Employee Future Detection

نظرة تجريدية:

في عالم إدارة الموارد البشرية يزداد الاهتمام أكثر فأكثر بالموظفين وخصوصاً أولئك الذين يشغلون مناصب مهمة وحساسة، ويزداد الاهتمام بالرحيل المفاجئ لأحدهم الأمر الذي يؤدي إلى خلل كبير في خطط واستراتيجيات الشركة وإنتاجها على المدى القريب. تعويض أحد الموظفين المهمين ليست بالمهمة السهلة لأنه حتى وإن وجد معوض بالخبرة والمهارة الكافية فإنه يحتاج إلى بعض الوقت الثمين من أجل أن ينخرط في أعمال الشركة. لذلك ظهرت دراسات نفسية وسلوكية عديدة من أجل تحليل الموظفين بناء على عدة معايير لتجنب مثل هذه الحالات، فظهرت الحاجة إلى الاهتمام بالموظفين الذين يعانون في عملهم أو في حياتهم الخاصة. وكما تساعد هذه الدراسات في معرفة مدى رضا الموظف عن عمله وأجره الشهري ودراسة العوامل التي من الممكن أن تؤدي إلى تفضيله العمل في مكان آخر. هنا تظهر أحد استخدامات الذكاء الصنعي في مجال إدارة الموارد البشرية حيث تستخدم خوارزميات التعلم الآلي للتنبؤ برحيل أحد الموظفين بناء على مشاهدات وقياسات ومعلومات تم جمعها وملاحظتها على الموظف. تهدف هذه الدراسة إلى استخدام خوارزميات الذكاء الصنعي والتعلم الآلي لمساعدة رؤساء قسم إدارة الموارد البشرية في الانتباه إلى الموظفين المحتمل مغادرتهم لمحاولة تصحيح الأمور قبل فوات الأوان.

مصطلحات:

ذكاء صنعي - تعلم آلي - هندسة الميزات - البيانات المصنفة - نموذج آلي - بحث شبكي - مخططات التحقق - مجموعة بيانات - خوارزميات التصنيف - تحليل المكونات الرئيسية

مقدمة:

في الوقت الحالي، انتشرت الدراسات والتحليلات في أقسام الموارد البشرية في الشركات حول الموظفين وجمعت كميات كبيرة من مجموعات البيانات تتضمن معلومات وميزات وخصائص مختلفة عن الموظفين. لذلك تحتاج أقسام الموارد البشرية في الشركات إلى دراسة مجموعات البيانات هذه عن قرب وتحليلها بدقة والاستفادة منها في دراسة مخاطر الفقدان المحتمل لأحد الموظفين مما يضمن سير خطط الشركة بالطريقة الصحيحة وعدم تأثر إنتاجها بمثل هكذا مخاطر.و لإنجاز ذلك تستخدم خوارزميات التصنيف. يشير مصطلح خوارزميات التصنيف إلى توزيع مجموعة من الأشياء على مجموعة من الأصناف، على سبيل المثال تصنيف الأشخاص إلى فقير - غني ، أو كما في دراستنا هذه تصنيف الموظفين إلى قسم يود المغادرة وآخر يود البقاء. تعتمد خوارزميات التصنيف على أسس التعلم الآلي في إنجازها للتصنيف حيث يحتوي التعلم الآلي على مجموعة من الخوارزميات التي تهدف إلى جعل الآلة تنجز مهام ينجزها الإنسان عادة بدقة قريبة أو ربما تفوق دقة الإنسان.

بقية الورقة البحثية مقسم على الشكل التالي:

- مجموعة البيانات : وفيها نستعرض مجموعة البيانات المستخدمة في هذه الدراسة لتحقيق الغاية المرجوة.
- الأعمال السابقة التي تمت على نفس مجموعة البيانات.
- المعالجة المسبقة لمجموعة البيانات: وفيها سيتم عرض مجموعة من التعديلات والتصحيحات التي تم إجراؤها على مجموعة البيانات.
- النماذج الآلية المستخدمة والنتائج: فيها يتم عرض مجموعة النماذج الآلية التي

تم استخدامها مع عرض نتائج هذه النماذج والمقارنة بينها.

- مقارنة النتائج مع الأعمال السابقة.
- الخاتمة والأعمال المستقبلية.

مجموعة البيانات:

تم الحصول على مجموعة البيانات من موقع Kaggle الشهير [1] ، وتتكون مجموعة البيانات من 9 ميزات (متضمنة الهدف) على الشكل التالي:

1- مستوى التعليم: جامعي - ماستر - دكتوراة

2- سنة الانضمام: العام الذي انضم فيه الموظف إلى الشركة.

3- المدينة: اسم المدينة التي يعمل فيها الموظف.

4- صنف الراتب: عالي - متوسط - منخفض

5- العمر

6- الجنس: ذكر أو أنثى

7- الوضع جانباً: هل وضع الموظف خارج المشاريع لمدة شهر أو أكثر أم لا.

8- الخبرة في المجال: مقدار الخبرة على مقياس بين الصفر وال7

9- المغادرة : هل سوف يغادر الموظف الشركة في العامين القادمين أو لا.

سيتم استخدام النقاط الثمان الأولى كـ ميزات ، والنقطة الأخيرة كـ هدف.

الأعمال السابقة التي تمت على نفس مجموعة البيانات:

قام [2] ADITYA PRATAMA PUTRA بالعمل على نفس مجموعة البيانات حيث قام بالمعالجة الأولية للبيانات بتحويل جميع البيانات المصنفة إلى أرقام لتصبح قابلة للعمل على خوارزميات التصنيف، وقام بتحويل الميزة سنة الانضمام إلى عدد السنوات التي قضاها الموظف في الشركة، ثم قام برد جميع الميزات إلى نفس المجال ودرب نموذج أشجار القرار وحصل على دقة 80.45% على بيانات الاختبار.

قام [3] ARJUN DATTARAJU بعمل استخراج للنقاط الشاذة من بعض الميزات ومن ثم قام بعمل بعض المقارنات والدراسات على كل ميزة على حدة (مثل نسبة كل صنف في هذه الميزة) ثم قام بتحويل جميع البيانات المصنفة إلى أرقام لتصبح قابلة للعمل على خوارزميات التصنيف، ثم قام برد جميع الميزات إلى نفس المجال، وقام ببناء عدد كبير من النماذج الآلية مثل Logistic Regression - KNN - SVM - adaboost وكانت أفضل نماذجه هي svm مع دقة 80.24%

قام [4] ABDALRHMAN MORSI بعمل تحليل استكشافي للبيانات، ودراسات تتعلق بارتباط الميزات مع بعضها ، وتعامل مع البيانات الشاذة والقيم المفقودة والتكرارات ، ثم قام بتحويل جميع البيانات المصنفة إلى أرقام لتصبح قابلة للعمل على خوارزميات التصنيف، ودرب نموذجين KNN مع دقة 80.87% وSVM مع دقة 82%.

المعالجة المسبقة لمجموعة البيانات:

1- تم عمل فحص على مجموعة البيانات من أجل معرفة فيما إذا كانت البيانات تحتوي على قيم مفقودة وتبين عدم وجود قيم مفقودة.

2- معالجة قيم الميزات (الأعمدة):

- من أجل عمود مستوى التعليم تم تحويل البيانات من بيانات مصنفة إلى بيانات رقمية بالشكل التالي:

0 = Bachelors, 1 = Masters, 2 = PHD

هذا يفرض أن:

Bachelors < Masters < PHD

وهذا الأمر حقيقي على أرض الواقع.

- من أجل عمود المدينة: لا يمكن استخدام الطريقة السابقة لتحويل

البيانات من مصنفة إلى رقمية والسبب في ذلك أنه لا يوجد تراتبية في أسماء المدن والطريقة السابقة تفرض التراتبية لذلك تمت معالجة هذا العمود عن طريق ما يسمى One Hot Encoding.

فيتم الاستعاضة عن هذا العمود بـ3 أعمدة (عدد المدن في مجموعة البيانات) بالشكل التالي إذا كانت المدينة Pune مثلاً:

Pune	New Delhi	Bangalore
1	0	0

- من أجل عمود صنف الراتب: في البيانات الأساسية تم اعتبار 1 بمعنى راتب عالي و2 متوسط و3 منخفض. هنا سنقوم باستبدال كل 1 ب 3 وكل 3 ب 1 من أجل أن يبقى مفهوم التراتبية صحيحاً.
- من أجل عمود الجنس:

Male = 1 , Female = 0

- من أجل عمود الوضع جانباً:

Yes = 1, No = 0

3- تقسيم البيانات إلى train و test بنسبة 80:20

4- هندسة الميزات:

- أفضل الميزات حسب تقنية الحذف العودي للميزات كانت التعليم - سنة الانضمام - Pune - صنف الراتب - العمر.
- أفضل الميزات حسب Random Forest كانت التعليم - سنة الانضمام - صنف الراتب - العمر - الجنس.

كما نلاحظ الحلين يشتركان بـ4 ميزات من أصل 5 وللوصول لحل فاصل بين الحلين تم استخدام اختيار الميزات بمتغير واحد والتي اختارت ميزات Random Forest.

5- تحليل المكونات الرئيسية PCA : تم تطبيق PCA لتقليل أبعاد فضاء الميزات إلى 5 أبعاد.

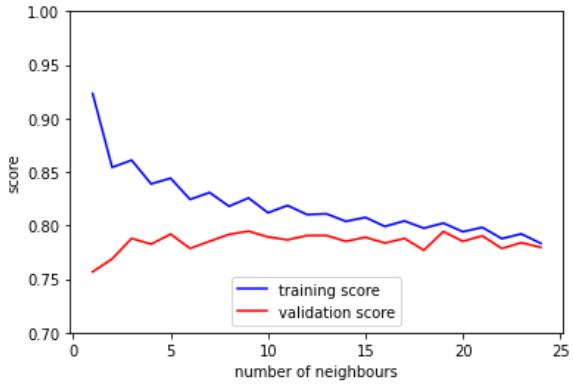
6- رد جميع الميزات لنفس المجال من القيم Normalization: باستخدام MinMaxScaler.

النماذج الآلية المستخدمة والنتائج:

أولاً: بالاعتماد على مجموعة البيانات مع جميع الميزات (السمات) وبدون Normalization:

تم اعتماد dummy classifier كمرجع baseline مع دقة وصلت إلى 66.48%.

- Logistic Regression: 71.1%
- Naive Bayes: 68.95%
- KNN : 73.79%



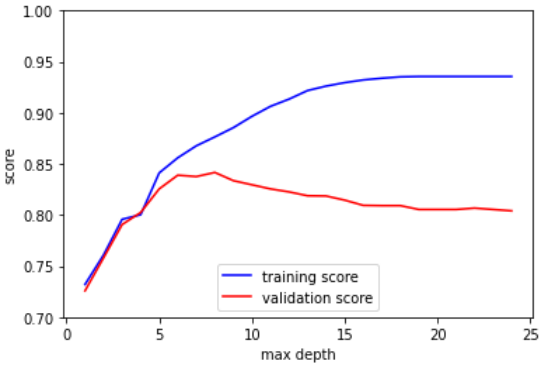
عدد المجاورين في KNN هو Hyper Parameter تم عمل ضبط له بالاعتماد على الvalidation curve مع cross validation يساوي 5 حيث نستطيع ملاحظة الoverfit عند عدد مجاورين 1 وكما نلاحظ أن دقة بيانات الvalidation تبلغ ذروتها عند عدد مجاورين 19.

- SVM : 81.41%

باستخدام grid search تم ضبط المعاملات الخاصة بهذا النموذج والوصول إلى المعاملات الأفضل التالية:

```
'C': 10
'gamma': 0.1
'kernel': 'rbf'
```

- Decision Tree: 83.45%



العمق الأعظمي في Decision Tree هو Hyper Parameter تم عمل ضبط له بالاعتماد على الvalidation curve مع cross validation يساوي 5 حيث نستطيع ملاحظة الoverfit يبدأ تقريباً عند عمق أعظمي 8 وكما نلاحظ أن دقة بيانات الvalidation تبلغ ذروتها عند عمق أعظمي 6.

- Random Forest : 84.31%

باستخدام grid search تم ضبط المعاملات الخاصة بهذا النموذج والوصول إلى المعاملات الأفضل التالية:

```
'max_depth': 9
'n_estimators': 18
```

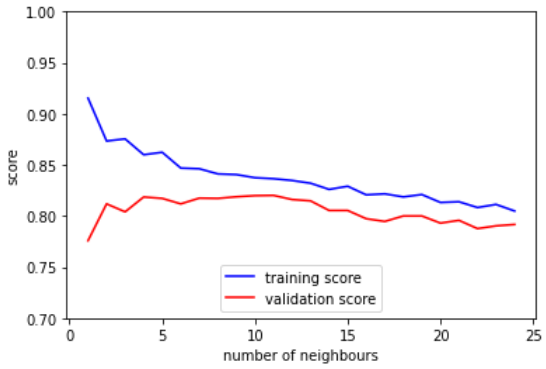
- Adaboost: 82.27%

باستخدام grid search تم ضبط المعاملات الخاصة بهذا النموذج والوصول إلى المعاملات الأفضل التالية:

```
'base_estimator__max_depth': 3
'learning_rate': 1
'n_estimators': 17
```

ثانياً: بالاعتماد على مجموعة البيانات مع جميع الميزات (السمات) مع Normalization:

- KNN: 80.12%



دقة بيانات الvalidation تبلغ ذروتها عند عدد مجاورين 11.

نلاحظ تحسن أداء نموذج الKNN بمقدار 7% تقريباً بعمل Normalization للبيانات.

- SVM: 84.1%

باستخدام grid search تم ضبط المعاملات الخاصة بهذا النموذج والوصول إلى المعاملات الأفضل التالية:

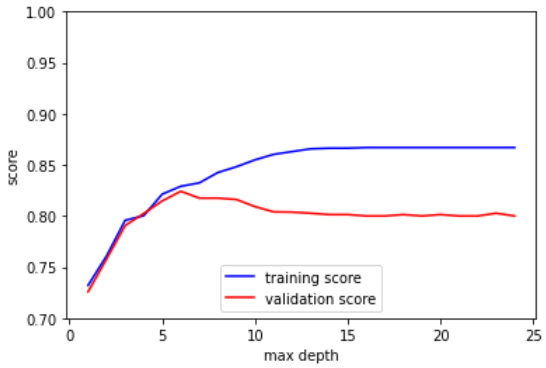
```
'C': 100
'coef0': 0.1
'degree': 3
'kernel': 'poly'
```

نلاحظ تحسن نموذج الSVM من 81.4 إلى 84.1 بمجرد عمل Normalization للبيانات.

ثالثاً: بالاعتماد على مجموعة البيانات مع الميزات المستخرجة من هندسة الميزات (أو PCA) بدون Normalization:

- Decision Tree:

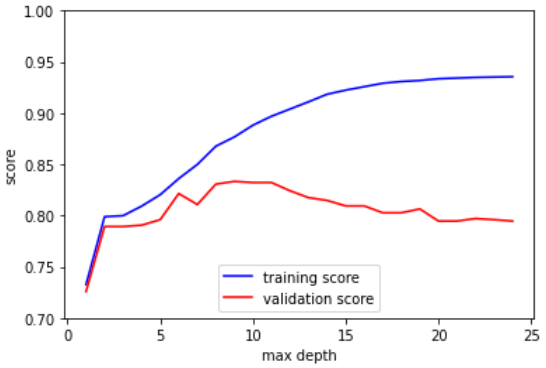
Feature Engineer Data: 80.77%



هنا لا يزال العمق الأعظمي 6 ولكن نلاحظ تراجع أداء النموذج بشكل عام مقارنة مع أدائه

على كامل الميزات حيث نلاحظ تراجع الأداء على منحني التدريب بشكل أساسي وهذا ربما يدل على حذف بعض الميزات المهمة في عملية هندسة الميزات.

PCA Data: 81.2%



نلاحظ أن الأداء تقريباً مشابه للأداء على كامل الميزات مع زيادة عمق الشجرة الأعظمي من 6 إلى 9. ولكن تشير النتائج على مجموعة الاختبار إلى تراجع أداء النموذج.

- adaboost:

Feature Engineer Data: 80.98%

باستخدام grid search تم ضبط المعاملات الخاصة بهذا النموذج والوصول إلى المعاملات الأفضل التالية:

```
'base_estimator__max_depth': 5
'learning_rate': 1
'n_estimators': 4
```

نلاحظ تراجع أداء النموذج وهذا يؤكد أن عملية هندسة الميزات أثرت سلباً على أداء النماذج.

PCA Data: 81.09%

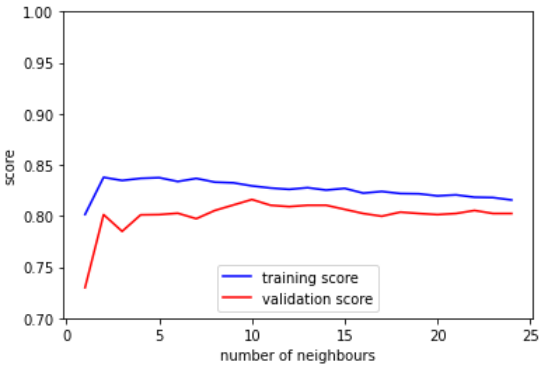
باستخدام grid search تم ضبط المعاملات الخاصة بهذا النموذج والوصول إلى المعاملات الأفضل التالية:

```
'base_estimator__max_depth': 3
'learning_rate': 0.5
'n_estimators': 25
```

رابعاً: بالاعتماد على مجموعة البيانات مع الميزات المستخرجة من هندسة الميزات (أو PCA) مع Normalization:

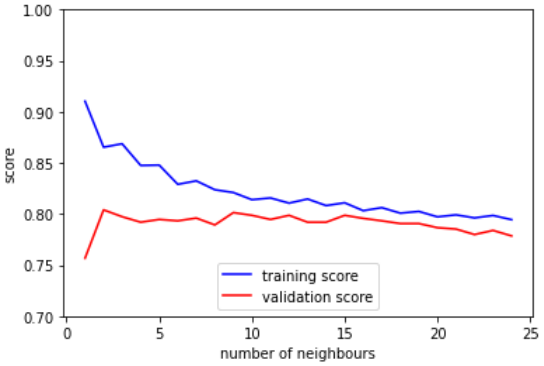
- KNN:

Feature Engineer Data: 80.55%



نلاحظ تحسن طفيف بالنتيجة مع عدد مجاورين 10.

PCA Data: 77.12%



نلاحظ تراجع الأداء مع عدد مجاورين 9.

خامساً: نماذج تجمع أفضل النماذج السابقة:

- Soft Voting: 85.39%

تم استخدام النماذج المدربة مسبقاً Decision Tree و Random Forest و Adaboost في بناء نموذج voting.

- Hard Voting: 84.31%

تم استخدام نفس النماذج السابقة لبناء هذا النموذج.

فيما يلي نتائج جميع النماذج التي جرى تدريبها مع تفاصيل هذه النماذج:

	Model	Data	Normalized	accuracy
0	dummy classifier - Baseline model	Original data	False	0.664876
1	LogisticRegression	Original data	False	0.711063
2	Naive Bayes	Original data	False	0.689581
3	KNN	Original data	False	0.737916
4	KNN	Original data	True	0.801289
5	KNN	Feature Engineering Data	True	0.805585
6	KNN	PCA Data	True	0.771214
7	SVM	Original data	False	0.814178
8	SVM	Original data	True	0.841031
9	Decision Tree	Original data	False	0.834586
10	Decision Tree	Feature Engineering data	False	0.807734
11	Decision Tree	PCA data	False	0.812030
12	Random Forest	Original data	False	0.843179
13	Adaboost	Original data	False	0.822771
14	Adaboost	Feature Engineering Data	False	0.809882
15	Adaboost	PCA Data	False	0.810956
16	Soft Voting (best 3 models)	Original data	False	0.853921
17	Hard Voting (best 3 models)	Original data	False	0.843179

مقارنة النتائج مع الأعمال السابقة:

نلاحظ أن نموذج Soft Voting مع دقة تصل لـ 85.39% يتفوق على جميع نتائج الدراسات السابقة.

وبمقارنة نموذج أشجار القرار الذي قام به ADITYA PRATAMA PUTRA وحصل على دقة 80.45% مع نموذج أشجار القرار الذي تم بناؤه في هذه الدراسة نلاحظ أيضاً تفوق النموذج المبني في هذه الدراسة مع دقة تصل 83.45%.

أما أفضل نماذج ARJUN DATTARAJU فكانت SVM مع دقة وصلت لـ 80.24% ونلاحظ تفوق نموذج SVM المبني في هذه الدراسة مع دقة وصلت لـ 84.1%.

قام ABDALRHMAN MORSI ببناء نموذج KNN بدقة 80.87% (مقابل 80.55% في هذه الدراسة) ونموذج SVM بدقة 82% (مقابل 84.1% في هذه الدراسة).

الخاتمة والأعمال المستقبلية:

المعرفة المسبقة بوقوع خطر محتمل تساعد على تقليل تأثيرات وقوعه أو ربما تلافيها. تساعد هذه الدراسة في التنبؤ بوقوع خطر محتمل لمغادرة أحد الموظفين المهمين أو رؤساء الأقسام مما يمنح رئيس قسم الموارد البشرية القدرة على تلافي آثار هذه المغادرة بزيادة الاهتمام بالموظف أو ربما تحسين راتبه الشهري أو ترقيته ومنحه مسؤوليات أكبر.في الأعمال المستقبلية، نهدف إلى إعطاء تفسيرات أكبر لأصحاب القرار بما يخص مغادرة أحد الموظفين، أي سنحاول معرفة الأسباب التي دفعت هذا الموظف إلى المغادرة حتى يتسنى لأصحاب القرار معالجة المشكلة عن طريق معالجة أسبابها.

المراجع:

[1] https://www.kaggle.com/datasets/tejashvi14/employee-future-prediction

[2] https://www.kaggle.com/code/adityaapp/employees-turnover-prediction

[3] https://www.kaggle.com/code/arjundattaraju/predictingempattrition-1-0

[4] https://www.kaggle.com/code/abdalrhmanmorsi/employee-future-eda-businesssolutions-prediction/notebook