**2024/12/29**

**اميمة محمد مصطفى دلحين**

**09211051**

**مقرر التنقيب في البيانات**

**تقرير المشروع النهائي**

**المقدمة :**

لنبدا حديثنا عن البيانات وهي كما يعرّفها Turban البيانات بأنها "حقائق رقمية أو رموز تُجمع من خلال العمليات المختلفة ويمكن معالجتها لاستخلاص معلومات ذات معنى"، تتخذ البيانات أشكالًا متنوعة، مثل الأرقام، النصوص، الصور، والأصوات، وهي تمثل الأساس لأي عملية تحليلية أو إحصائية.

والبيانات تنقسم الى :

#### 1. البيانات المهيكلة (Structured Data): هي بيانات منظمة ومُرتبة في جداول أو قواعد بيانات، بحيث تكون لها تنسيق محدد يسهل على الأنظمة معالجتها. تتكون عادة من صفوف وأعمدة وتمثل علاقات واضحة بين البيانات.

#### 2. البيانات غير المهيكلة (Unstructured Data):وهي البيانات التي لا تتبع نموذجًا محددًا أو بنية موحدة. تشمل النصوص الحرة، الصور، الفيديوهات، والصوتيات، وتكون عادة أكثر تعقيدًا في التحليل والمعالجة.

#### 3. البيانات شبه المهيكلة (Semi-Structured Data): هي بيانات تجمع بين خصائص البيانات المهيكلة وغير المهيكلة. تكون منظمة جزئيًا لكنها لا تتبع نموذجًا صلبًا، مثل XML وJSON، حيث تحتوي على علامات تدل على معنى البيانات.

بالاضافة الى انواع اخرى ولكن نكتفي بذكر هذه الثلاثة المشهورة ، الان لنعرّف احد اهم علوم العصر وهو علم البيانات ،"علم البيانات يُعرف بأنه فنّ التعامل مع البيانات وتحليلها باستخدام الأدوات الحديثة مثل البرمجة، الإحصاء، والنماذج التنبؤية لتحويل البيانات إلى معلومات ذات قيمة." بعد التعرف على علم البيانات يجب ان نفهم لماذا اصبح علم البيانات من اكثر العلوم شهرة واهمية في هذا العصر، يوفر علم البيانات أدوات تحليل قوية تساعد الباحثين على كشف أنماط غير مرئية سابقًا، مما يفتح آفاقًا جديدة للأبحاث في مجالات مثل الطب، الهندسة، والعلوم الاجتماعية. يستخدم علم البيانات بشكل واسع في المؤسسات ومجال الاعمال بشكل عام حيث يساعد علم البيانات الشركات في تحليل كميات ضخمة من البيانات لتحديد الاتجاهات السوقية، فهم سلوك العملاء، واتخاذ قرارات استراتيجية دقيقة. بالاضافة الى انه باستخدام علم البيانات، يمكن للشركات تحسين خدماتها من خلال تخصيص العروض، توقع احتياجات العملاء، وتحسين رضاهم.

ستنحدث بالتفصيل عن مشكلة من اهم مشكلات العصر التي تهم المؤسسات التي تريد الصعود والقيام بانجازات عالية ،سنتحدث عن الحل المقترح الذي اجتهدت في تكوينه وهو نموذج لتحليل بيانات الموظفين التي نستنج من خلالها وتقسيم للموظفين بناءً على ادائهم ثم سنتحدث عن العمليات التي تمت لمعالجة البيانات وجعلها في صورة مقبولة ومناسبة للنموذج وعن الاختيار للخوارزمية والذي لم يكن بالقرار السهل، سنكشف التفاصيل عن عملية التدريب والتحسيين والتقييم والاختبار ختامًا تنتهي رحلتنا في هذه الجولة لتحليل البيانات عند الحديث عن النتائج والتوصيات التي توصلنا اليها

نموذج تحليلي لتحديد الفرق ذات الأداء العالي والاداء المتوسط والاداء الادنى بناءً على الرواتب والمكافآت وعوامل الإدارة العليا

المشكلة:

كنت ابحث على بيانات على موقع البيانات الشهير kaggle،فوجدت البيانات المسمّاه بعنوان Employees، وجدت بانها تحتوي على 1000 عينة وبها العديد من القيم الفارغة حالة البيانات هي كالاتي :

1. الاسم الأول: الأسماء الأولى للموظفين، مع بعض القيم المفقودة.

2. الجنس: جنس الموظف، مع بعض القيم المفقودة.

3. تاريخ البدء : التاريخ الذي بدأ فيه الموظف.

4. وقت تسجيل الدخول الأخير: وقت تسجيل الدخول الأخير للنظام.

5. الراتب: رواتب الموظفين.

6. نسبة المكافأة: النسبة المئوية للمكافأة.

7. الإدارة العليا: تشير إلى ما إذا كان الموظف في الإدارة العليا (صواب/خطأ)، مع وجود قيم مفقودة.

8. الفريق: فريق الموظف، مع بعض القيم المفقودة.

ولكن بمجرد الاطلاع عليها بدات تساؤلات عديدة تمرّ بخاطري وهي :

الشركات والمؤسسات لديها العديد من الموظفين ولكن هل لديهم معلومات كافية عن ادائهم وتقارير سريعة وسهلة ومعرفة شاملة بالذي يؤثر على هذا الاداء؟ وذلك بناءً على بياناتهم المسجلة لديهم بالفعل ،بالتالي لديهم المشكلة التالية: كيف تؤثر العوامل المختلفة مثل الرواتب، نسب المكافآت، والانتماء للإدارة العليا على أداء الفرق في المؤسسة؟ وما هي الخصائص المشتركة التي تميز الفرق الأكثر إنتاجية، بحيث يمكن استخدامها كمعايير لتحديد وتحفيز الأداء العالي عبر المؤسسة؟

هذه الاسئلة حفزتني للبحث عن حل مقترح لهذه المشكلة وذلك لغرض تحسين أداء المؤسسة: عبر تحديد الفرق الناجحة يمكن اتخاذ قرارات استراتيجية واكتشاف اسباب الضعف ونقاط القوة .

### بدأنا بتحميل بيانات الموظفين من ملف CSV لفهم الهيكل الأساسي للبيانات. استخدمنا مكتبة pandas لقراءة الملف وعرض بعض الصفوف الأولى من البيانات لتحليل الأعمدة والمحتويات.

فقمت باستكشاف العلاقة بين العوامل (الراتب، المكافآت، الإدارة العليا) ومحاولة ايجاد تقييم أداء الفرق وذلك باستخدام تقنيات مثل التجميع (Clustering) لتصنيف الفرق حسب أدائها بناءً على الميزات المحددة.

بدات في اول خطوة هي التعامل مع القيم المفقودة ومعالجة البيانات الاولية، عدد القيم المفقودة في كل عمود:

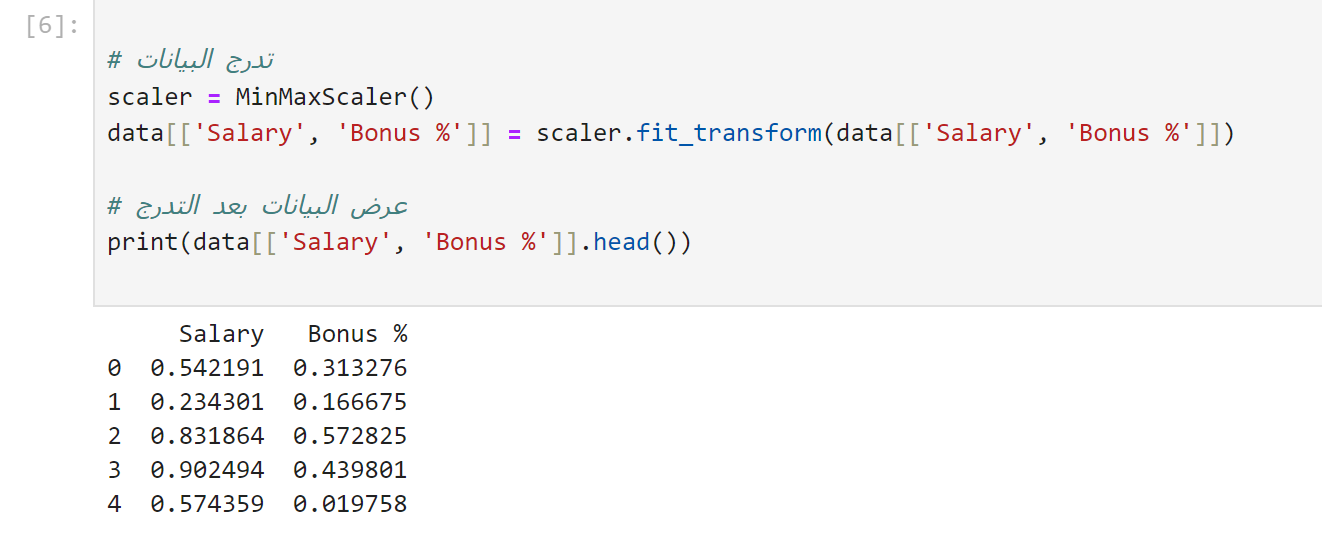
First Name 67 | Gender 145 | Start Date 0

Last Login Time 0 | Salary 0 | Bonus % 0

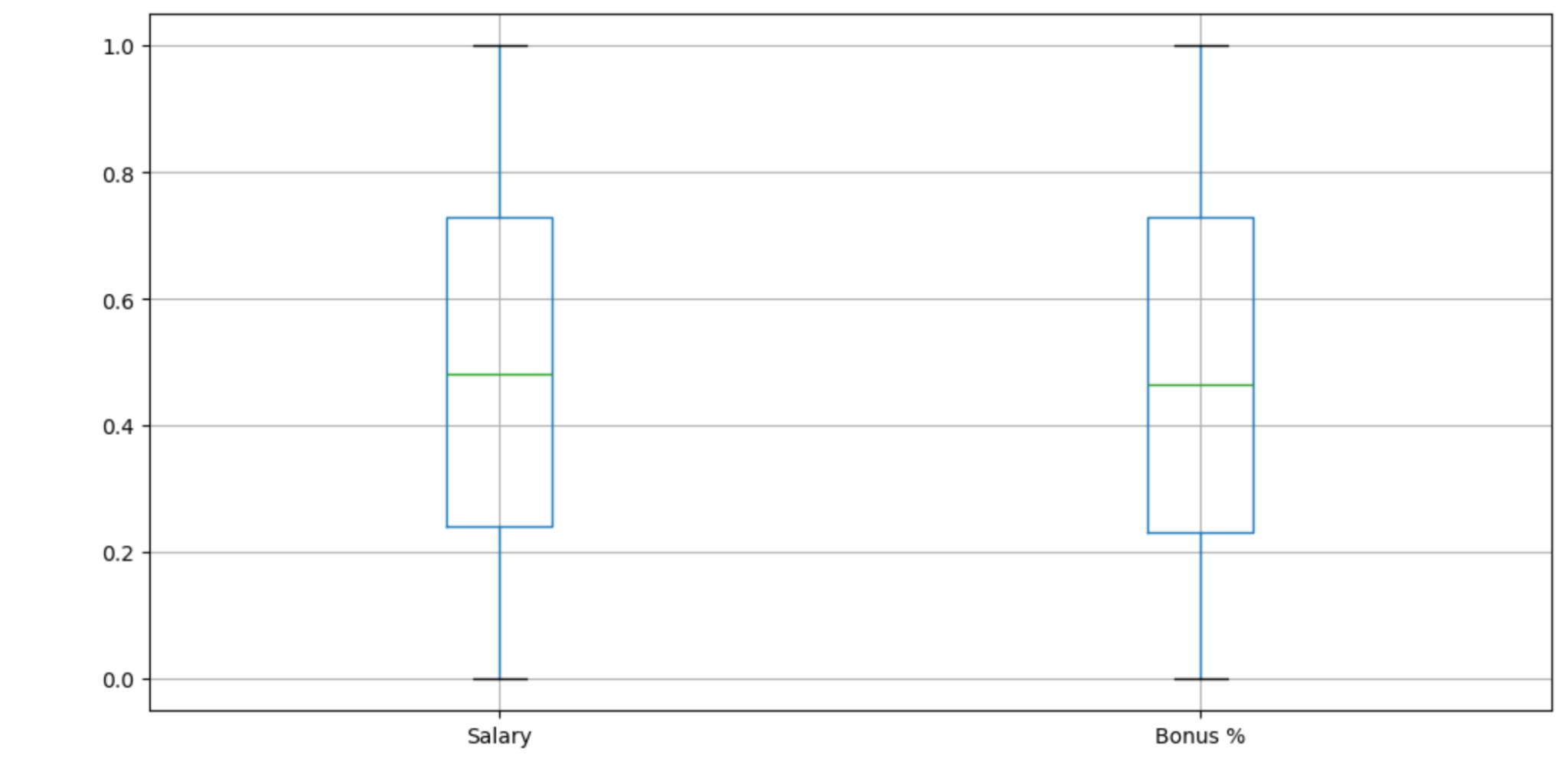
Senior Management 67 | Team 43

ثم قمت بتحويل النصوص إلى تنسيقات datetime لتحليلها بسهولة باستخدام مكتبة panda .

بعد ذلك استخدمت MinMaxScaler وذلك لعمودين قيمهما رقمية Salary،Bonus



الخطوة التالية هي التحليل البصري باستخدام Boxplot للقيم المتطرفة:



يظهر الرسم عدم وجود قيم متطرفة للبيانات بعد المعالجة .

تاكدت بعد ذلك من ان البيانات منطقية ولا تحتوي على اخطاء وادخالات خاطئة قد تتسبب في تشتت وتشويش على النموذج.

وهذه هي حالة البيانات بعد المعالجة :

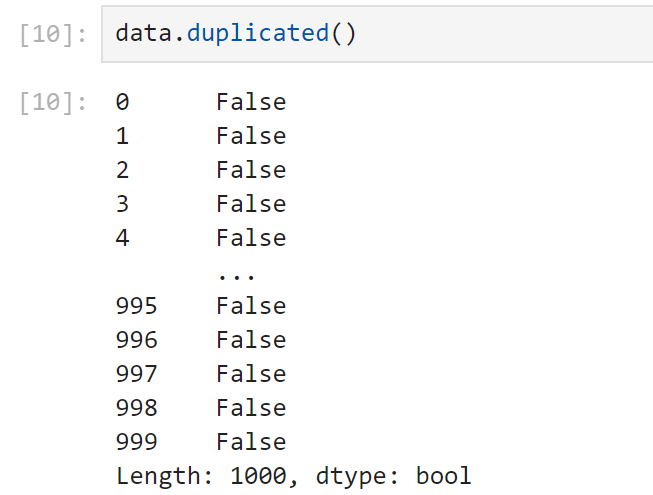
First Name 0 | Gender 0

Start Date 0 | Last Login Time 0

Salary 0 | Bonus % 0

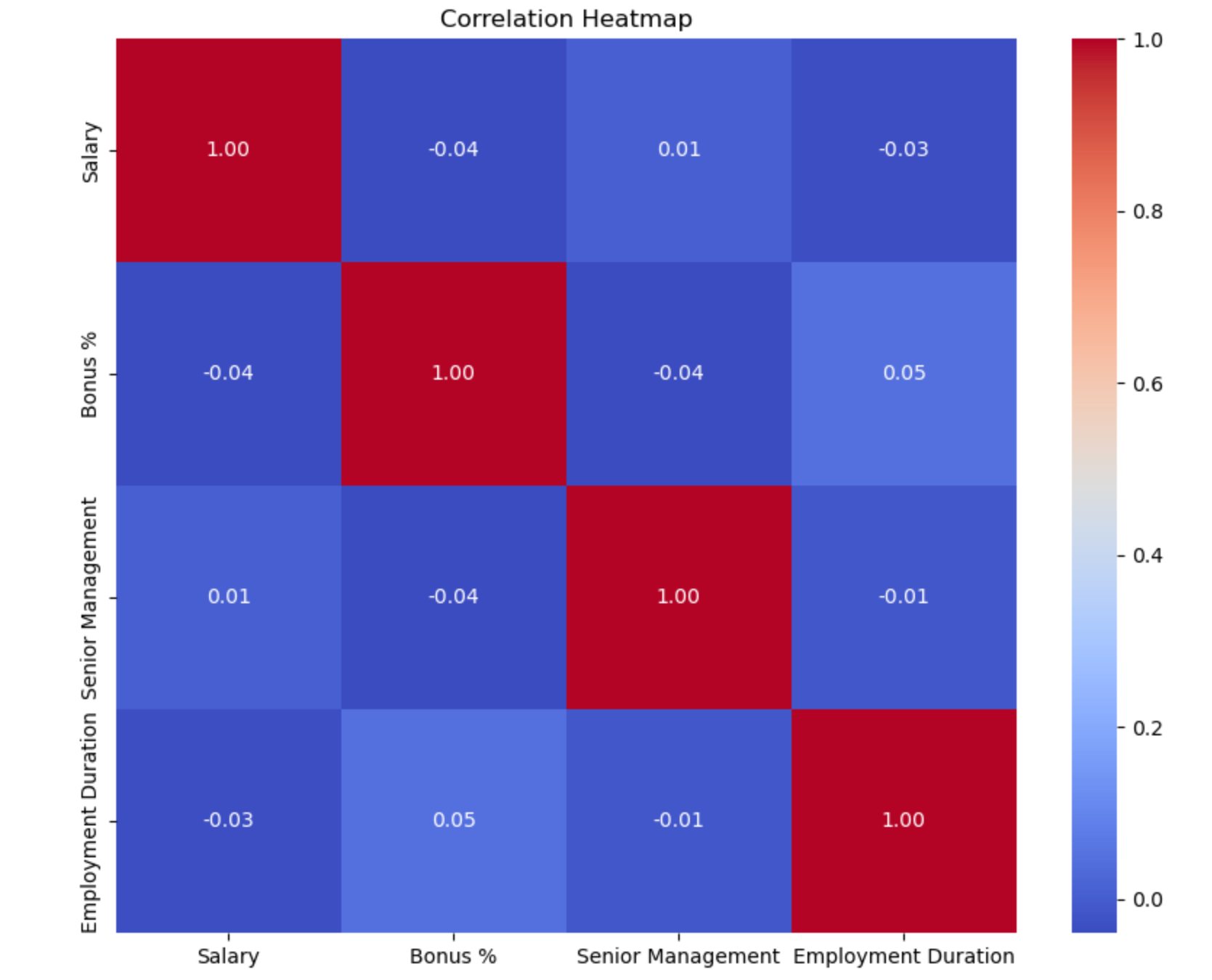
Senior Management 0 | Team 0

بعد ذلك تصنيف البيانات وتحويلها إلى متجهات استخدمت : from sklearn.preprocessing import LabelEncoder تاكدت من عدم وجود صفوف مكررة قد تسبب الخطا للنموذج .



قمت بإضافة ميزة جديدة (مدة العمل) والتي هي ناتج الطرح من start date و last login time وذلك لانها ستكون ميزة مؤثرة في تقييم الاداء للموظف فهي تعكس لنا خبرة الموظف وتاثيرها.

 ثم قمت باكتشاف العلاقات بين الميزات الموجودة في البيانات والتي ستهمني في بناء النموذج وذلك بعرضها باستخدام حساب مصفوفة الارتباط

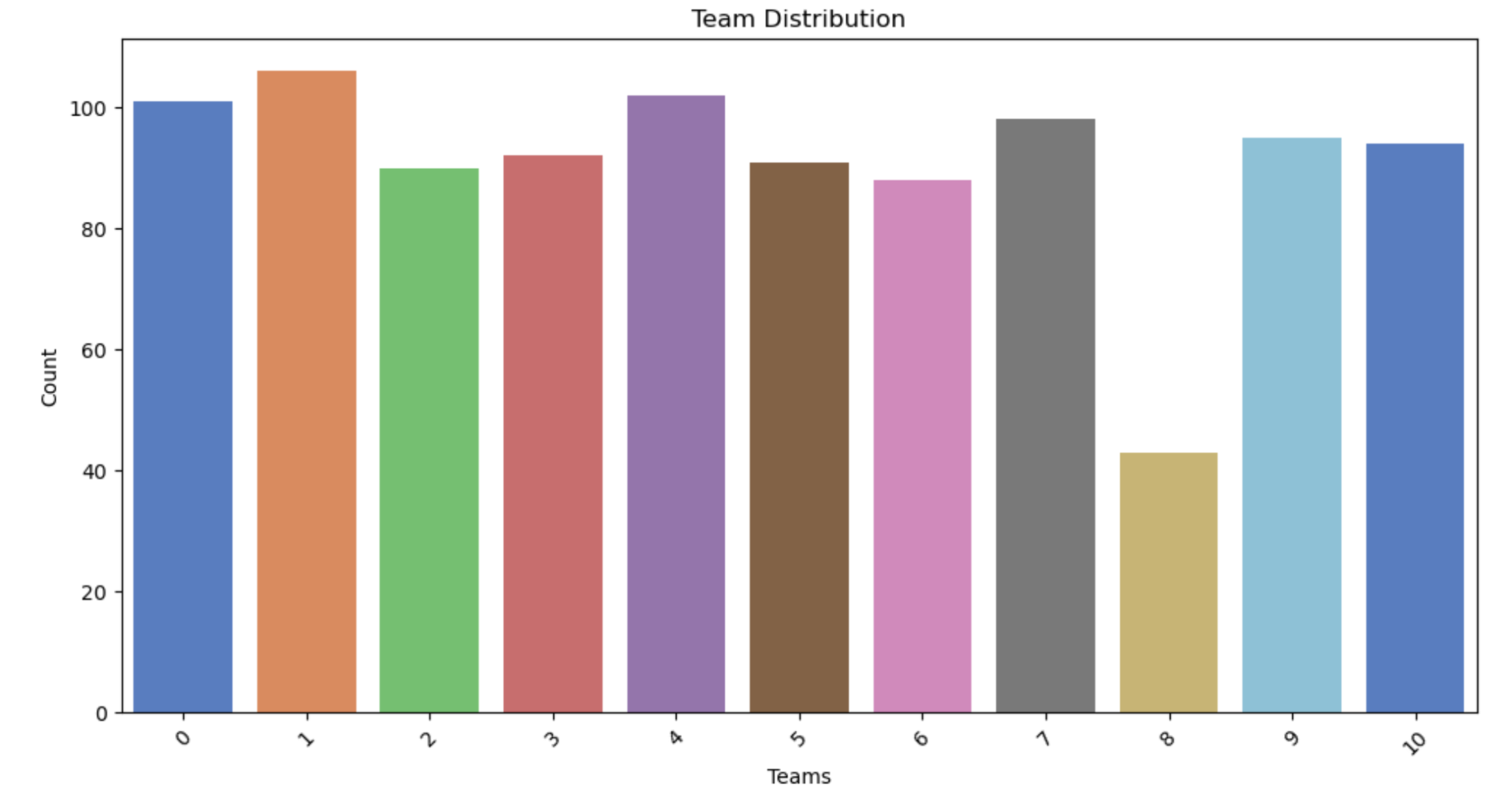


الارتباطات بين المتغيرات هي كما يلي:

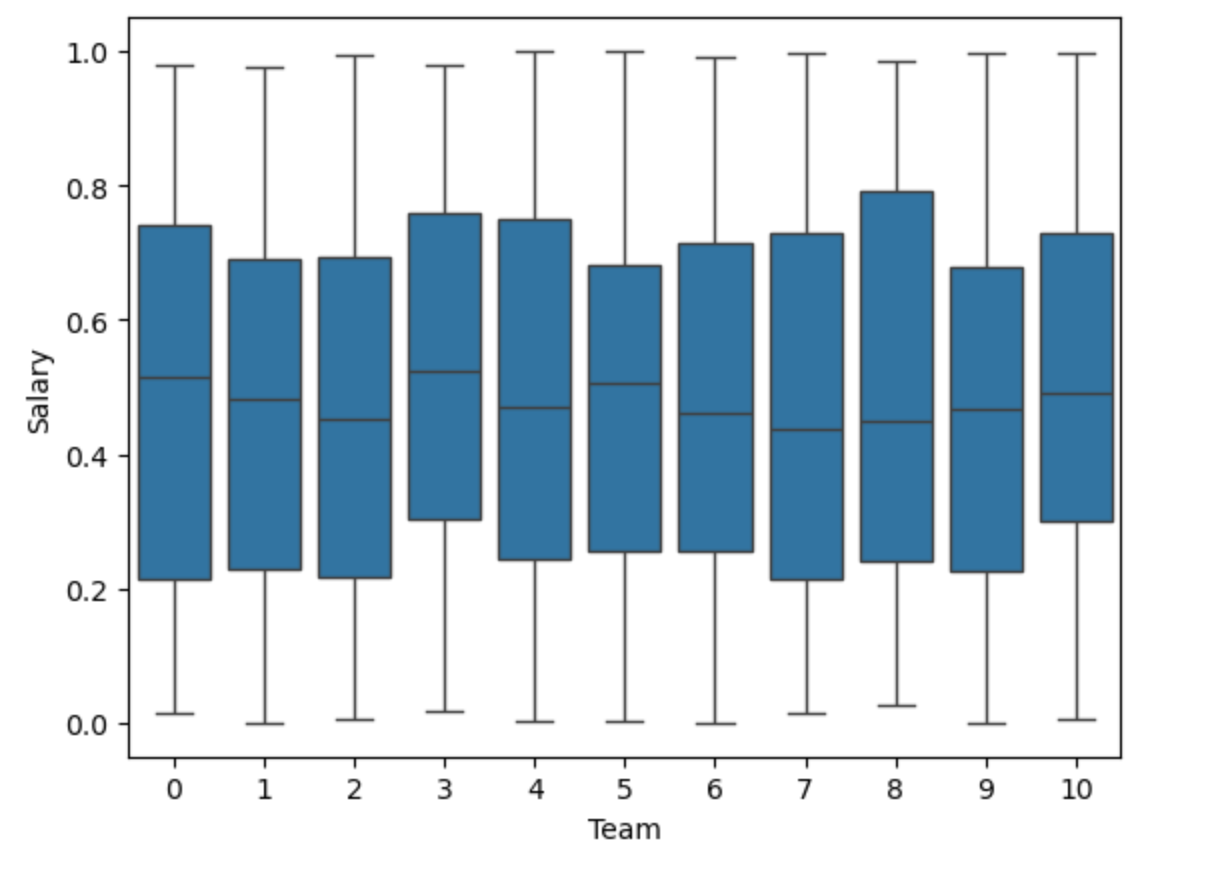
* الراتب (Salary) مع نسبة المكافأة (Bonus %): -0.04
* الراتب (Salary) مع الإدارة العليا (Senior Management): 0.01
* الراتب (Salary) مع مدة التوظيف (Employment Duration): -0.03
* نسبة المكافأة (Bonus %) مع الإدارة العليا (Senior Management): -0.04
* نسبة المكافأة (Bonus %) مع مدة التوظيف (Employment Duration): 0.05
* الإدارة العليا (Senior Management) مع مدة التوظيف (Employment Duration): -0.01
* المتغيرات الأربعة تبدو مستقلة نسبيًا عن بعضها البعض، أي أن أحدها لا يؤثر بشكل كبير على الآخر.

في الحقيقة جربت حساب المصفوفة على اغلب الميزات قمت بحسابه عدة مرات وذلك لاكتشاف الميزات التي لها قيم مناسبة للنموذج وقد تكون مرتبطة وتؤثر بالاخراجات المطلوبة

استعرضت بعد ذلك توزيع البيانات داخل الاعمدة وذلك ايضا للتاكد من ان النموذج لن يقوم بالتحيز لوجود توزيع غير جيّد للبيانات :



وهنا التاكد لاعمدة اخرى وهي salary,team :



بعد ذلك اخترت الميزات بهذا الشكل :

Columns before dropping redundant features:

Index(['First Name', 'Gender', 'Start Date', 'Last Login Time', 'Salary',

'Bonus %', 'Senior Management', 'Team', 'Employment Duration'],

dtype='object')

Columns after dropping redundant features:

Index(['Gender', 'Start Date', 'Salary', 'Bonus %', 'Senior Management',

'Team', 'Employment Duration'],

dtype='object')

في هذه المرحلة لم اكن متاكدة مما اذا كانت ميزات مثل الجنس والفريق مهمة ام لا لذلك لم اقم بحذفها هنا ،بعض الأعمدة مثل Gender قد تُسبب تحيزًا غير مرغوب فيه في النموذج إذا لم تكن مرتبطة بالمشكلة.

* تم اختيار الخصائص الرئيسية:
  + Salary

-الراتب يعكس مستوى دخل الموظف وهو مؤشر مهم لتحديد مستوى الخبرة والمهارات.

* + Bonus %

-تعطي فكرة عن تقدير الشركة للموظف بناءً على أدائه.

* + Senior Management

-هذه الخاصية تميّز بين الموظفين الذين يتولون مناصب قيادية أو إدارية وبين غيرهم.

* + Employment Duration

-مدة عمل الموظف تعكس خبرته واستقراره الوظيفي.

هذه الخصائص تغطي جوانب مالية (Salary, Bonus %) وإدارية (Senior Management) وزمنية (Employment Duration).

* هذه الخصائص قابلة للتحليل بسهولة بعد تنظيف البيانات وتمثل متغيرات متباينة يمكن استخدامها لإنشاء مجموعات مختلفة تحقق الغرض المطلوب.

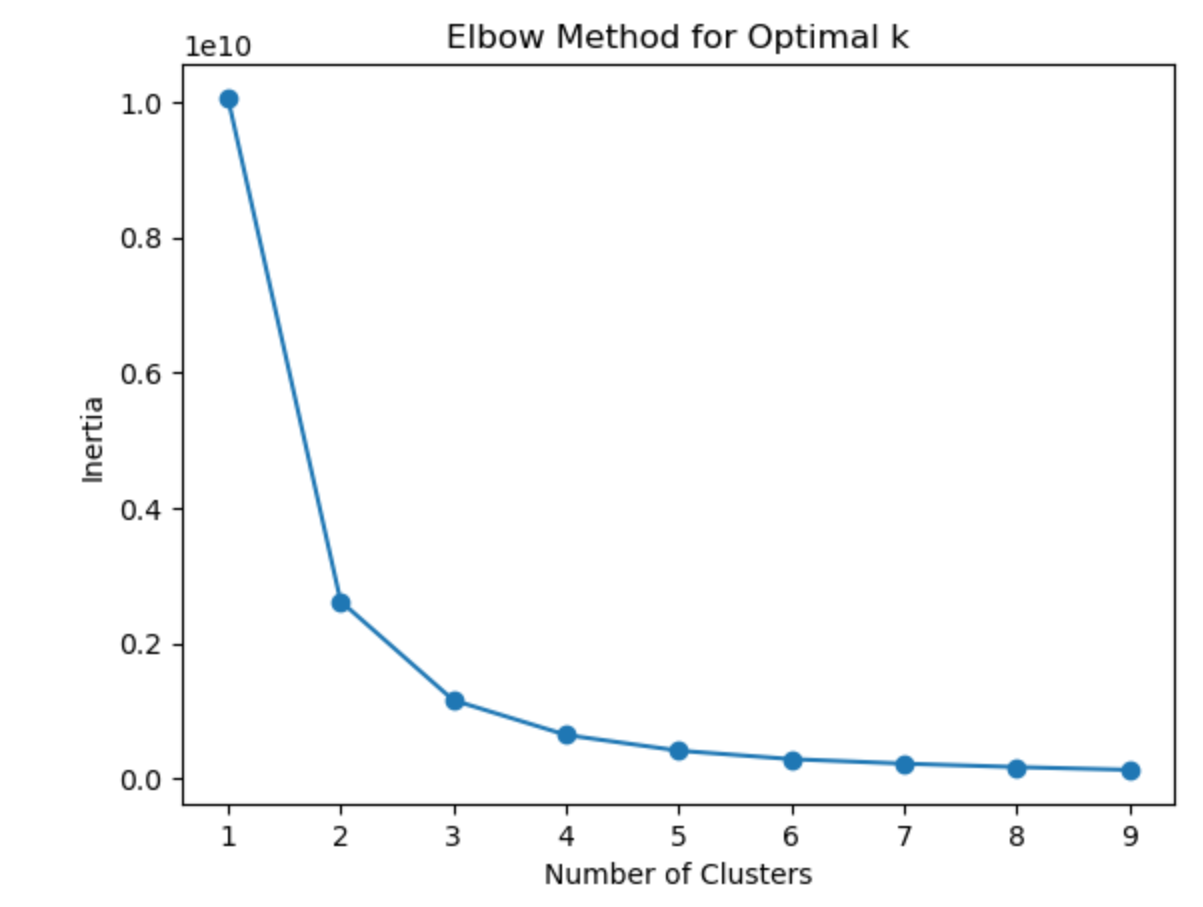
بعد المعالجات الشاملة التي قمت بها بدات رحلة البحث عن الخوارزمية المناسبة فوجدت ان:

التجميع هو تقنية غير خاضعة للإشراف تُستخدم لاكتشاف الأنماط الطبيعية داخل البيانات. في مشروعك، يمكن استخدامه لتصنيف الفرق بناءً على الأداء (مرتفع، متوسط، منخفض) دون الحاجة إلى معلومات مسبقة عن النتائج.

الخوارزميات مثل K-Means تقدم نتائج يمكن تفسيرها بسهولة عن طريق تصور البيانات (مثل الرسوم البيانية) لفهم الفروقات بين المجموعات.

استخراج القيمة من البيانات، اتخاذ قرارات مدروسة، وتحقيق أهداف المشروع بكفاءة هذه هي المعايير التي كنت ابحث عنها بالتالي تم استخدام خوارزميات K-Means، DBSCAN، وAgglomerative Clustering لتجميع البيانات بعد معالجة البيانات

بشكل اساسي تم استخدام K-Means Clustering كخوارزمية تحليل بيانات للتجميع.

استخدمت دالة elbow لايجاد افضل عدد للتجمعات : 

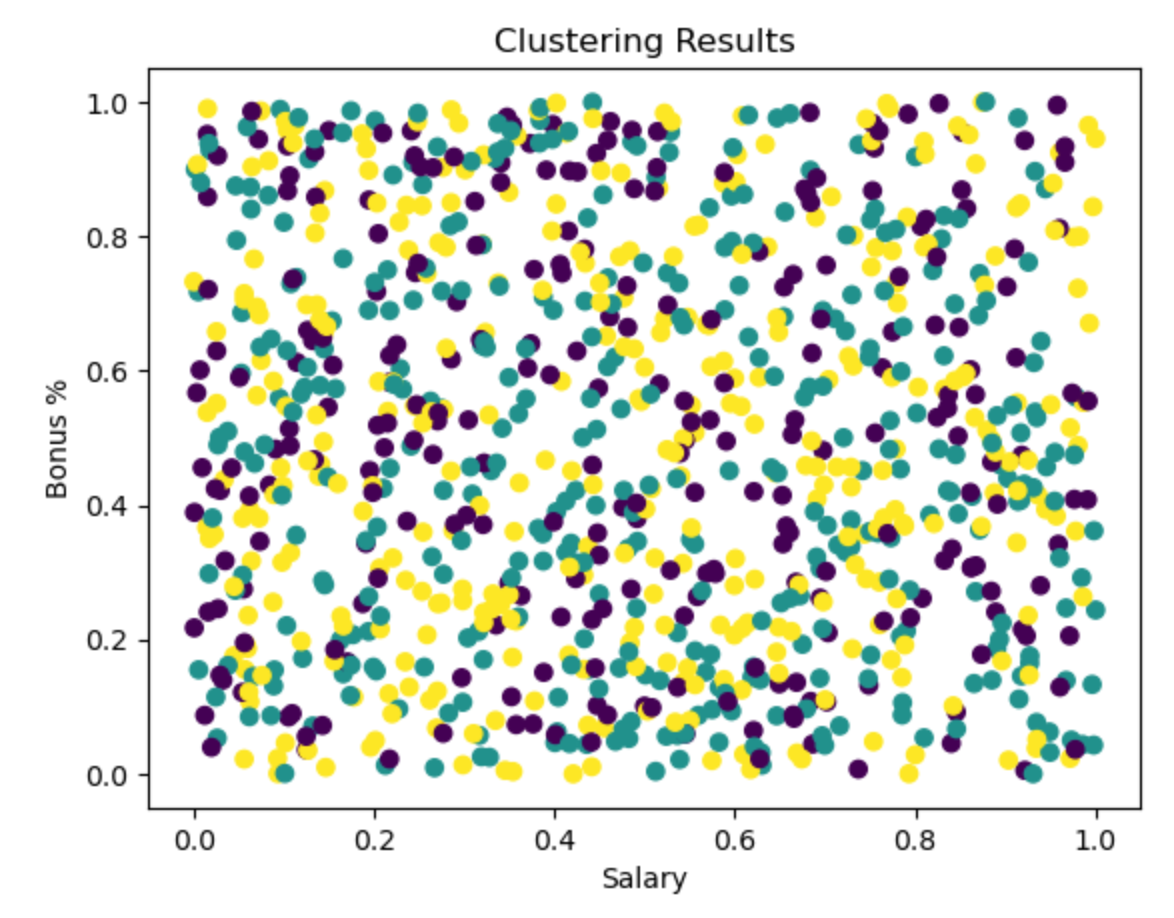
استنادا إلى الرسم البياني لطريقة الكوع ، فإن العدد الأمثل للمجموعات هو 3.

هناك "كوع" ملحوظ عند k = 3 ، مما يشير إلى أن القصور الذاتي ينخفض بشكل كبير حتى هذه النقطة ثم ينخفض المستوى، يشير هذا إلى أن ثلاث مجموعات هي الخيار الأفضل.

### بدايةً تم تقسيم البيانات الى :

### 70% training , 15% validation , 15% testing

### ثم تدريب الخوارزمية على البيانات واستعراض النتائج للنموذج :

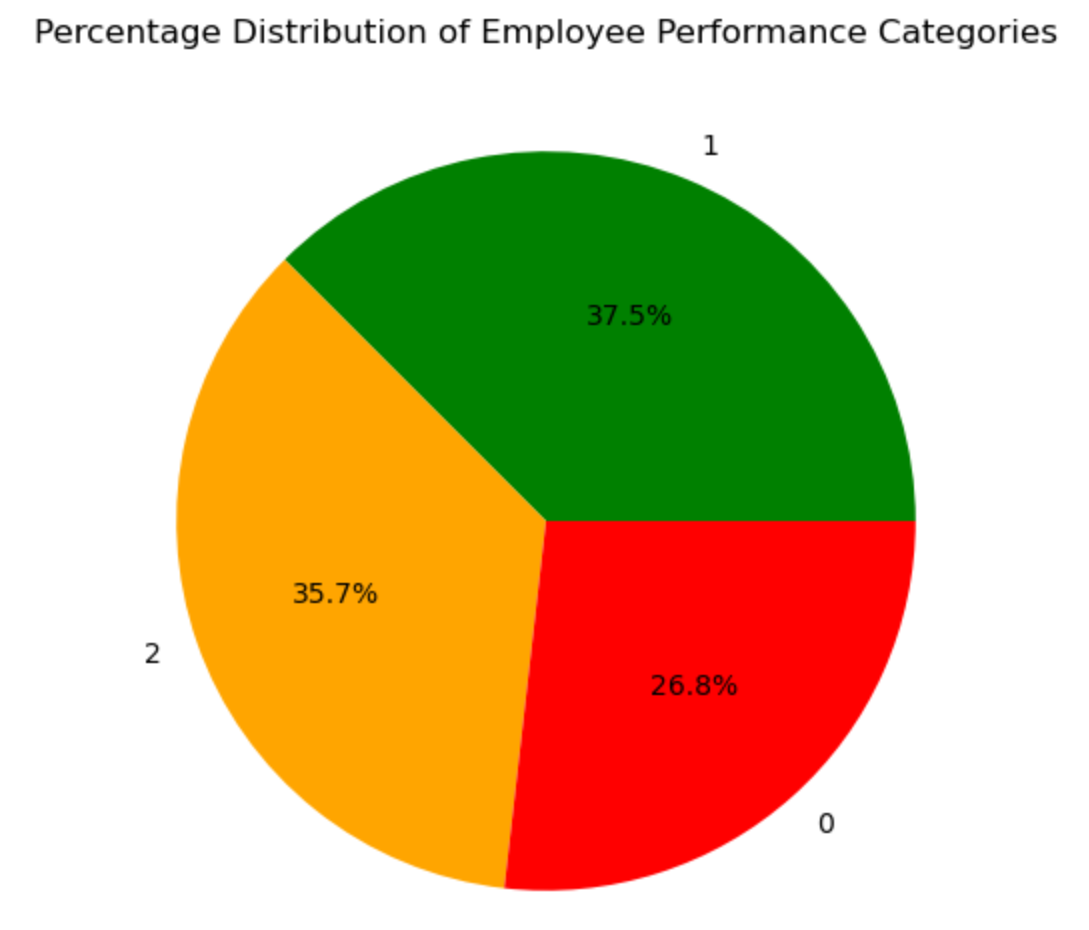
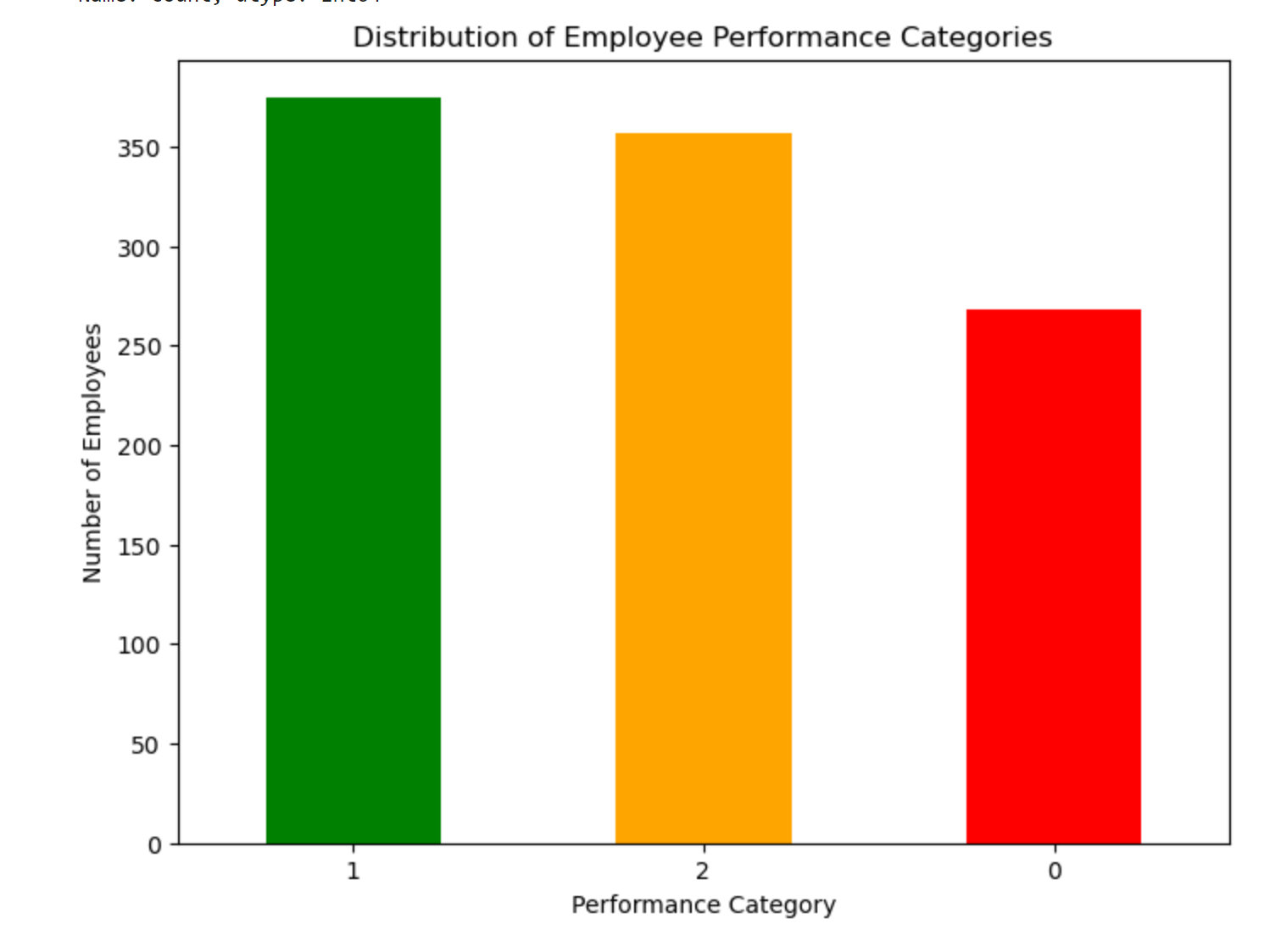


ثم ايجاد دقة النموذج على بيانات التدريب ,والاختبار والتقييم باستخدام Silhouette:

Training Silhouette Score: 0.35

Validation Silhouette Score: 0.60

Test Silhouette Score: 0.59

تمثل الصورة التالية جودة نتائج نموذج K-MEANS : 

ثم لتقييم النتائج قمت ب تحديد الفئة الأكبر ومقارنتها بالباقي وكانت النتيجة :

The largest category is '1' with 37.50% of employees.

The classification appears balanced.

بعد ذلك حساب المسافات داخل التجمعات

Mean Intra-Cluster Distance: 0.82

أن التجمعات ليست شديدة التقارب، لكنها ليست متباعدة بشكل مفرط.

خلال التقييم يجب ان ننتبه الى تقييم جودة التجمعات: درجة Silhouette تعطي فكرة عن كيفية توزيع النقاط داخل التجمعات. درجة Silhouette العالية (قريبة من 1) تشير إلى أن النقاط متجمعة بشكل جيد حول مركز التجمع الخاص بها وبعيدة عن التجمعات الأخرى.

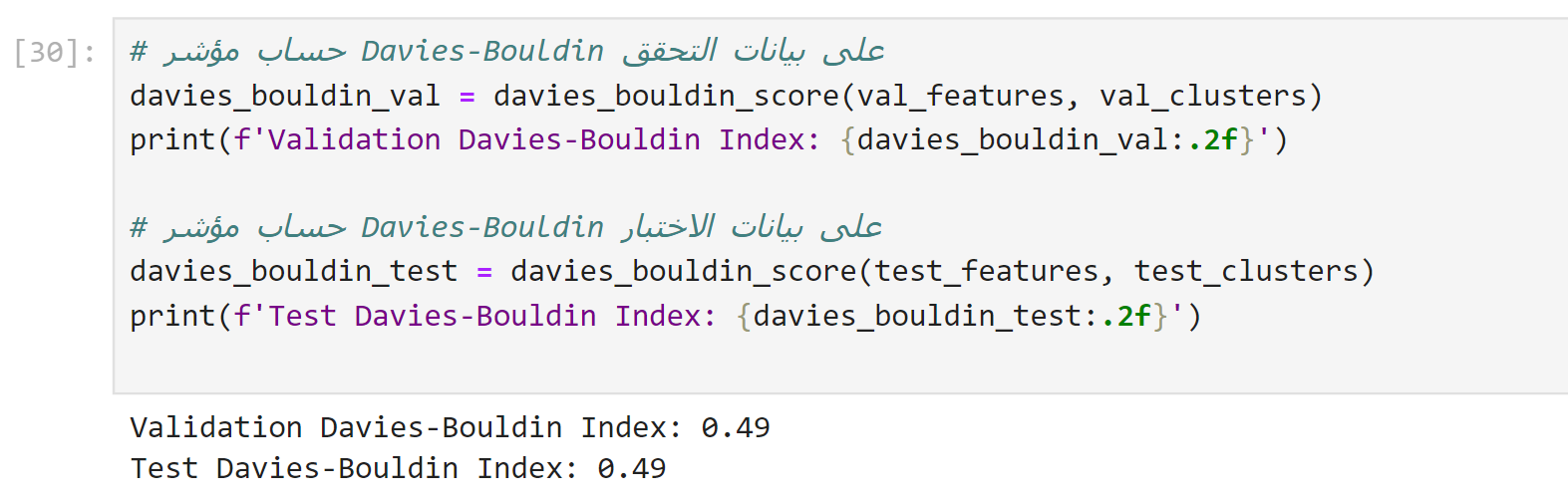
حساب الـ Silhouette لكل نقطة

0=>0.692526

1=>0.593132

2=>0.527636

حساب مؤشر Davies-Bouldin على بيانات التحقق : هذا المؤشر هو مقياس اخر لتقييم جودة التجمعات والقيم المنخفضة منه تعني الاداء الافضل :



ثم قمت بتحليل الاداء للنموذج مقابل عدد التجمعات باستخدام الرسم والذي يساعد في ضبط عدد العناقيد المثالي لتحسين جودة النموذج. بعد تحديد العدد الأمثل، يمكنك استخدامه في تطبيق النموذج على بيانات جديدة للتنبؤ بالعناقيد.

التغيير في الأداء مع عدد المجموعات يشير إلى أن النموذج حساس لتغيير عدد المجموعات. يجب اختيار عدد المجموعات الأمثل للحصول على أداء جيد.

بعد ذلك لاستعراض وفهم طريقة عمل النموذج :



الخريطة الحرارية تُظهر القيم المركزية للميزات الثلاثة لكل مجموعة:

Cluster 0:

Salary: 0.40/ Bonus: 0.34 /Employment Duration: 0.77

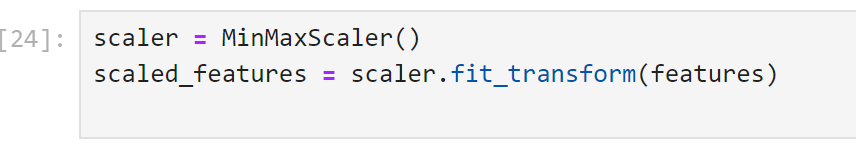
Cluster 1:

Salary: 0.45/Bonus: 0.79/Employment Duration: 0.45

Cluster 2:

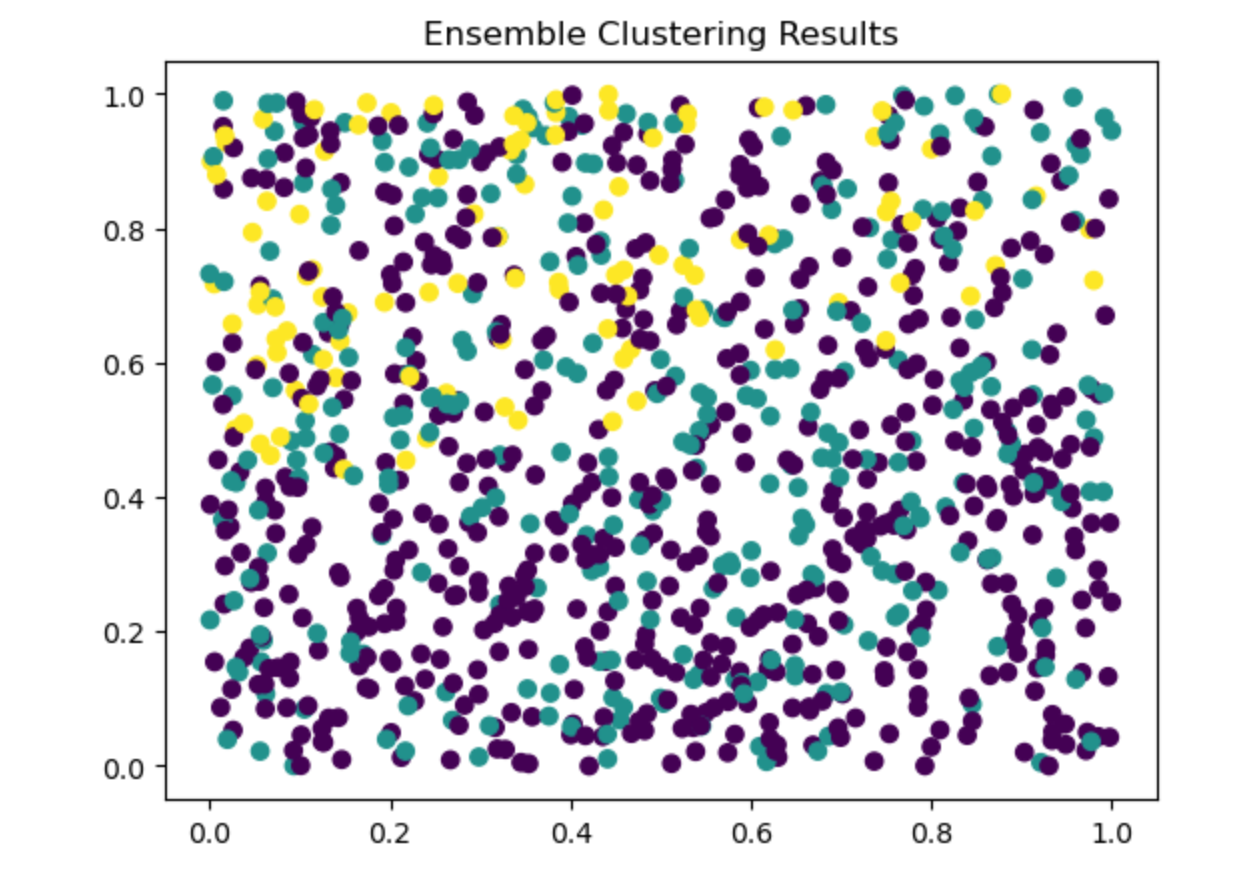
Salary: 0.58/Bonus: 0.26/Employment Duration: 0.27

استخدمت تطبيع البيانات :



ثم ادخلت هذه البيانات على النماذج الثلاثة : kmeans و DBSCAN و Agglomerative ثم تم دمج نتائج النماذج الثلاثة معًا واظهرت النتيجة التالية : Silhouette Score (Ensemble): 0.2219229063173543 ،هذه القيمة تمثل نتيجة النماذج الثلاثة معا وهي ضعيفة بسبب نموذج DBSCAN بشكل اساسي

هذه الخطوة هي لتحسين النتائج المرجوة من النماذج .



كما نلاحظ من الرسم تداخل المجاميع وهذا مؤشر على ضعف اداء النماذج الثلاثة معا ..

وبما ان الاداء ضعيف يجب ان احاول في تحسين نتائجه فوجدت وسيلة للتحسين هي تقليل الأبعاد: لتسهيل التحليل والتصور.

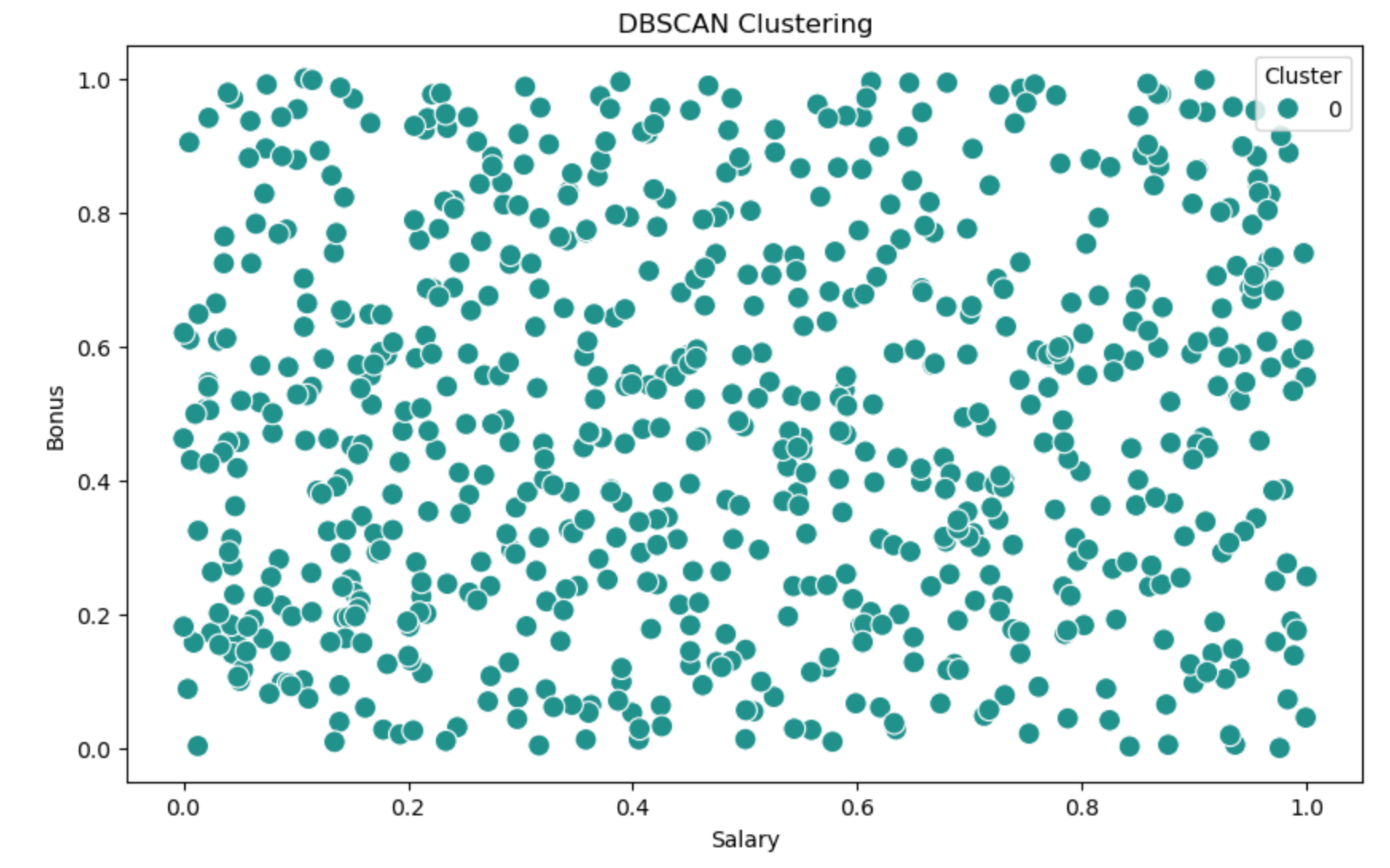
ثم التاكد من التحسين بحساب Silhouette Score: لتقييم جودة التجميع النهائي.

Silhouette Score (EnsembleSilhouette Score (Ensemble): 0.26993778425084175

قيمة 0.27 تقريبًا، تشير إلى أن التجميع لا يزال بحاجة إلى تحسين ولكنه يعكس بعض التمايز بين المجموعات.

بالاضافة الى ذلك جربت تغيير المعلمات لكلا النموذجين وخاصة DBSCAN فقد جربت معلمات بقيم اعلى واقل ولكن دون جدوى .

وكان نموذج DBSCAN يظهر النتيجة التالية:

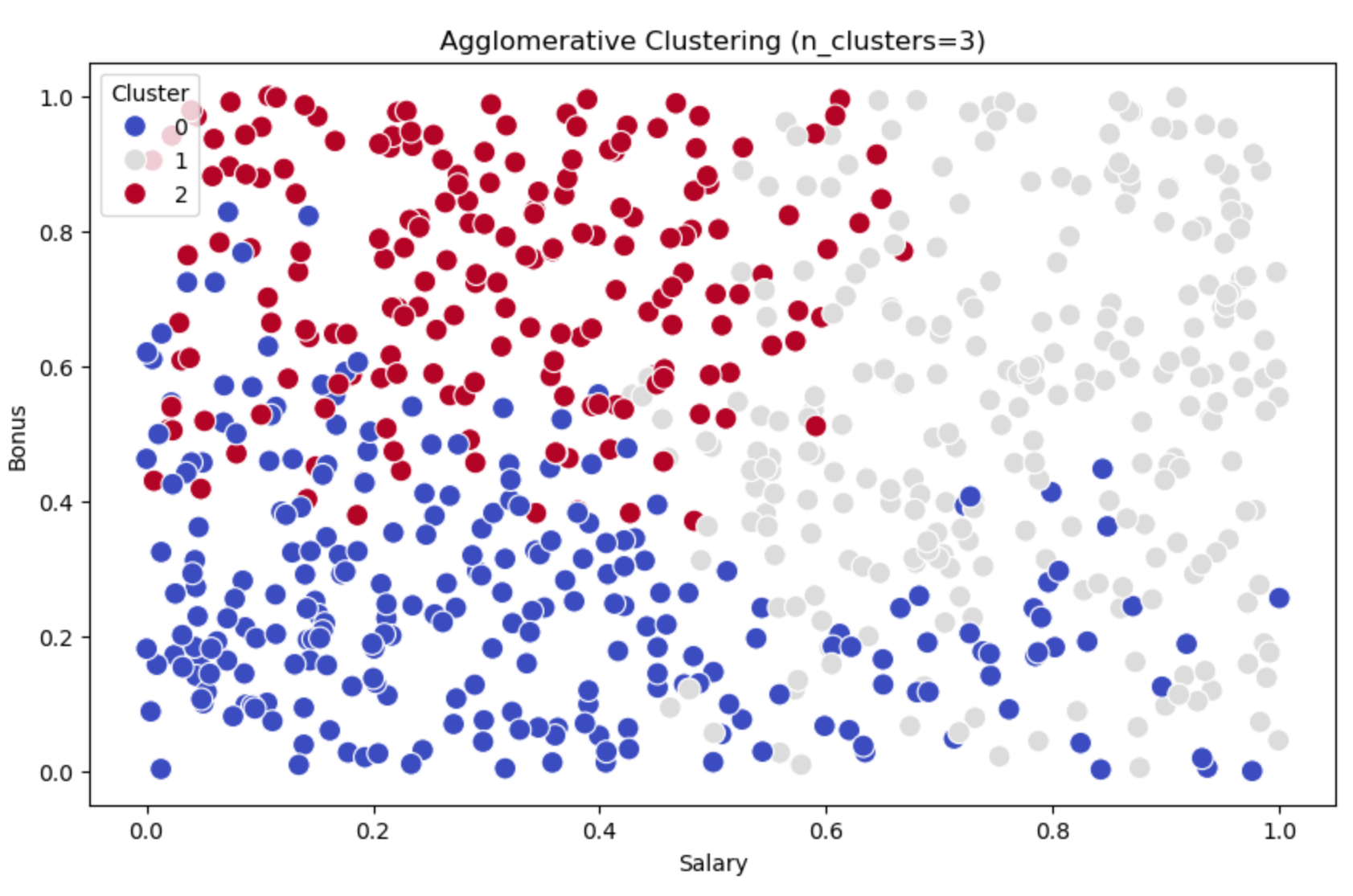
DBSCAN Evaluation:

Silhouette Score: N/A

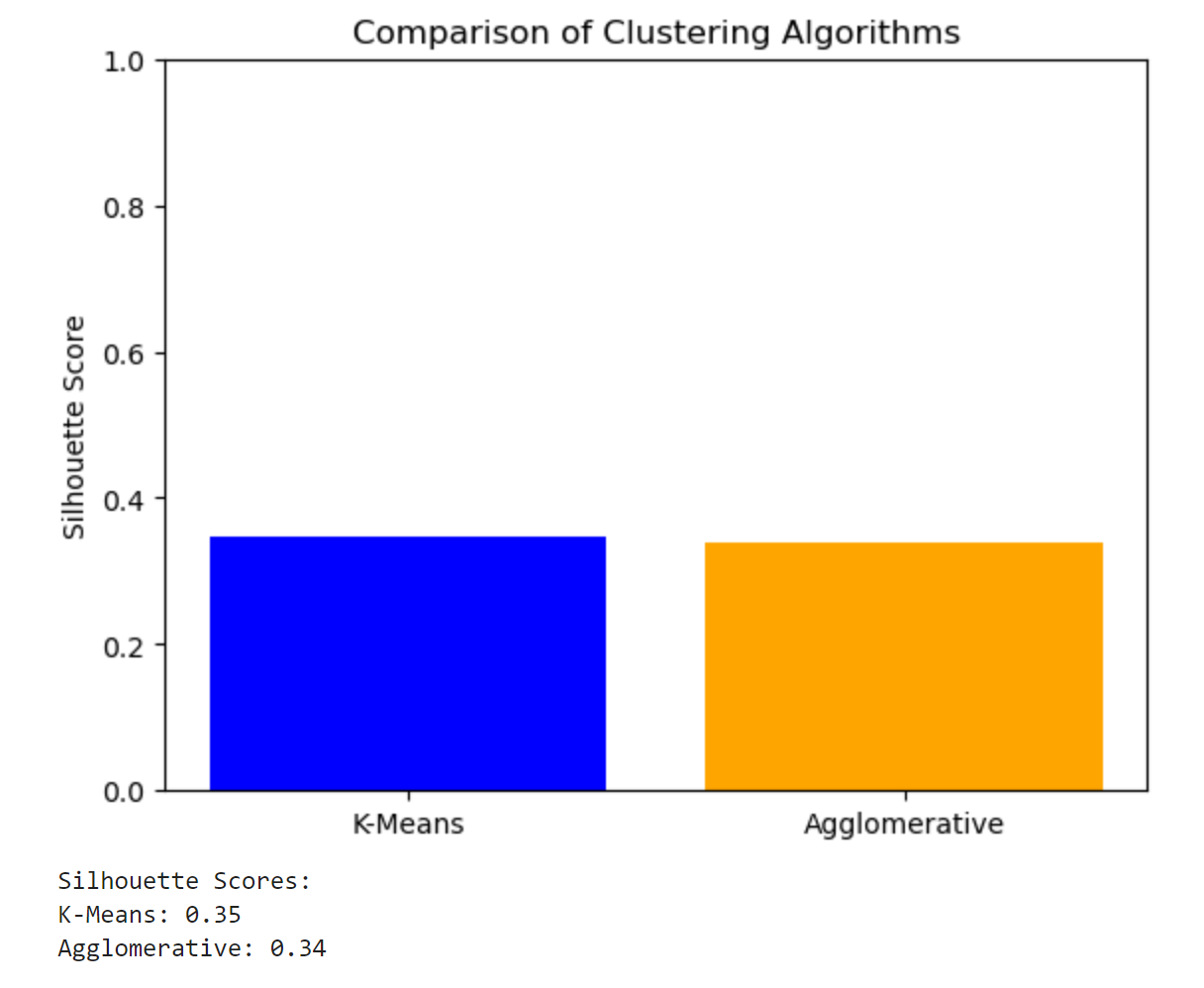
Davies-Bouldin Index: N/A

Calinski-Harabasz Index: N/A

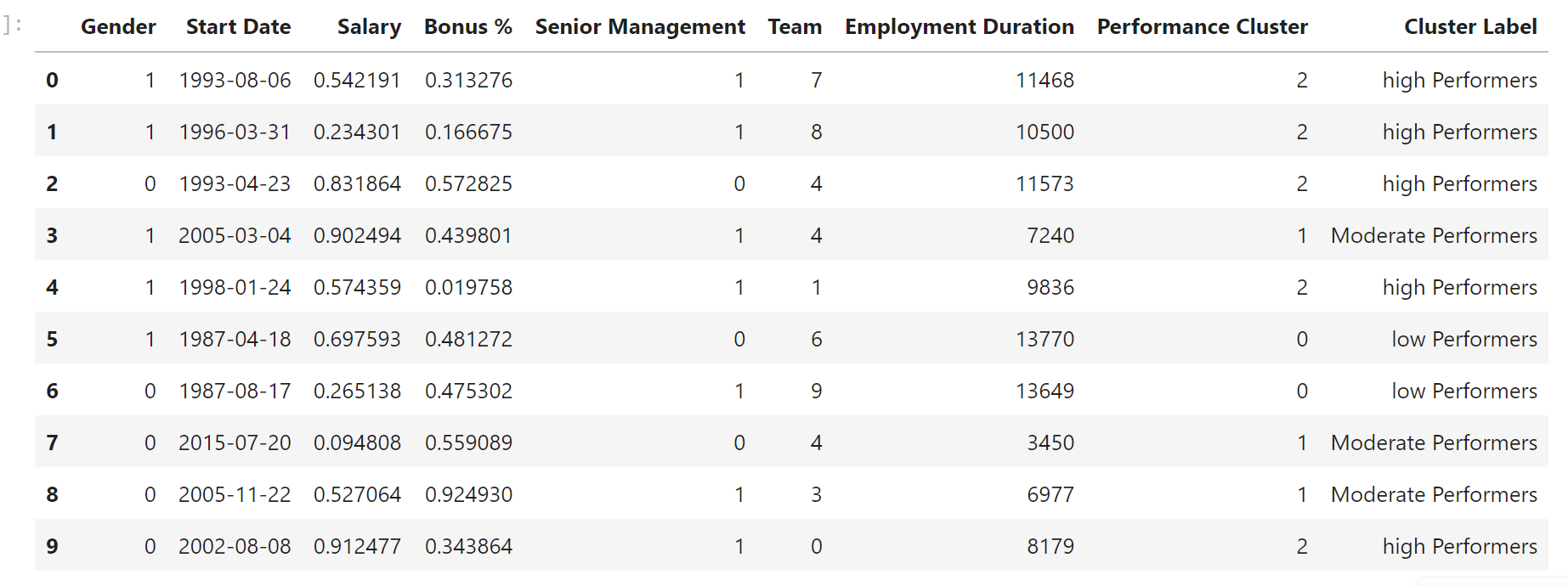
بينما نموذج Agglomerative :



بما ان النموذج DBSCAN لا يتناسب مع البيانات ونوع المشكلة قمت بمقارنة النموذجين الاخرين :



ختامًا هذه عينة من اداء نموذج K-MEANS بعد اضافة تسمية للتجمعات لجعلها اقرب للفهم :



الخطوات النهائية التي قمت بها هي :

حفظ البيانات من الضياع وجعلها في صورة جاهزة للعرض:

Dataset with clusters saved as 'clustered\_employees.csv'

# حفظ النموذج

joblib.dump(kmeans, 'final\_model.pkl')

# تحميل النموذج عند الحاجة

loaded\_model = joblib.load('final\_model.pkl')

عند حفظ النموذج، لن نحتاج إلى إعادة تدريبه في كل مرة نريد استخدامه. هذا يقلل من استهلاك الوقت والطاقة الحاسوبية.

إذا اردت دمج النموذج مع تطبيق حقيقي (مثل واجهة ويب أو تطبيق هاتف محمول)، يمكنك ببساطة تحميل النموذج الجاهز بدلاً من تدريبه أثناء تشغيل التطبيق.

يمكنك تحميل النموذج واستخدامه على بيانات جديدة لتوقع أو تصنيف النتائج.

**الخلاصة :**

في هذا المشروع، تم تطوير نموذج تحليلي لتصنيف الموظفين وذلك على مجموعة بيانات تحتوي 1000 عينة وتصنفهم إلى 3 فرق{ ذات أداء عالي، متوسط، ومنخفض}، اعتمادًا على عوامل مثل الرواتب، المكافآت، والانتماء للإدارة العليا. بدأت الفكرة بعد العثور على بيانات موظفين على منصة Kaggle تحتوي على عينات ناقصة. تمثلت المشكلة الأساسية في كيفية تحليل البيانات المتوفرة لفهم تأثير هذه العوامل على الأداء المؤسسي واتخاذ قرارات استراتيجية لتحسينه.

تمت معالجة البيانات أولاً بالتنظيف والتأكد من تكاملها، بما في ذلك التعامل مع القيم المفقودة وتحويل النصوص إلى تنسيقات قابلة للتحليل. أضيفت ميزة جديدة تُعبر عن مدة العمل، كونها تعكس الخبرة، وأُجريت تحليلات إحصائية لاستكشاف العلاقات بين المتغيرات.

استخدمت تقنيات التجميع مثل K-Means لتحليل البيانات وتصنيفها. باستخدام طريقة الكوع (Elbow Method)، تم تحديد العدد الأمثل للتجمعات بثلاث مجموعات. خضعت النتائج لتقييم شامل عبر مؤشرات مثل Silhouette وDavies-Bouldin لتحسين جودة التجميعات. اظهرت النتائج ضعف أداء نموذج DBSCAN، ولكن كان نموذج K-Means فعّألاً وتم استخدام تقنيات تقليل الأبعاد لتحسين النتائجواستعراض النتائج بطريقة مرئية ومكتوبة .

تم حفظ النموذج المُدرب كملف جاهز للاستخدام، ما يتيح إمكانية تطبيقه مباشرة في أنظمة حقيقية مثل واجهات الويب أو تطبيقات الهاتف المحمول، مما يوفر الوقت والموارد عند تنفيذ المشروع مستقبلًا.

**الراي الشخصي والتوصيات :**

في الحقيقة عند البحث عن الحلول المقترحة لهذه المشكلة هناك بعض التحفّظ على مثل هذه المشاريع لخصوصية الشركات والسياسات ولكن هذا يسبب عائق للطلاب من ناحية الانسجام في المشاريع العملية المفيدة التي تقرّبهم من سوق العمل، لقد اجتهدت في البحث عن مصادر توضح لي كيف يمكنني بناء نموذج تجميع بالتفصيل كانت هذه مهمة صعبة بعض الشيء لاختلاف مشروعي عن المشاريع الموجودة وطبيعة بياناتي ،تتوفر البيانات الاكبر والامكانيات الاعلى للاشخاص الموجودين داخل المؤسسات للبحث والابداع وابتكار الحلول اكثر من الاشخاص الذين هم في مرحلة النمو التعليمي والبحثي مثلي انا شخصيّا ولكنني اعتقد ان ايجاد افكار اكثر ابداعًا وانشاء نماذج مختلفة بافكار شتّى ممكن حتى على مجموعة البيانات البسيطة التي وجدتها وما يشبهها من البيانات ، ان العمل على مثل هذه المشاريع البرمجية سيعطي فرص أكبرعلى الصعيد العملي والعلمي ويفتح مجالات واسعة لمن يحاول الدخول والانطلاق في مجال تحليل البيانات وعلم البيانات بشكل عام .

اوصي بان تكون هناك عناية اكبر بالبيانات التي يتم جمعها عن الموظفين والاعتناء بتوفير هيكل واضح للرواتب مع مراجعات دورية للتأكد من أنها تعكس أداء الموظفين وخبراتهم، حيثأظهرت البيانات أن الرواتب والمكافآت لها تأثير على أداء الفرق، مما يعني أن تحسين الرواتب يمكن أن يسهم في تعزيز الأداء.

**و**ربط المكافآت بأهداف واضحة وقابلة للقياس، مع توزيعها بناءً على أداء الفرق، لتحفيز الإنتاجية وذلك لان الموظفين الذين يتحصلون على مكافآت مناسبة يظهر لديها أداء أعلى نسبيًا.

كما ان بناء نظام متابعة مستمر لأداء الفرق بناءً على العوامل التي تم تحليلها، مثل الرواتب والمكافآت. التجميع يظهر أن الأداء يمكن أن يُصنف بناءً على هذه العوامل، مما يجعل مراقبتها مهمة والتي ستعود على المؤسسة بالنفع واتخاذ القرارات بطرق مدروسة اكثر.

بالاضافة الى ان تصميم خطط لتحفيز الفرق ذات الأداء المنخفض من خلال تحديد أسباب ضعفها وتقديم الدعم اللازم سيؤدي الى الارتقاء بمستوى اداء الموظفين والشركة ككل .

اوصي ايضا باستخدام نماذج مختلفة من نماذج التجميع على نفس مجموعة البيانات لايجاد نماذج اكثر دقة وذات اداء عالي المستوى، ان النتائج التي وصلت لها هي نتائج متوسطة ومتواضعة جدا جدا حيث ان درجة Silhouette المنخفضة تشير إلى وجود فرصة لتحسين جودة النماذج والتوصيات بناءً على تحليلات إضافية. يمكن البحث بشكل اكبر في خوارزميات التجميع المختلفة والتي تعتمد كلّ منها على طريقة مميزة في المعالجة والتعامل مع البيانات .

**المراجع :**

Turban, E., Sharda, R., & Delen, D. (2010). *Decision Support and Business Intelligence Systems*. Pearson Education.

Schutt, R., & O’Neil, C. (2013). Doing Data Science: Straight Talk from the Frontline. O'Reilly Media.

Cao, L. (2017). Data Science: A Comprehensive Overview. ACM Computing Surveys.

Davenport, T. H., & Harris, J. G. (2007). Competing on Analytics: The New Science of Winning. Harvard Business Review Press.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.