラム

　MBTI別に各特徴量をヒストグラムにした結果を図3-3から図3-xまで示す。

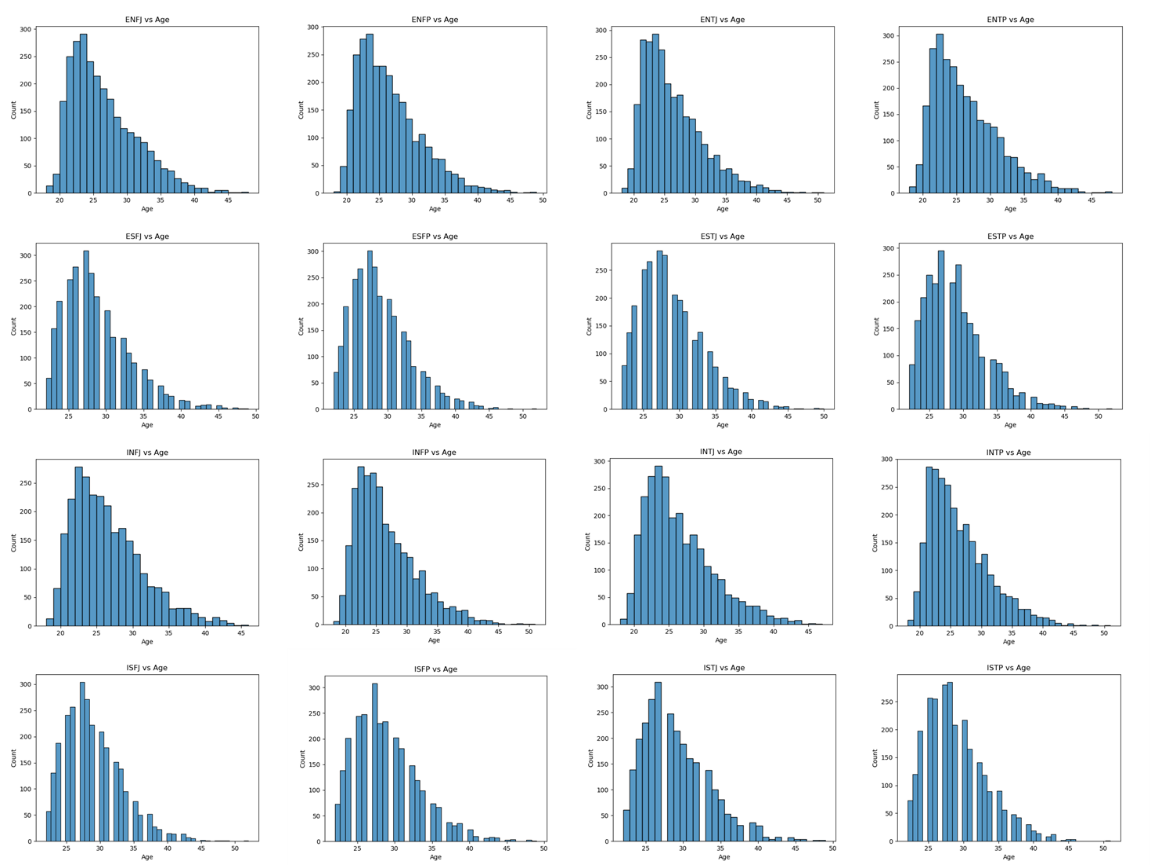


図3-3　MBTI別年齢

図3-3からMBTIにおいて年齢による差は少ないが、分布から若年層を中心としたデータセットであることが分かる。

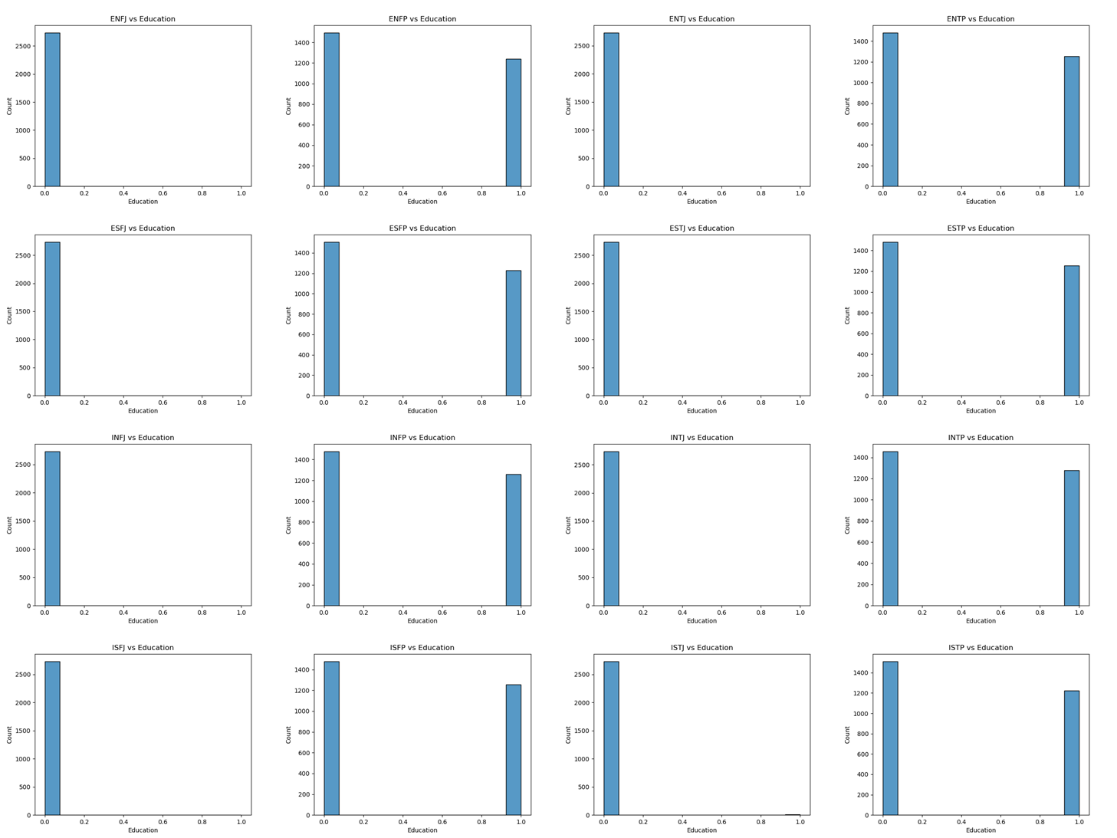


図3-4　MBTI別教育

本データセットにおいてJ(判断型)とP(知覚型)では知覚型のPは大学院進学あるいはそれ以上の結果になりJは大学進学までという結果になった。

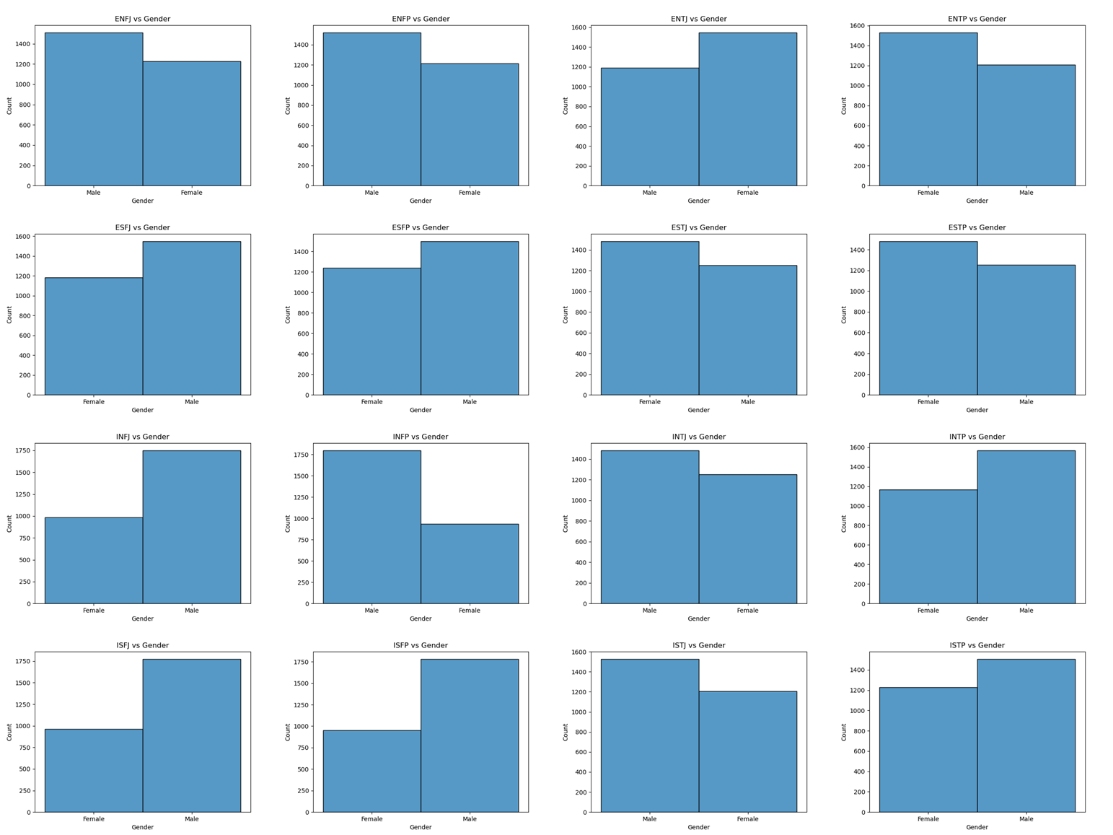


図3-5　MBTI別性別

E(外向的)の性別はS(感覚型)かN(直感的)かで男女比が逆転しておりTJ(論理的思考)の場合、他のSかNと比率が逆転している。また、Nの時は男性が多く、Sの時は女性が多い。

また、I(内向的)の性別はINFJとISFJで同じ性別の多さは同じ程度の比率だがEと比べて男女比が大きく異なる。同様に、INFPとISFPは性別が逆転しているがEと比べて男女比が大きく異なる。

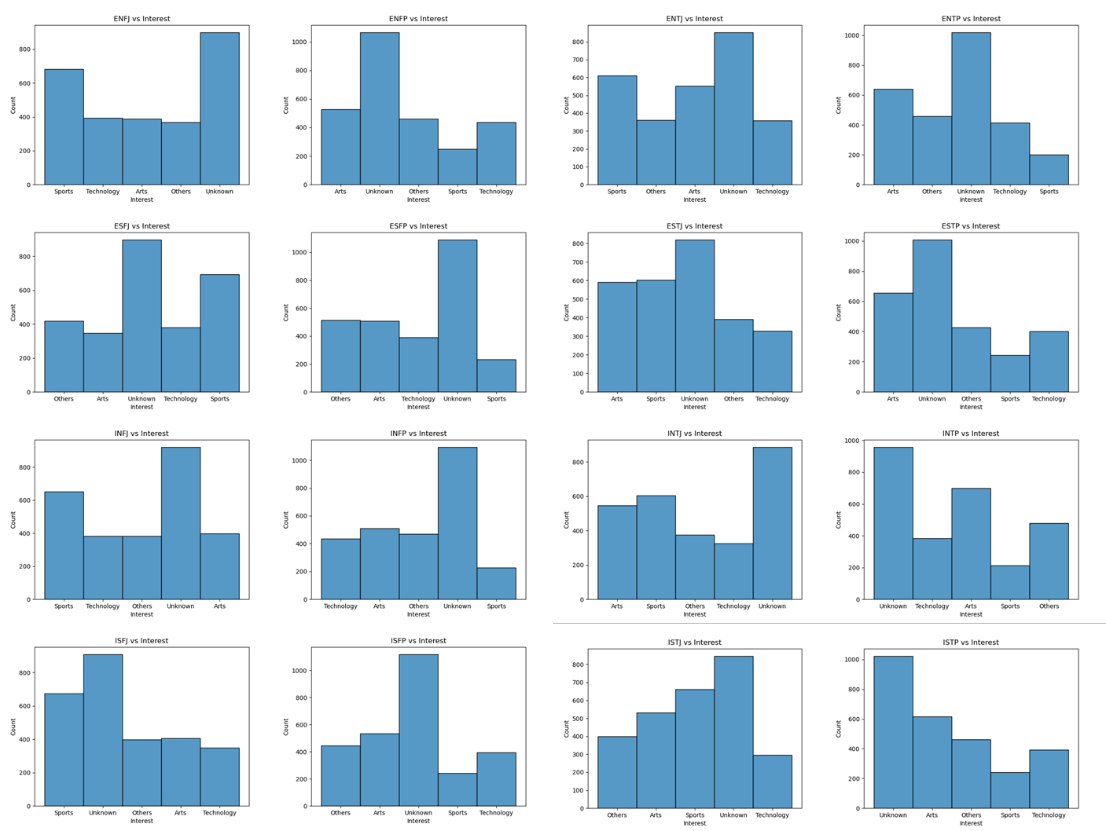


図3-6　MBTI別興味

図3-6より目立った法則性自体は確認できなかった。ただし、MBTI別に何に興味があるかは分かれていることが分かった。

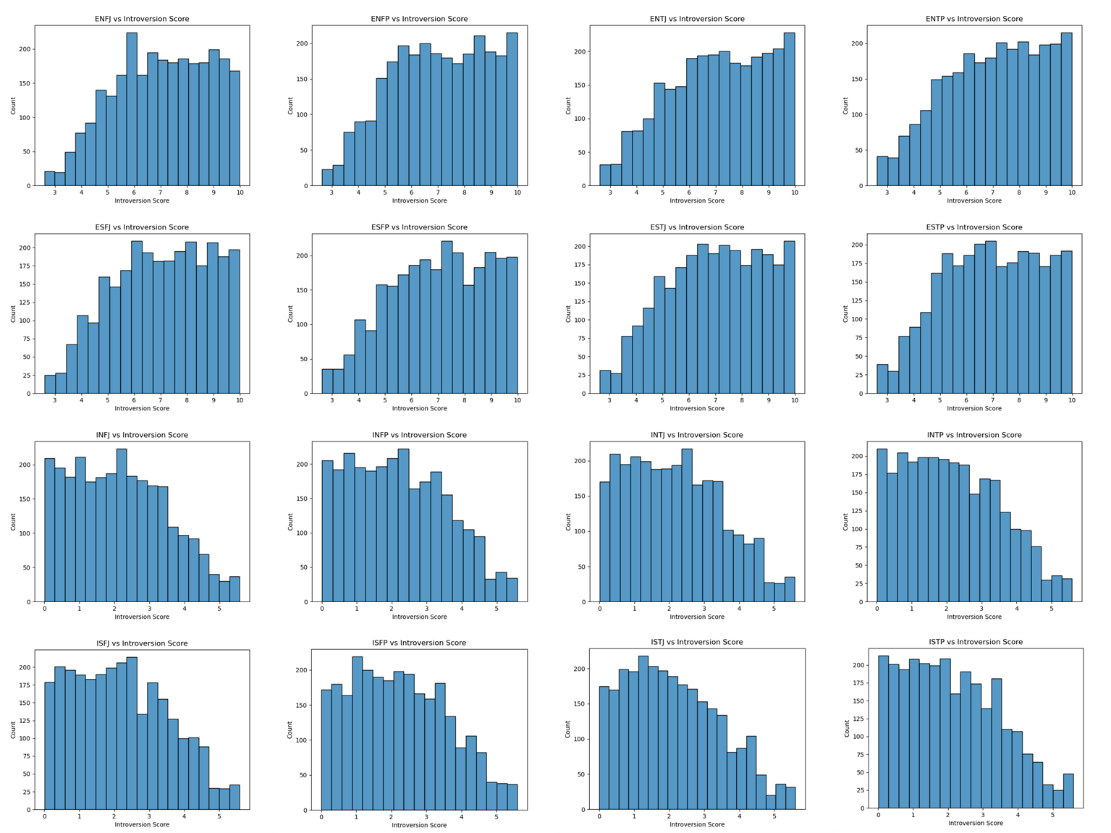


図3-7　MBTI別内向スコア

図3-7から基本的にE(外向的)は右肩上がりの結果だったのに対して、I(内向的)は右肩下がりの結果になった。

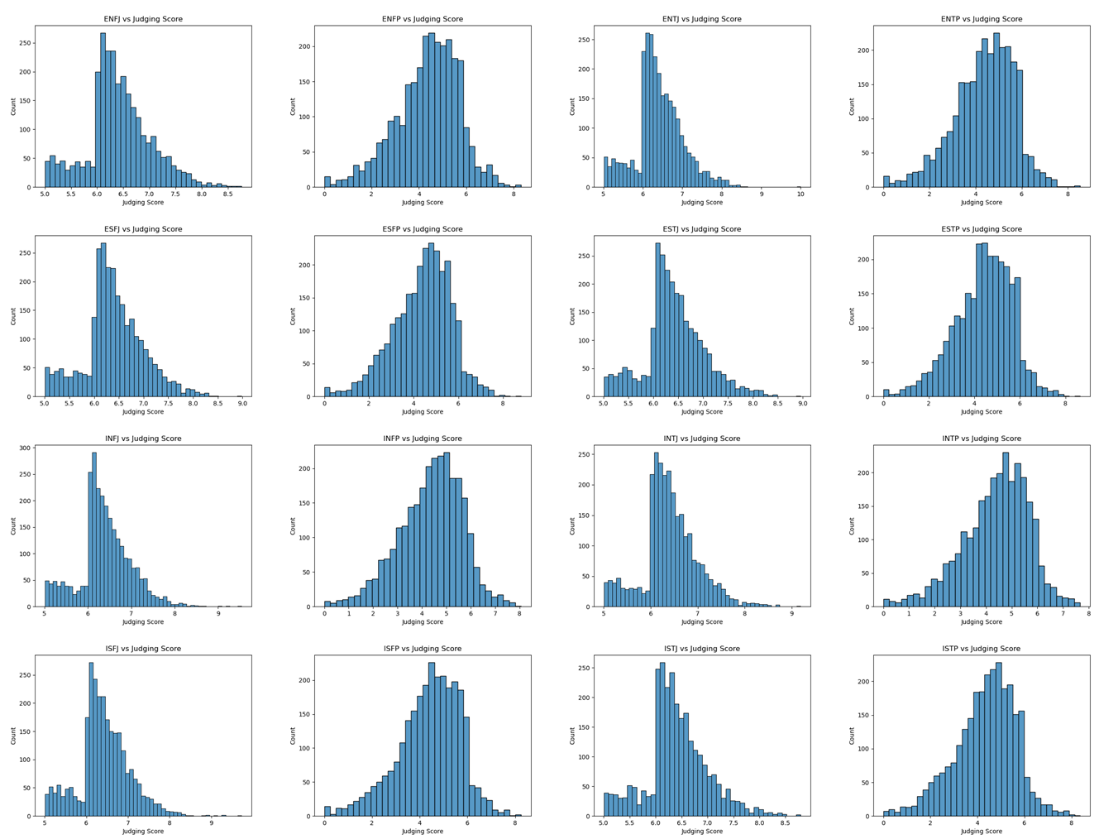


図3-8　MBTI別判断スコア

図3-8よりJ(判断型)はスコアが6辺りから急激に値が上昇しているおり、その後すぐ下降している。それに対してP(知覚型)はどれも近似した分布をしていることが分かる。

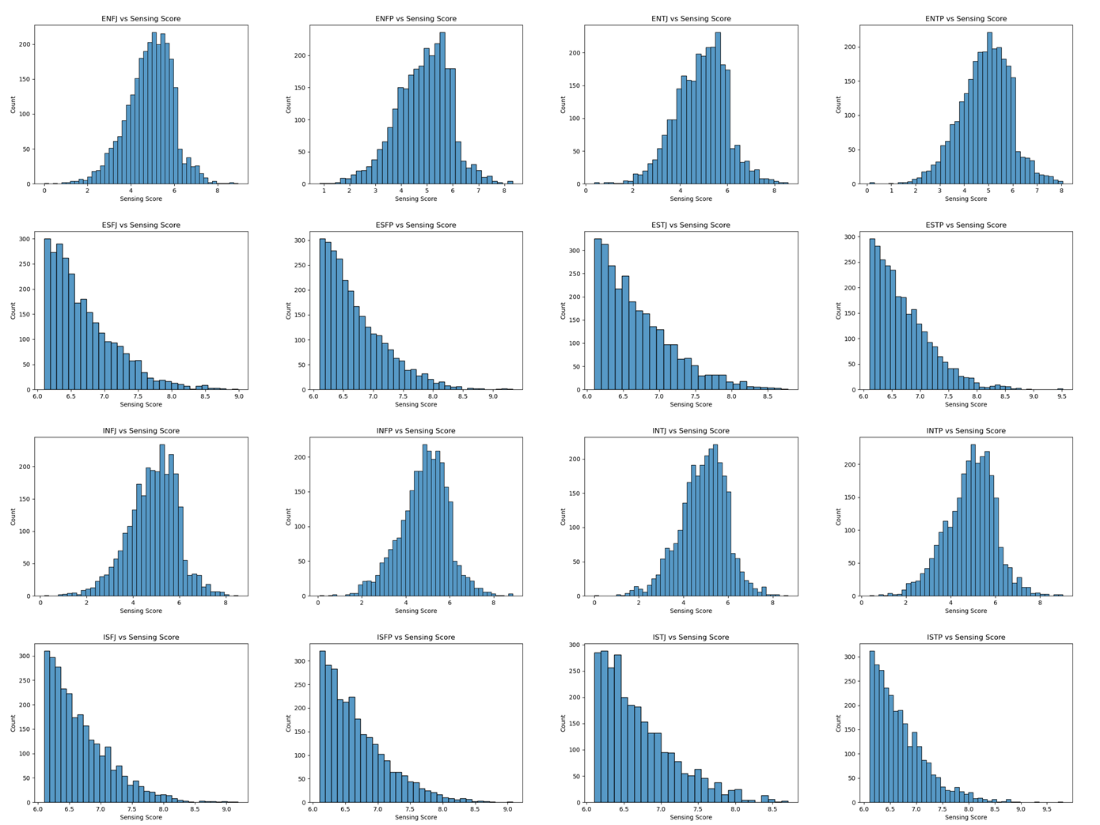


図3-9　MBTI別感覚スコア

図3-9からS(感覚型)はスコア6近くに最頻値がありそこから下降している。一方でN(直感型)は基本的に似た分布をしている。

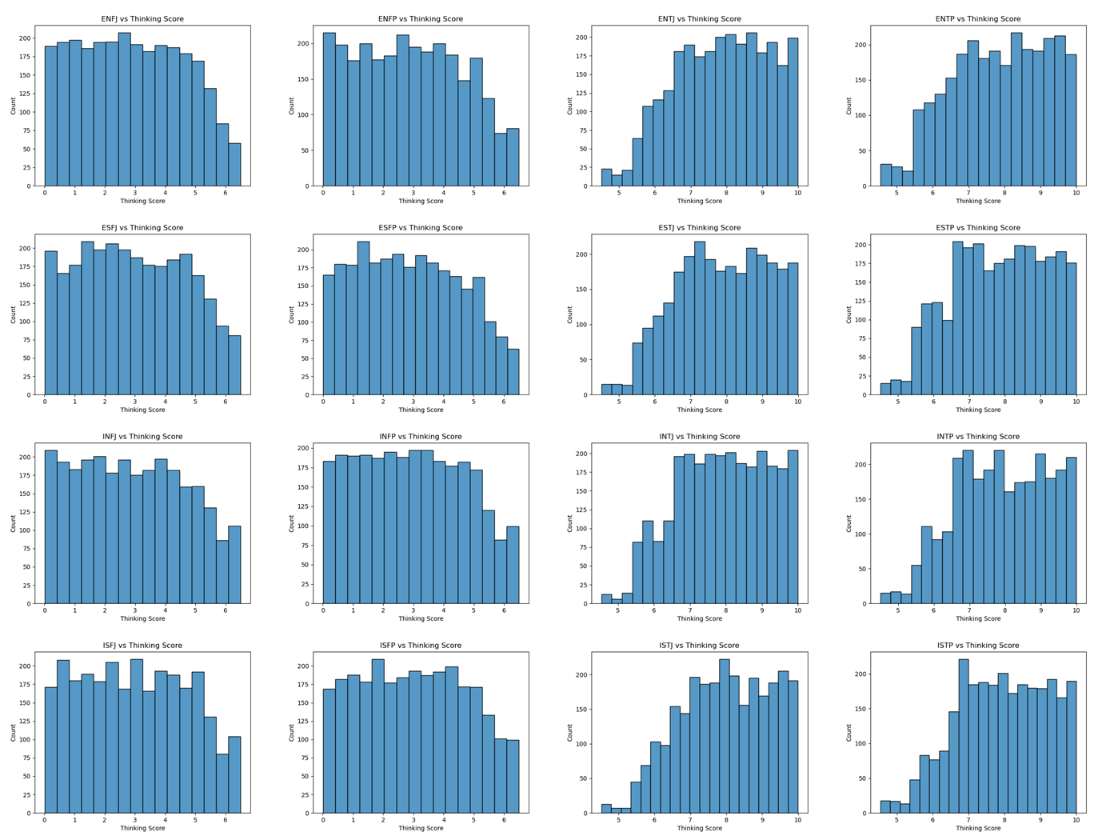


図3-10　MBTI別思考スコア

図3-10からT(思考型)はある地点からヒストグラムが右肩上がるがどこから上がるかがMBTIによって異なるのに対してF(感覚型)はある地点まで一様分布のような形をしてそこから右肩下がりのヒストグラムになっている。

1. クラスタについて検証

よりデータを深く分析するために主成分分析を用いて次元圧縮しデータを2次元に可視化した。この時に第一主成分と第二主成分の因子負荷量を分布とともにプロットするbiplotの結果を図4-1に示す。

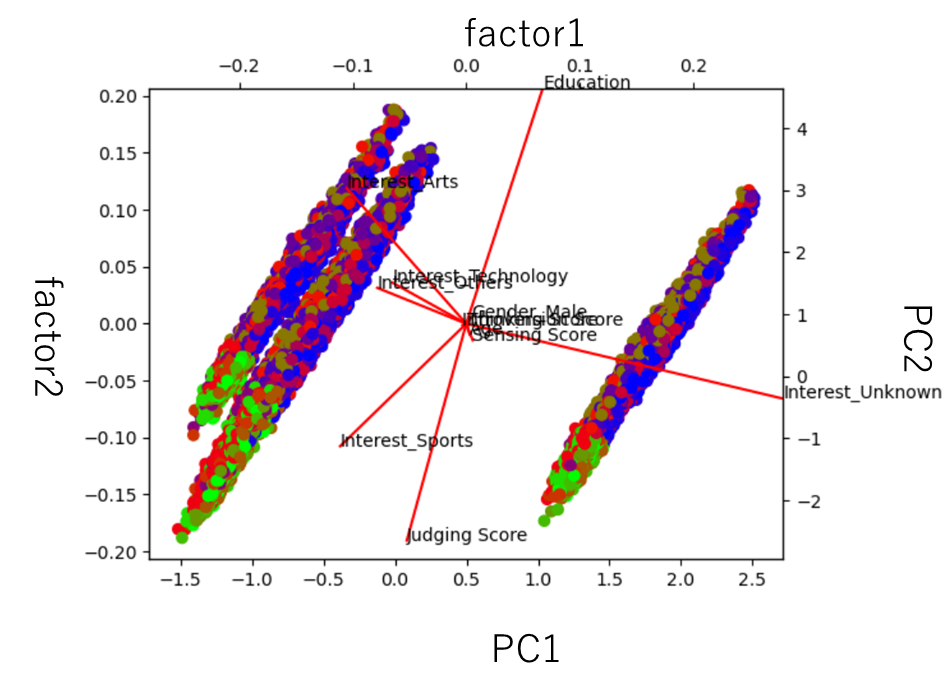


図4-1　データセットの主成分分析と因子負荷量

　なお主成分分析をするにあたり、尺度を統一するため標準化を行った。

　また図4-1より3つのクラスタが確認できた。しかしMBTIについてはクラスタによって規則性(MBTIの値は色)が確認できないことが分かった。

　biplotにおける因子負荷量の出力結果から横軸(第一主成分)は値が高いとInterest\_UnknownがTrue(ダミー変数のため)になりInterest\_ArtsがFalseになる。また、因子負荷量から縦軸(第二主成分)は値が高いと教育の値が1になり、判断スコアが下がりInterest\_SportsがFalseになる。

* 1. クラスタリング

　この3つのクラスタについてクラスタリングを行いラベル付けを行う。

　結果として行った主成分分析における第一主成分から第三主成分までを特徴量とし混合ガウスモデルを使用することで適切なクラスタリングができた(図4-2)。

グラフ

AI によって生成されたコンテンツは間違っている可能性があります。

図4-2　第一から第三主成分でクラスタリングをしてラベル付けを行った結果

* 1. 第三主成分までの因子負荷量

　図4-2のクラスタリングは第三主成分まで使用している。そこで第三主成分までを可視化したグラフを図4-3に示す。

グラフ, 折れ線グラフ, ウォーターフォール図

AI によって生成されたコンテンツは間違っている可能性があります。

図4-3　第三主成分までの因子負荷量

ここで、目立った第三主成分の因子負荷量については興味(芸術)が負に高く興味(他)と興味(科学技術)が正に高いことが分かる。

　つまり第三主成分では値が正に大きいほど興味(他)・興味(科学技術)がTrueになり興味(芸術)がFalseになることが分かる。

* 1. クラスタリング結果を教師データにして分類

　4.1項で行ったクラスタリングによりラベル付けがされた。そこで元のデータセットを用いてクラスタの分類に寄与する特徴量を探す前段階としてクラスタリングされてできたラベルを教師データとして分類し精度を測定して分類寄与率の信ぴょう性を確かめる。

　ここで訓練データ50%でテストデータ50%において得られた精度を図4-4に示す。

テーブル

AI によって生成されたコンテンツは間違っている可能性があります。

図4-4　クラスタリング結果を教師データとしたときの精度

　図4-4より精度は100%の結果になった。この結果としての信ぴょう性を元に計測した分類寄与率を表4-1に記す。

表4-1　クラスタリング結果で測定した分類寄与率



測定した結果から上位から「年齢」「内向スコア」「興味(芸術)」「興味(不明)」「感覚スコア」「思考スコア」「判断スコア」「教育」という結果になった。

　ここで上位になった特徴量のヒスとグラムを図4-5から図13まで示す。

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図4-5　年齢のヒストグラム

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図4-6　内向スコアのヒストグラム

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図4-7　興味(芸術)のヒストグラム

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図4-8　興味(不明)のヒストグラム

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図4-9　感覚スコアのヒストグラム

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図4-10　思考スコアのヒストグラム

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図4-11　判断スコアのヒストグラム

グラフ, ヒストグラム

AI によって生成されたコンテンツは間違っている可能性があります。

図4-12　教育のヒストグラム

1. 考察
   1. MBTIデータセットについて

図3-3から図3-10において、MBTIでは陽性陰性の基準が明確にあり、それがデータセットに現れたことが表3-2における分類寄与率に現れたことが考えられる。

また、教育についてはPが大学院まで行きJが大学院まで行かない傾向があることがわかる。

E(外向的)の性別はS(感覚型)かN(直感的)かで男女比が逆転しておりTJ(論理的思考)の場合、他のSかNと比率が逆転している。また、Nの時は男性が多く、Sの時は女性が多い。また、I(内向的)の性別はINFJとISFJで同じ性別の多さは同じ程度の比率だがEと比べて男女比が大きく異なる。同様に、INFPとISFPは性別が逆転しているがEと比べて男女比が大きく異なる。

* 1. クラスタリング

　図4-1のbiplotからクラスタ2つと1つを分ける特徴量は興味であることが分かる(特に興味(不明)、興味(芸術))。

　図4-2からクラスタリングできた事の証左として図5-1にクラスタリングラベルごとの分布を示す。

グラフ

AI によって生成されたコンテンツは間違っている可能性があります。グラフ

AI によって生成されたコンテンツは間違っている可能性があります。グラフ, 散布図

AI によって生成されたコンテンツは間違っている可能性があります。

図5-1　各ラベル毎のクラスタリング結果(左からラベル「0」「1」「2」)

　この結果からクラスタリングは正常に行えた事が考えられる。

図4-3から各クラスタを特徴づけた因子としては「教育」「判断スコア」「興味(芸術)」「趣味(他)」「興味(スポーツ)」「興味(科学技術)」「興味(不明)」等が挙がっている。また、分類寄与率とヒストグラムから興味軸は特に「興味(芸術)」「興味(不明)」がクラスタを分ける要因になっていることが分かる。その他では判断スコアと教育が若干分布が異なっており、明確に分かれるのは「興味(芸術)」「興味(不明)」の2つであることからこのクラスタは「興味(芸術)」「興味(不明)」を分けるものとなっており判断スコアと教育は特徴量を補完していると考えられる。

1. まとめ

　MBTIの93個の質問が完璧だと仮定したとき、今回の8個の特徴量では代表値の中で最頻値を採用するとしたときに10%程度精度が落ちてしまうことが分かった。

　また、MBTIとは関係ない事項(年齢・性別)においては性別は外向的(E)の時に関係しやすいことが分かったが内向的(I)の時は性別において規則性は確認できず、年齢については細かい分布の違いはあれど目視では分類に寄与するほどの大きな違いが確認できなかった。

　また、趣味趣向については分類においては寄与が確認できなかったがMBTI毎に異なっており、これについて法則性は確認できなかった。

　次に主成分分析をした結果3つのクラスタが現れ、第三主成分まで特徴量としてクラスタリングに有効であることが分かった。

　これについて、因子負荷量とGMMで作られたクラス別ヒストグラムからは3つのクラスタは興味についてがメインで判断スコアと教育がデータを補完していることが分かった。