MondrianForestのテストを行った。

<http://archive.ics.uci.edu/ml/index.php>というなんか色々データがあるサイトから落としてきた、以前学習用に遊んでいたadultというデータ(エクセルで修正)を使ってテストをしてみる。

データの内容はAge, Workclass, Fnlwgt, Education, Education\_number, Marital\_Status, Occupation, Relationship, Race, Sex, Capital\_Gain, Capital\_Loss, Hours\_pre\_week, Native\_Countryの14項目から、Imcomeが>50Kか、<=50Kかを判別するというもの。

データ数は32561個で、テストデータは答え付きで16281個ある。

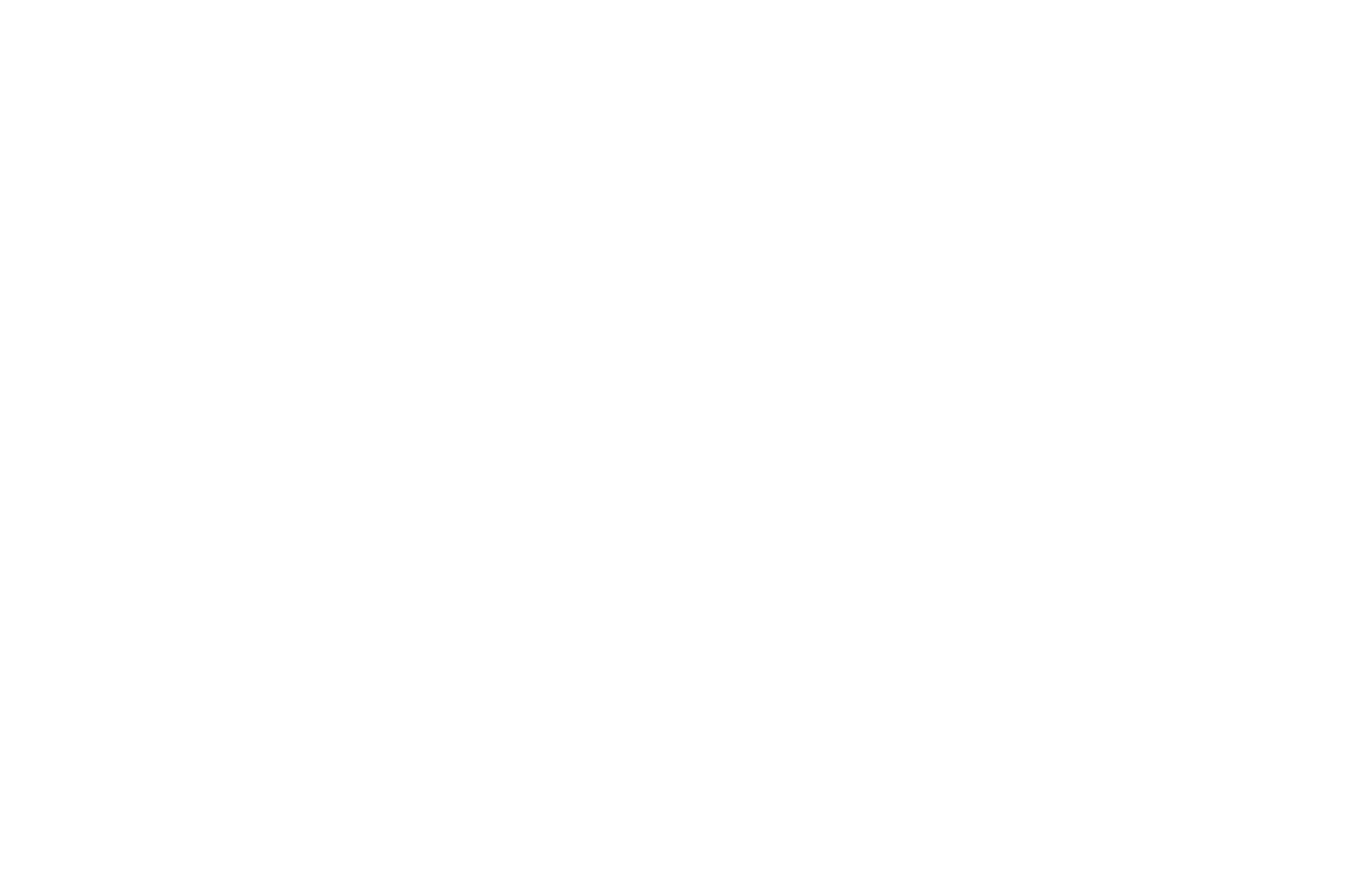
(計算時間なのでデータ数が多い方がいいと思ってこうしたのですが、処理が重すぎて死にました。勉強になりました。)

欠損値などあったので、seminarのP04S06\_Scoring process-2に従ってget\_dummies, RFEによる特徴量選択などのデータの前処理をした。

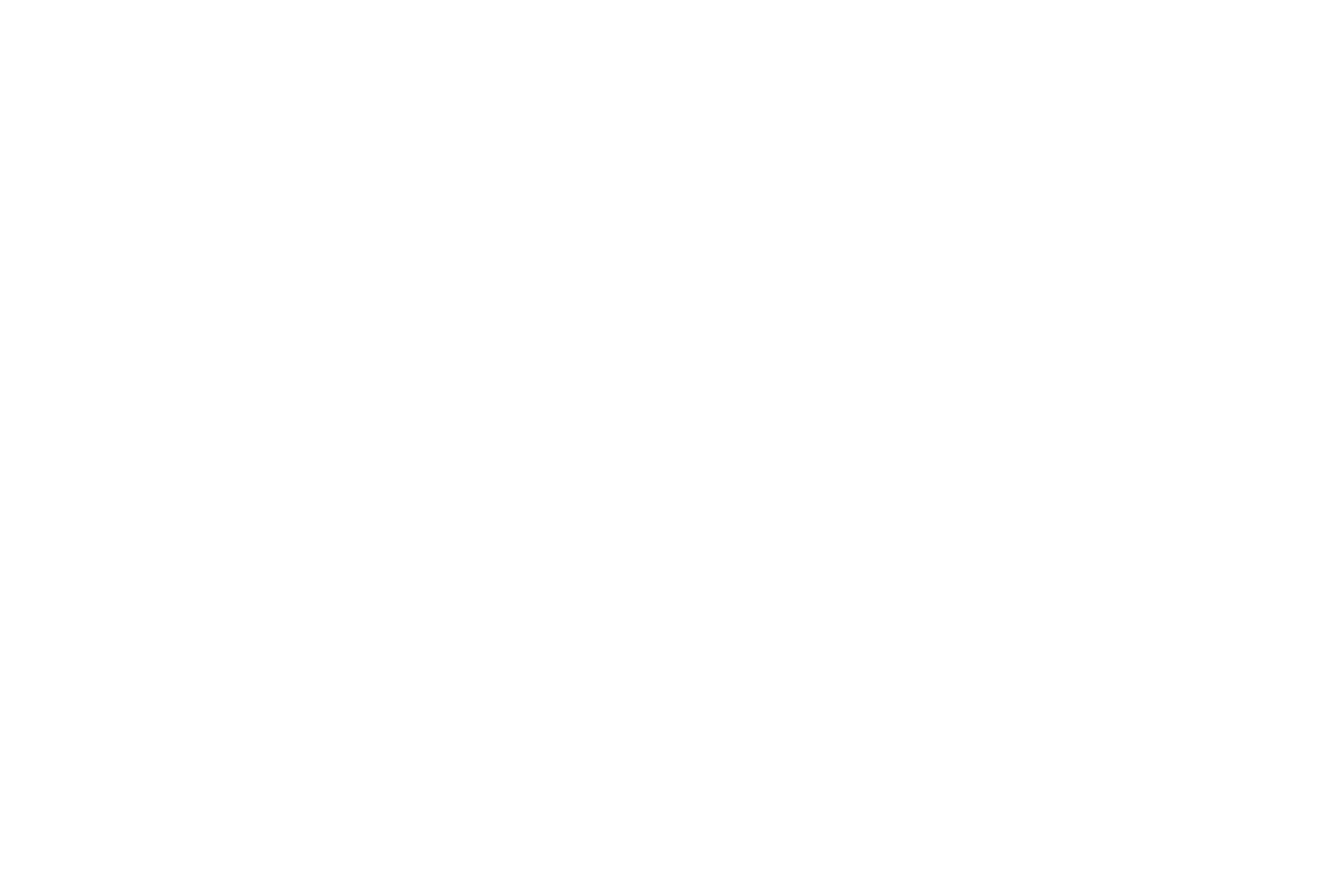
特徴量は15個に絞り、残ったのは、Age, Fnlwgt, Education\_number, Capital\_Gain, Capital\_loss, Hours\_pre\_week, Workclass\_Self-emp-not-inc,…,など。

Partial\_fitはpipelineできないため、事前にStanderdSchlerによる標準化を行い、「fit」など自体の処理を計測した。

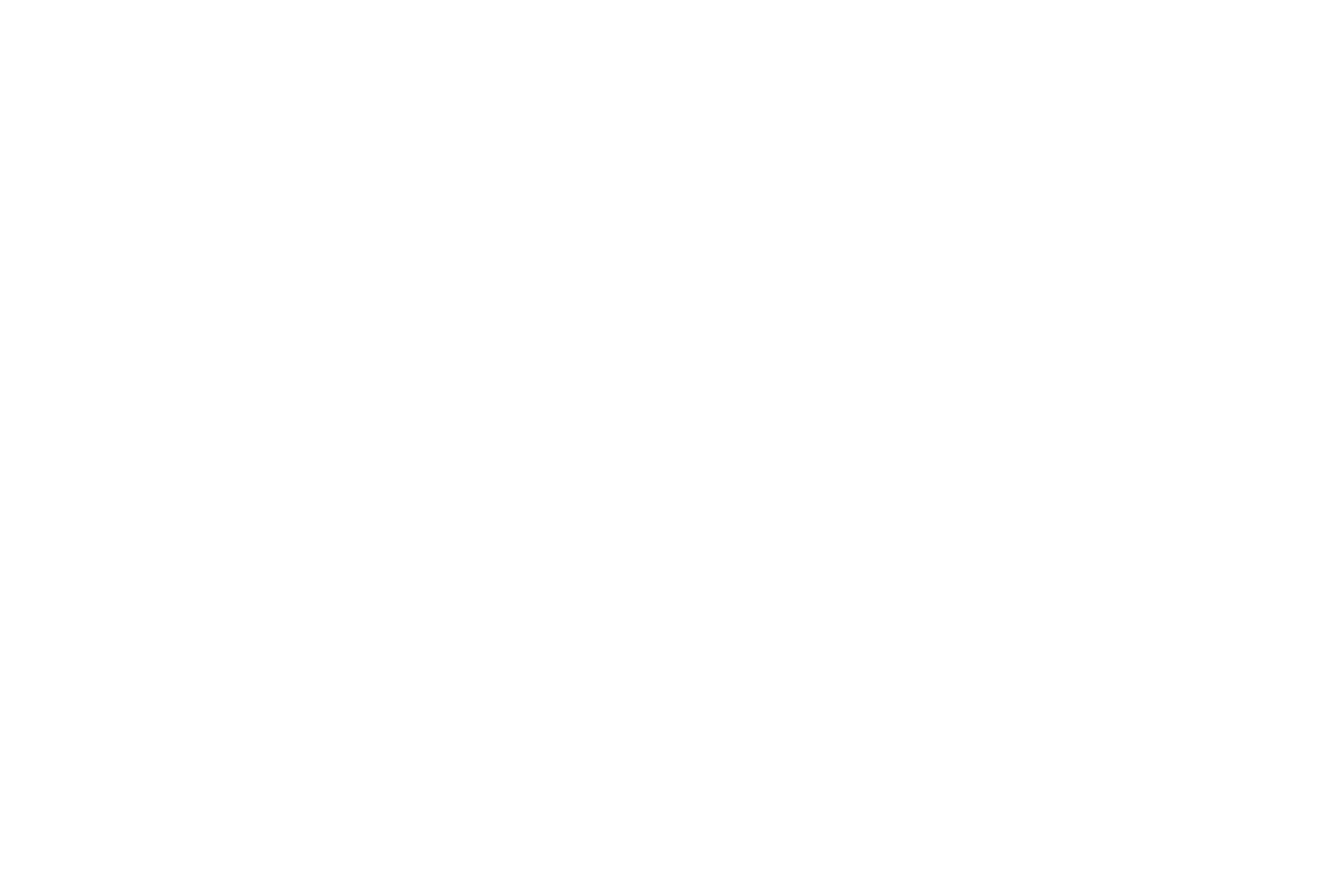
まず、そもそもの性能を見るために、一度データ数に伴って各学習器の計算時間とスコアがどうなるかを調べる。ここでは、MondrianForestのPartial\_fitは毎回リセットし処理を行っているので、蓄積されたデータ数がゼロの時から、データを入れるとどれくらい時間がかかるかを示している。



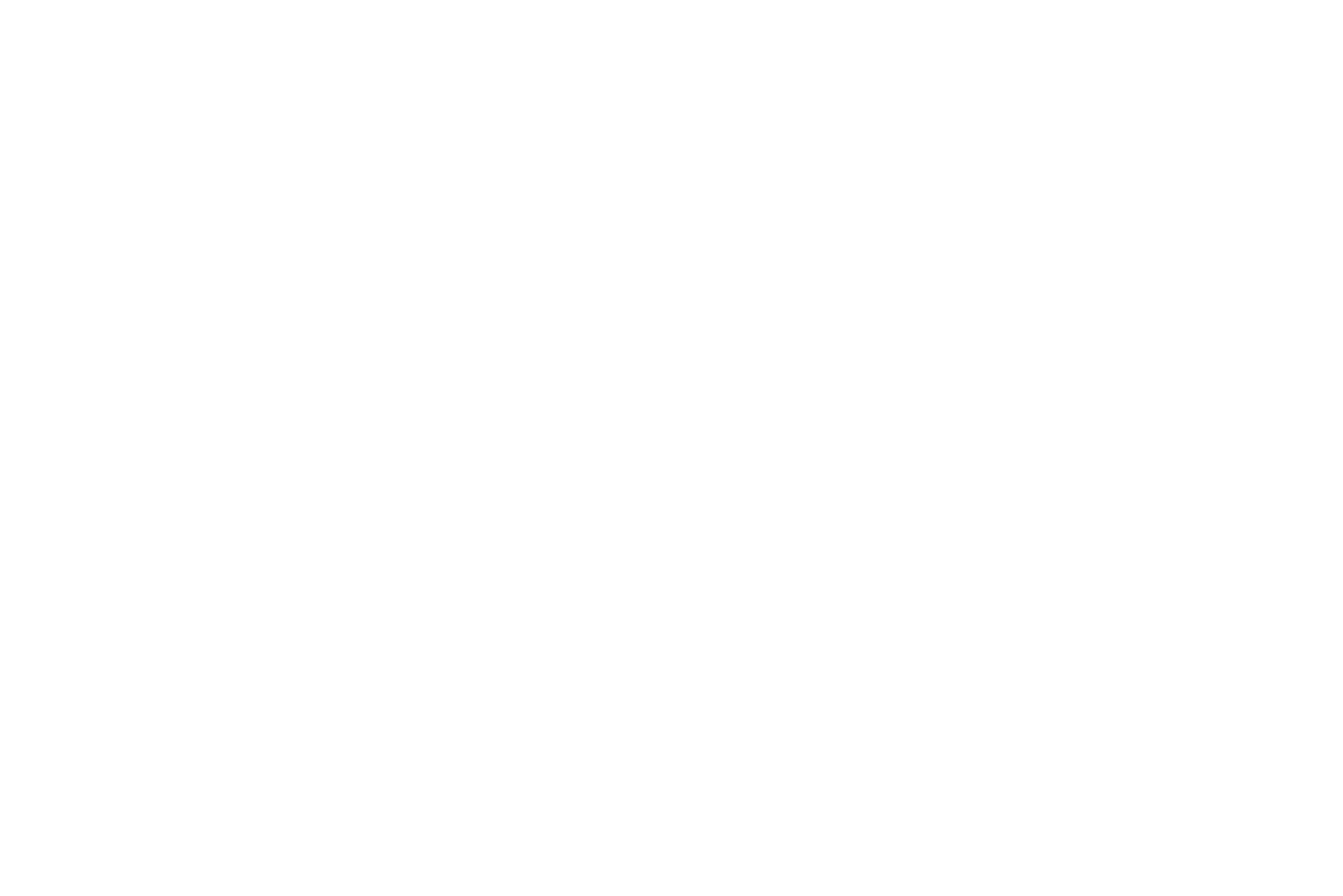
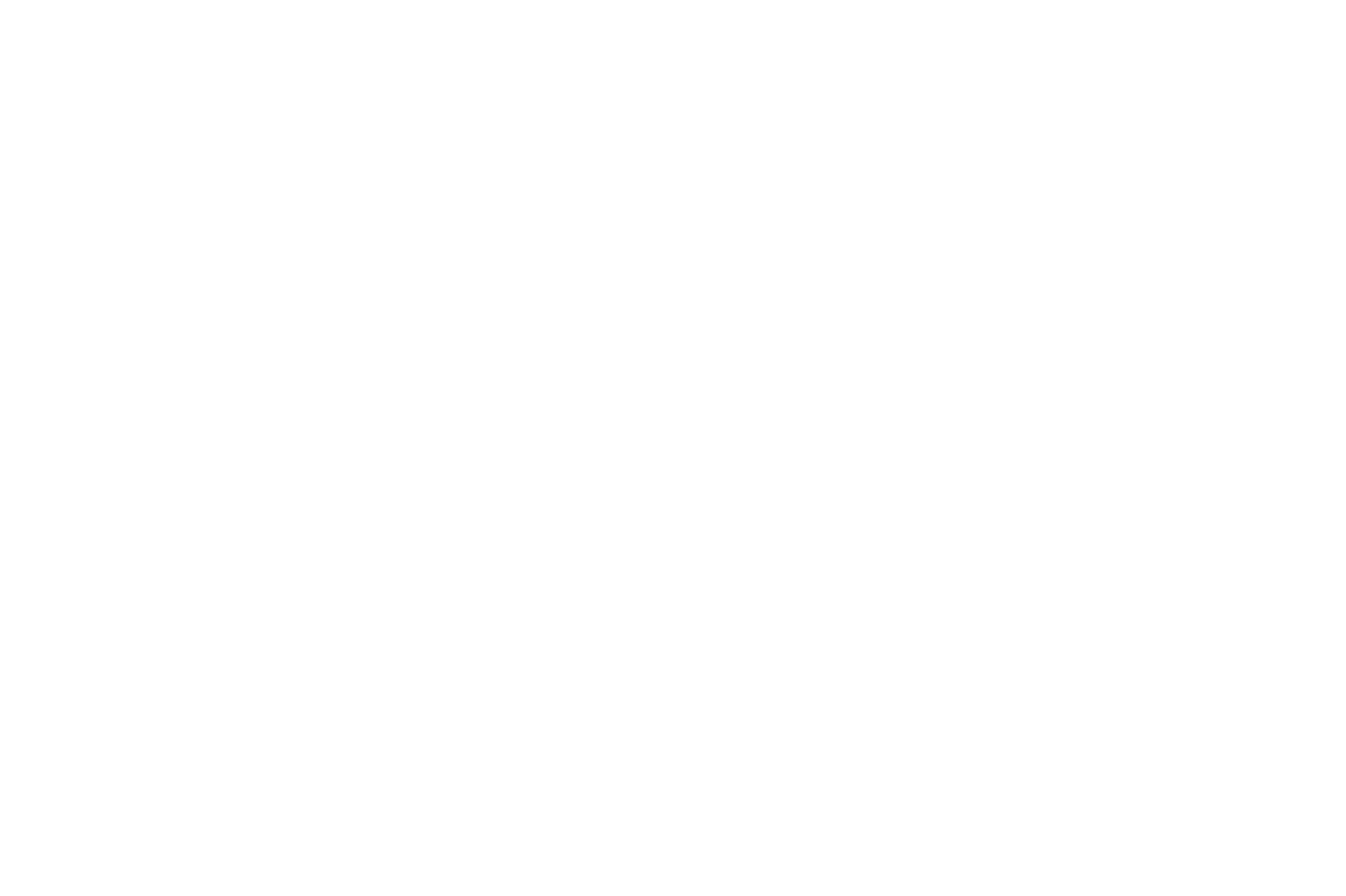
この時のaccuracy score。



また、f1 score



最初の方でほぼ収束してしまっているので、軸を変えて。



とりあえず、ここから読み取れることは、

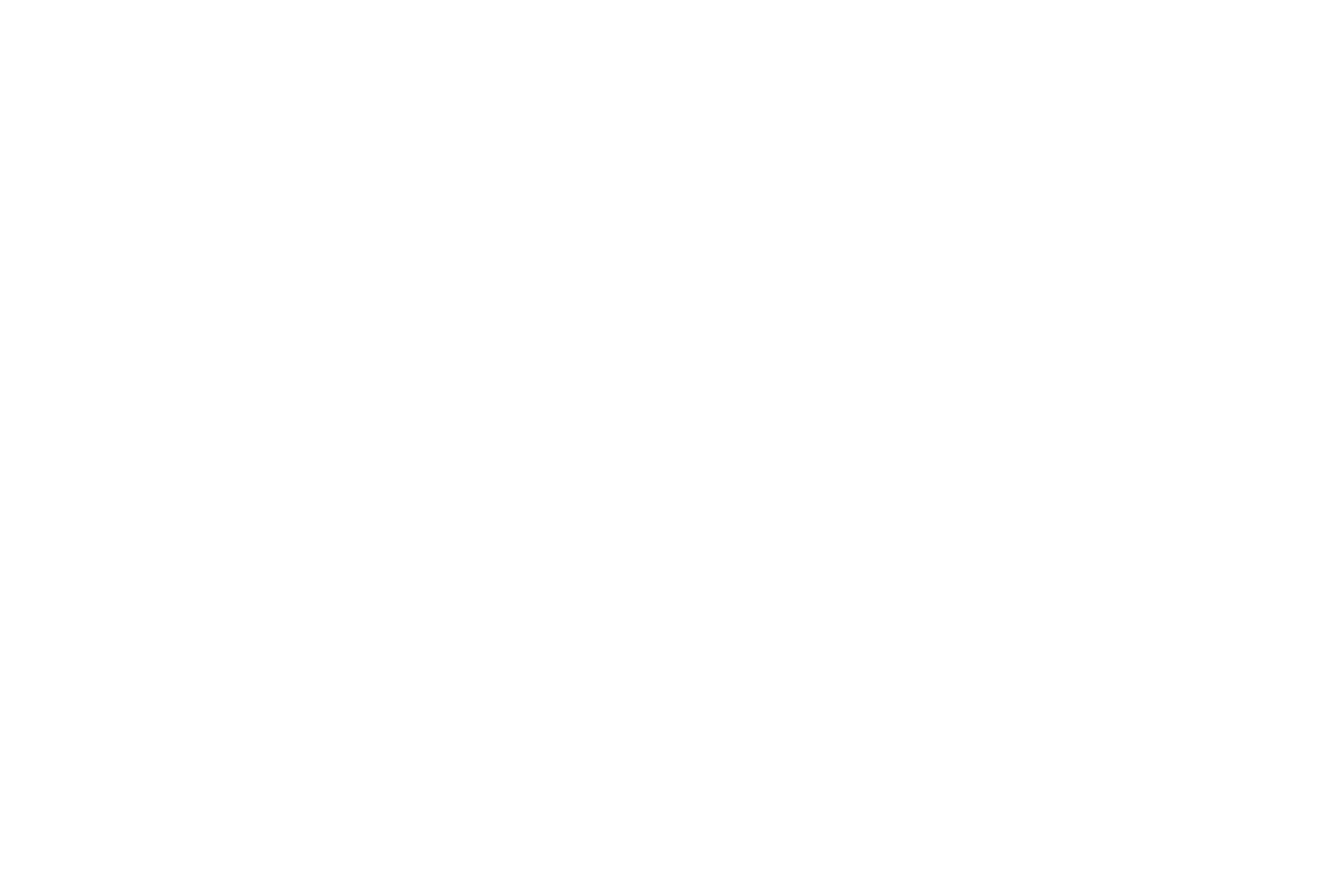
・MondrianのPartial\_fitとfitはほぼ精度が変わらない(fitの方が若干高い)上に、partial\_fitの方がだいぶ遅い。

・MondrianForestはRandomForestと比べるとそこまで精度に差はないが、GradientBoostingと比べると有意な差が出てしまっている。

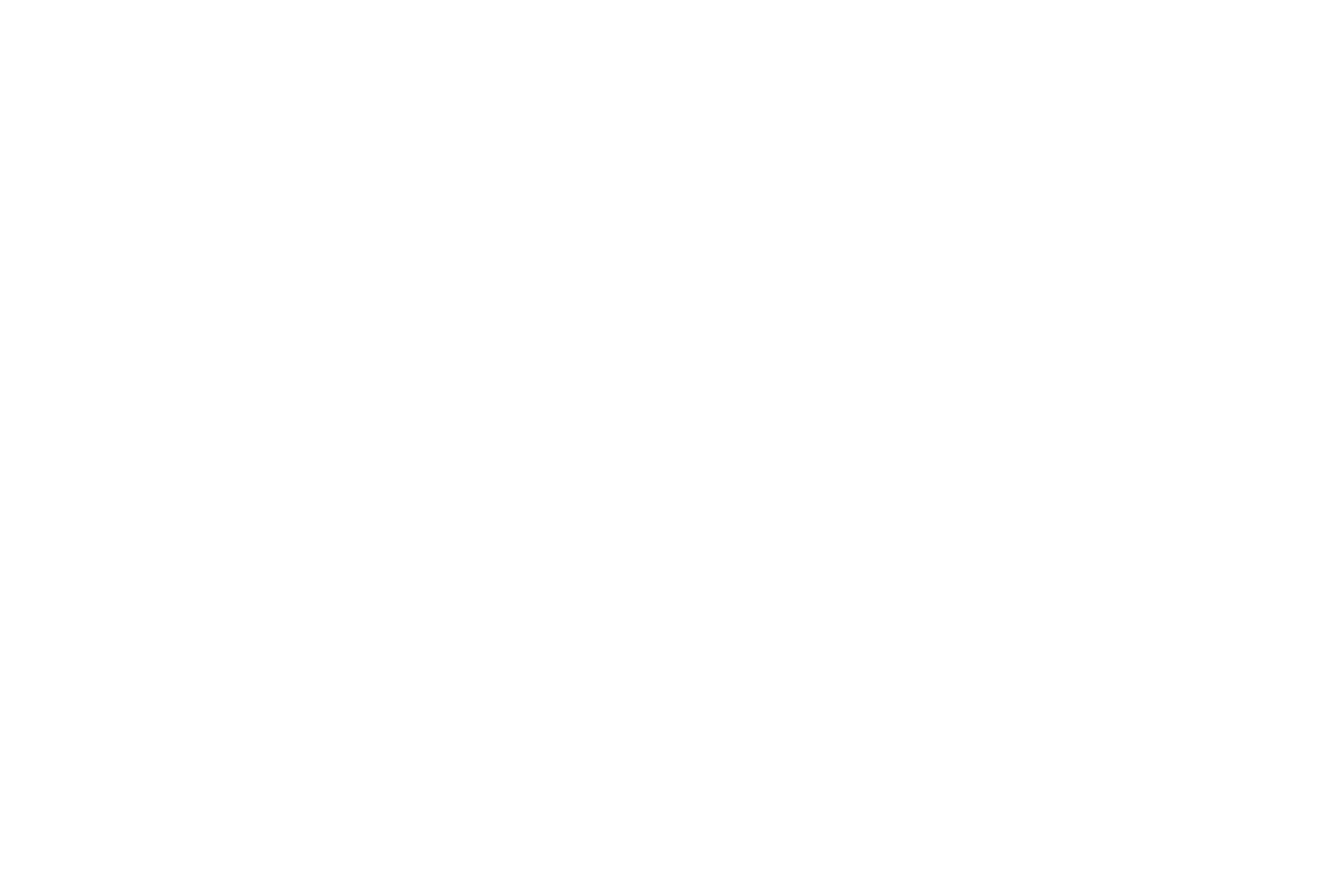
GradientBoostingはα(学習率)など全てデフォルト値でやっているので、スピードより精度重視の設定になってはいるが。

次に、データが1000個づつ次々に到着して来たという状況を仮定する。

その場合の合計の計算時間を調べる。

ここでは、各学習器は(そこまでひどい状況はないと思うが)データが1000個追加するたびに再学習する。Mondrianのpartial\_fitは1000個来るたびに、追加学習する。

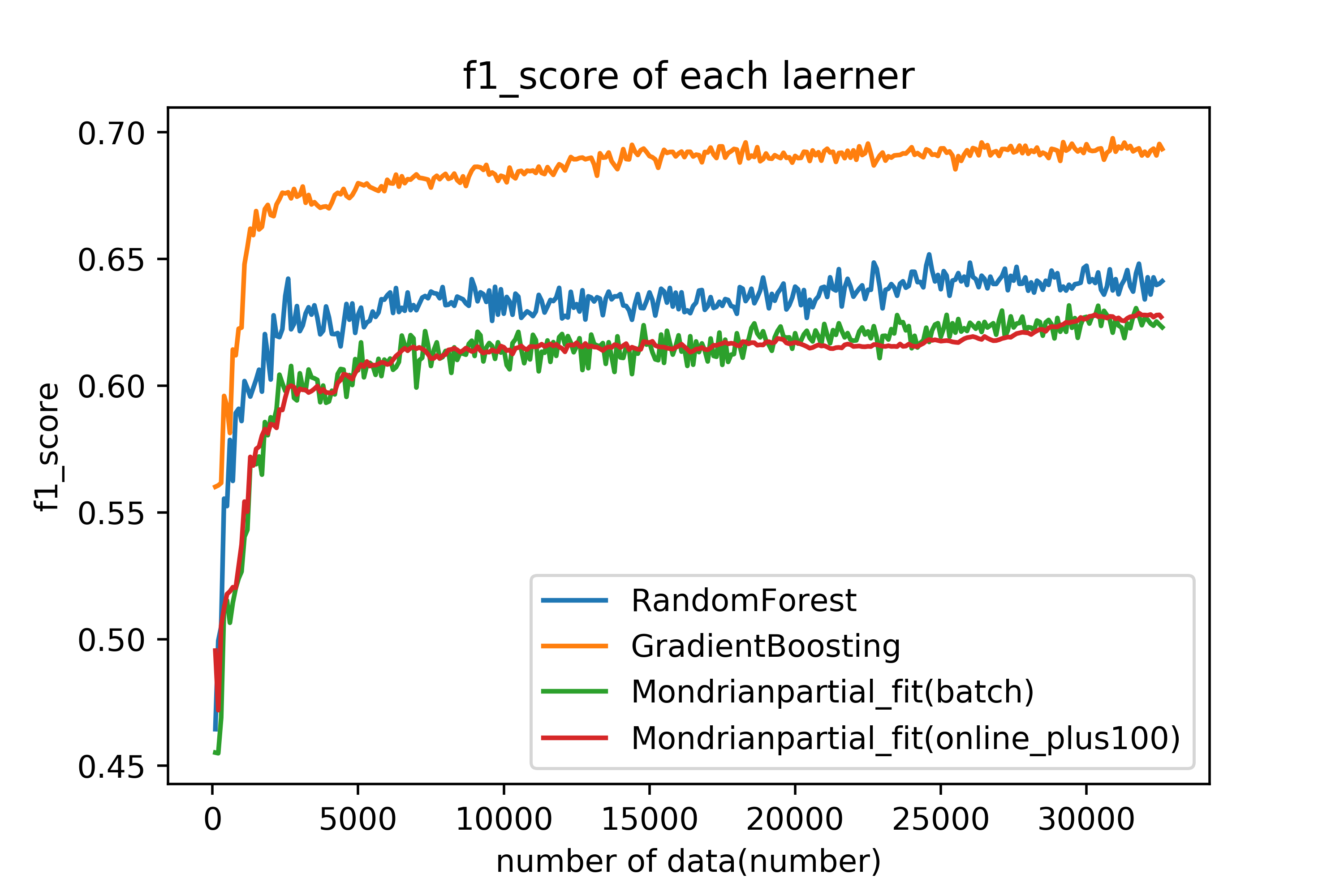
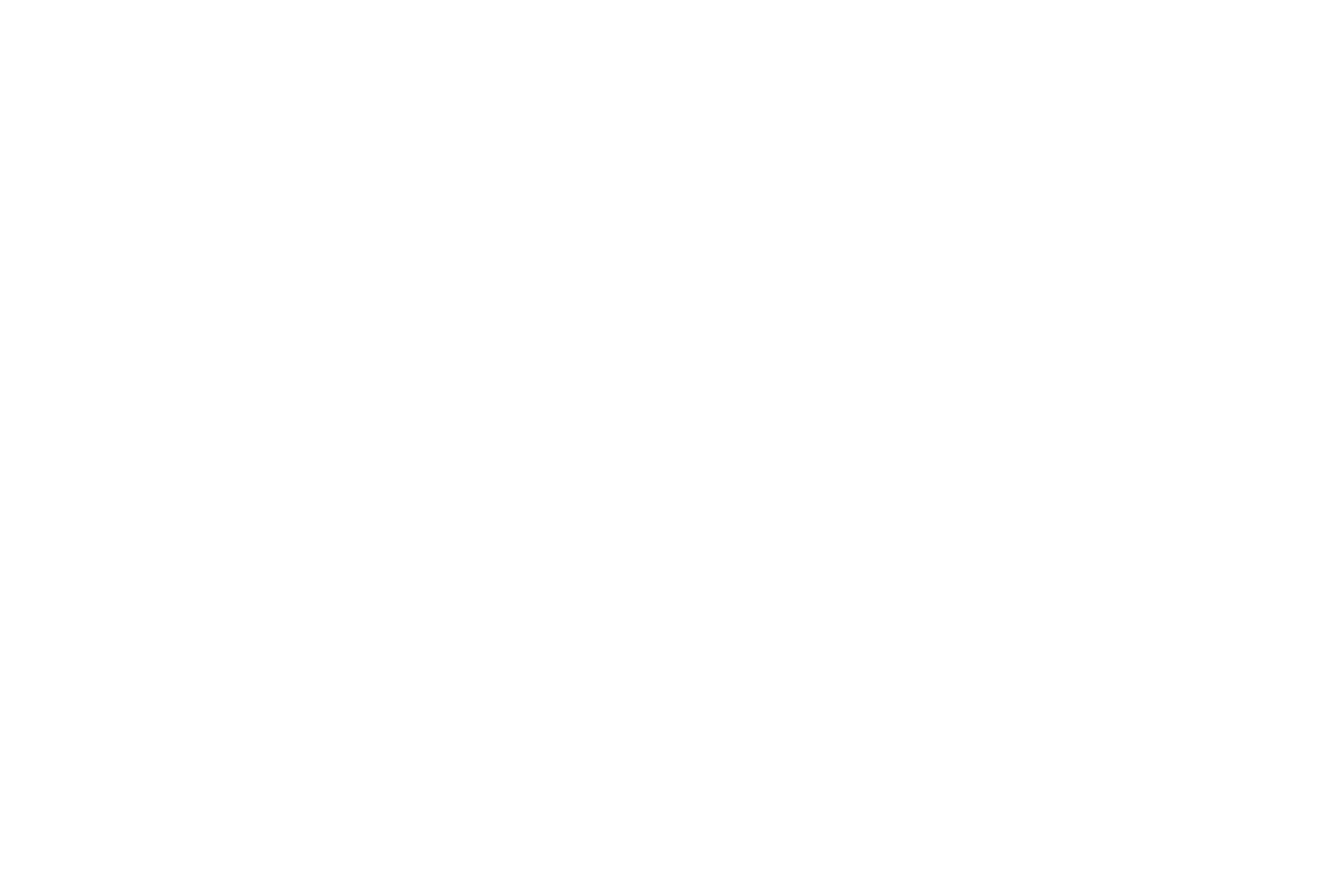
一応データが100個づつ来た場合。



さすがにMondrianが最も、そして大きく計算時間が早かった。

気になるのはこの時、精度が出ているかということ。

以下で、Mondrianpartial\_fit(batch)は、最初に性能を見た時のようにデータ数がゼロから学習させたもの、Mondrianpartial\_fit(online\_plus100)は100個ずつデータを追加学習させたものである。

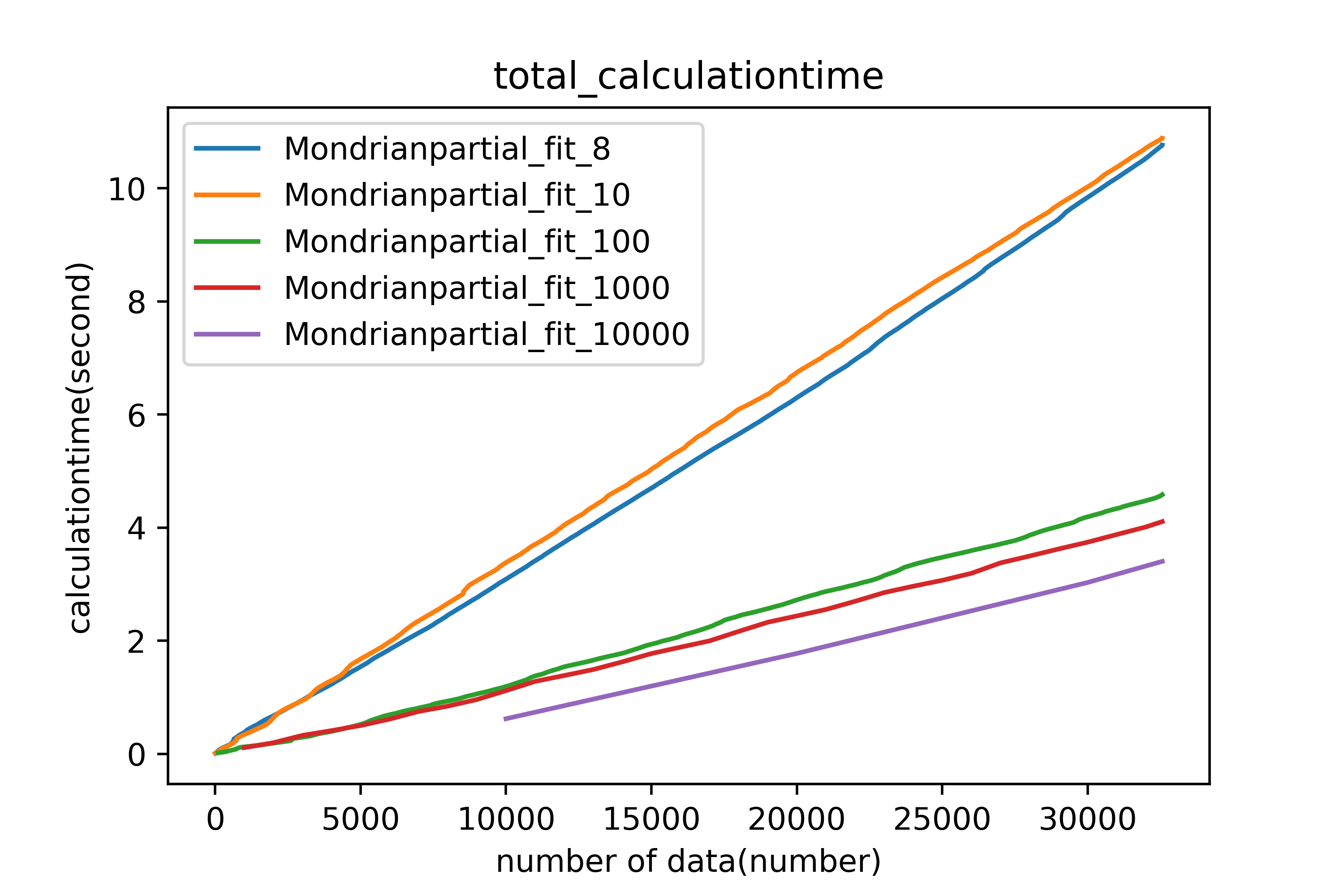


MondrianForestは一気に学習させた場合と追加学習させた場合でほぼ同じスコアであった。

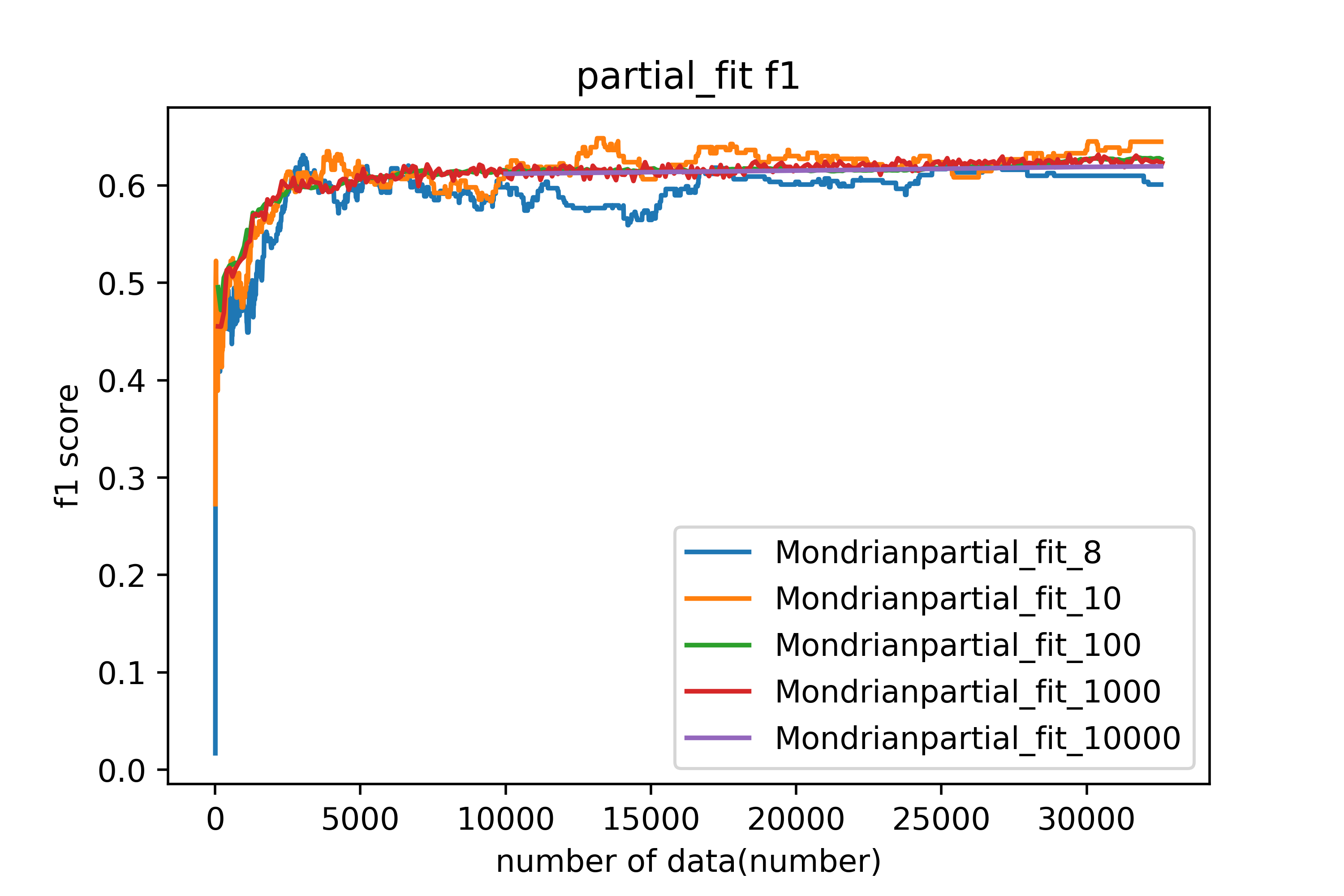
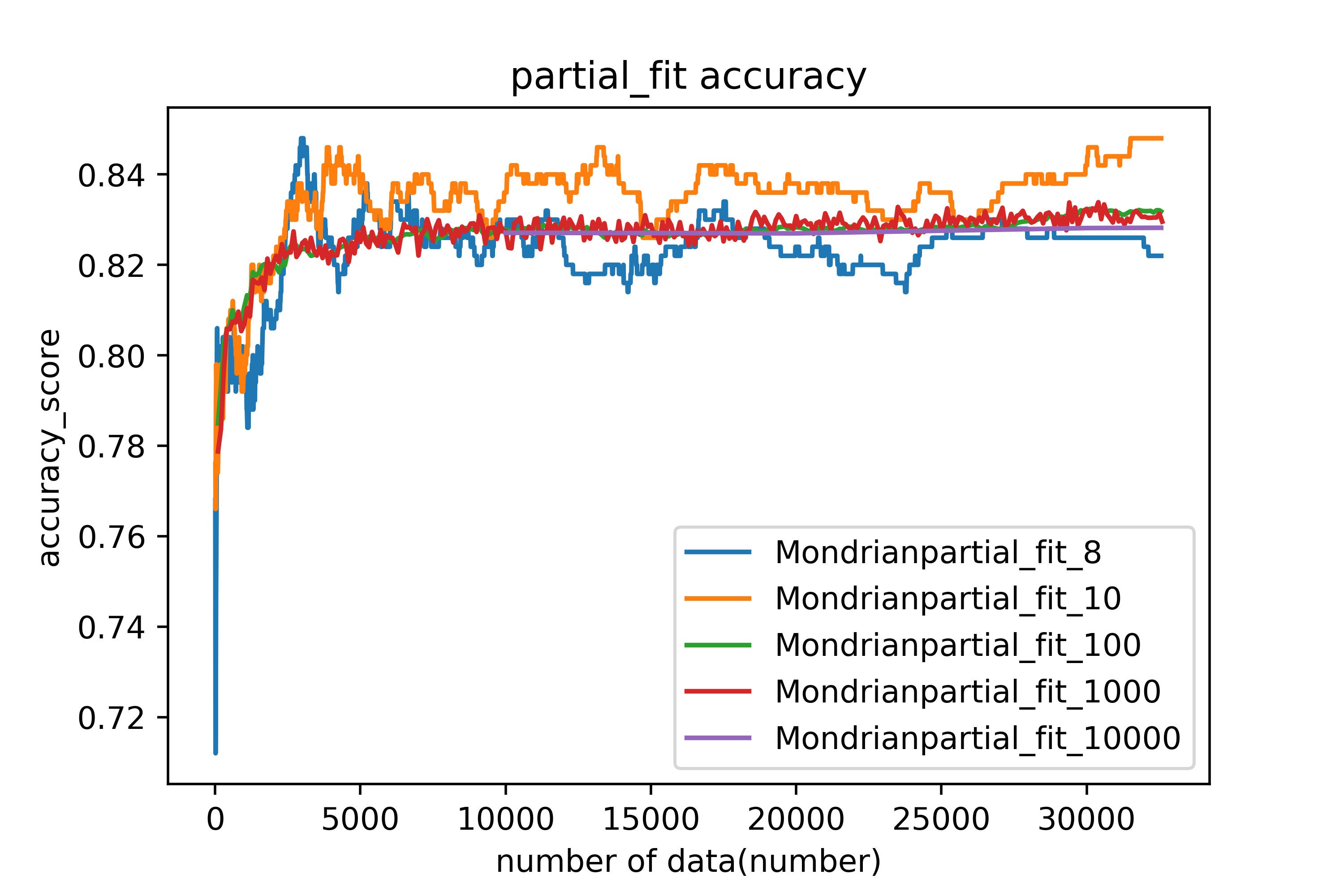
ちゃんとonline学習ができていると思われる。

最後に、8個ずつ、10個ずつ、100個ずつ、1000個ずつ、10000個ずつ足していった時の合計の計算時間を調べる。

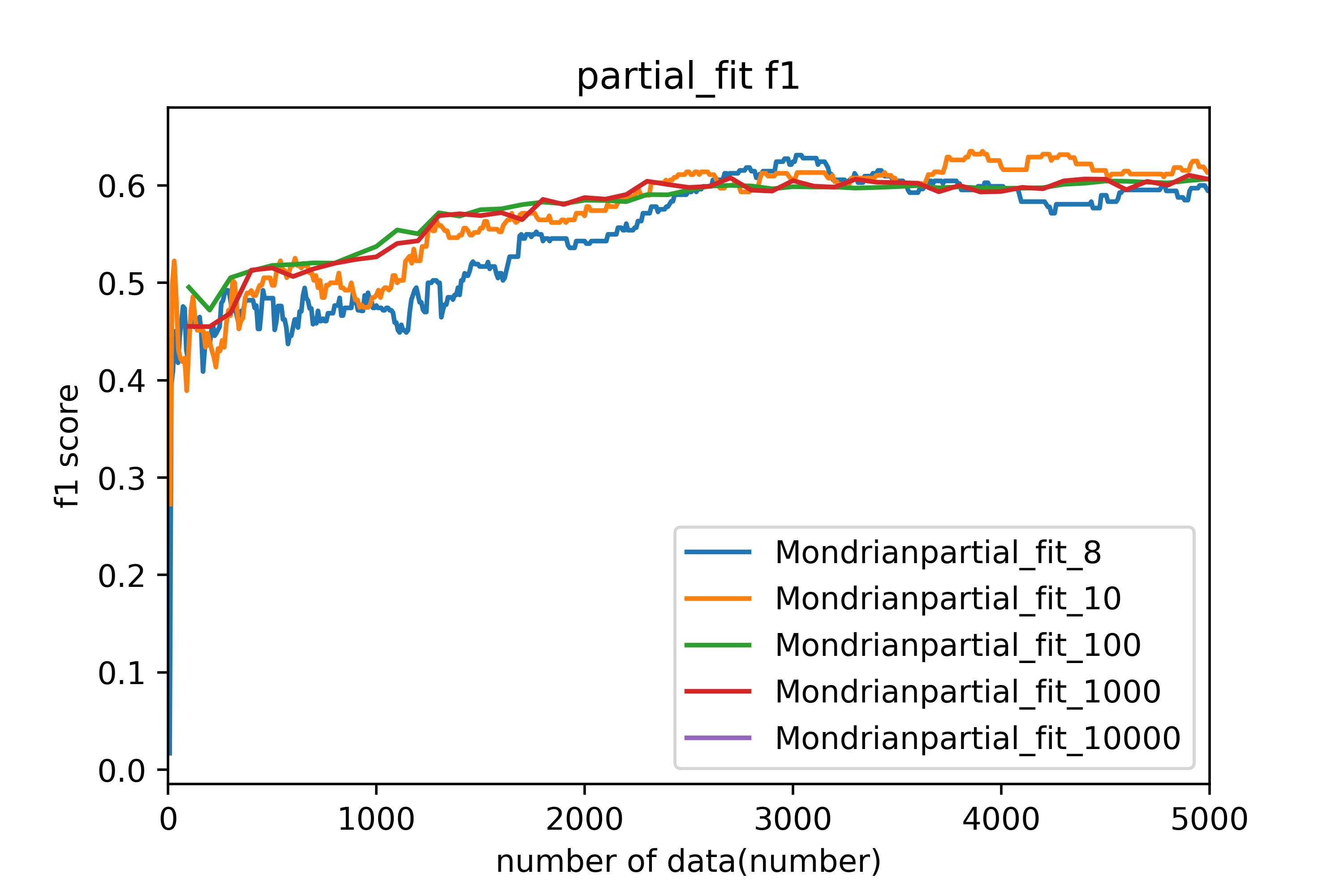
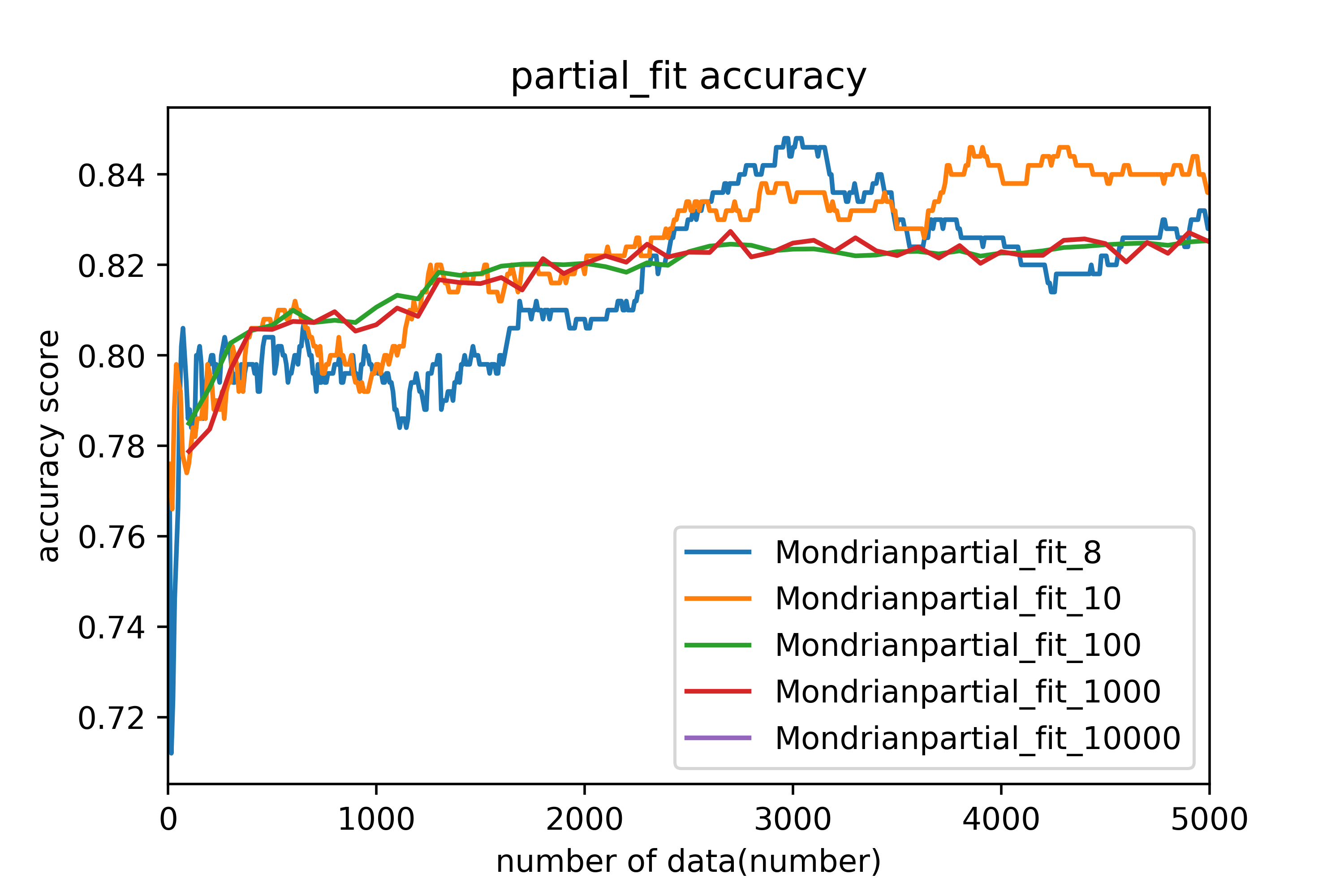
8個ずつなのは、一度に読み込ませるデータが7個以下だとエラーになるため8個ずつにした。



また、この時のスコアを示す。ただし、10個ずつと8個ずつのものは、他のものはテストデータ16281個でスコアを計算しているが、計算が重すぎたので500個で行なっているため乱数性がある。



軸を変えて



一気に多くのデータを読み込ませた方が時間が少ないことがわかった。

また、先ほども同じようなものを示したが、このデータではデータを一度に取り込む量とスコアに相関性はなかった。

まとめ。

・MondrianForestのpartial\_fitやfit自体の計算時間は早くはなく、fitの方が早かった。

・MondrianForestは精度に少し懸念がある。

・MondrianForsetで一気に学習させた場合(batch)と追加学習の場合(online)では精度に差はみられなかった。

・MondrianForestでは一度に読み込ませるデータ数で精度の変化がみられず。また、一度にたくさんのデータを読み込ませた方が計算が早かった。