جامعة دمشق – كلية الهندسة المعلوماتية (1442هـ – 2021م)

**توقع رضى المسافرين عبر خطوط الطيران باستخدام تقنيات تعلم الآلة**

**محمد عبد العزيز – مصعب مهنا**

**المهندسة زينة الدلال**

**الملخص**

يعتبر التنبؤ برضى المسافرين عبر خطوط الطيران من الأمور الهامة لنجاح أي شركة طيران، كما يعتبر جذب المسافرين وطريقة الحفاظ عليهم جزء أساسي من نظام هذه الشركات، كما أصبح مراقبة أداء المسافرين أحد أهم الأولويات لصناع القرار في شركات القرار، الفهم الشامل لأسباب جنوح المسافرين عن خطوط الطيران يؤدي الى تدخل مناسب لضمان بقائهم. تستخدم العديد من تقنيات التعلم الآلي مثل: التصنيف، الانحدار، التجميع والتنبؤ لمساعدة صناع القرار في انجاز مهامهم، في هذه الدراسة تم تقديم نموذج التنبؤ برضى المسافر بالاعتماد على الميزات التي لها تأثير كبير في رضاه، وقد اعتمدت هذه الدراسة في آلية عملها على بعض خوارزميات التعلم الآلة باستخدام بيانات تم جمعها من خلال استبيانات المسافرين عبر خطوط الطيران الأمريكية، حقق نظام التنبؤ بعد عدد من التجارب العملية وتحليلها ومقارنة عدد من خوارزميات تعلم الآلة دقة أقدر بـ 96%.

**الكلمات المفتاحية**: التنبؤ، التصنيف، التجميع، خطوط الطيران، مسافرين، التعلم التلقائي

**1 - المقدمة**

ان بيانات رحلات الطيران تنموا يوميا وبشكل متزايد تزامنا مع ازدياد اعداد خطوط الطيران وهذا يؤدي الى استدامة البحوث عن تحليلات وتصنيفات لهذه البيانات.

التصنيف هو إجرائية تستخدم العديد من تقنيات التعلم التلقائي التي تكشف الغطاء عن العديد من الحقائق الغامضة، نحن سنقوم ببناء عدة نماذج تعلم تلقائي مألوفة ومقارنة النتائج وعرضها في الجدول [[1](#comparison)]، وكما نعلم أن زيادة حجم مجموعة البيانات التي نقوم بالتدريب عليها ربما يؤدي الى دقة أفضل ولكن عملية التصنيف والتنبؤ على أحجام كبيرة من البيانات تستغرق وقتا طويلا لذلك اخترنا مجموعة البيانات الخاصة بنا بشكل يناسب وقت التدريب ويعطي نتائج جيدة.

**1 -1 هدف البحث**

الهدف من هذه الدراسة ملخص على عدة مراحل:

• معالجة مجموعة البيانات لتصبح جاهزة لعملية النمذجة

• عرض مجموعة البيانات بطرق مختلفة.

• إيجاد النموذج المناسب لتصنيف شعور المسافرين عبر خطوط الطيران بين راضٍ وغير راضٍ (طبيعي).

**1 -2 بنية البحث:**

هذه الحلقة مقسمة كالتالي: القسم 2 يناقش الاعمال السابقة والمرتبطة بالمشكلة التي نعمل على حلها، القسم 3 شرح عن مصادر البيانات وسمات البيانات

القسم 4 تقديم تفاصيل عن البيانات والعلاقات بينها، القسم 5 التنبؤ برضى المسافر باستخدام عدة نماذج تعلم تلقائي مختلفة، القسم 6 اختيار النموذج المناسب.

**2 – الدراسات السابقة:**

إن وفرة البيانات الموجودة نتج عنها العديد من الأعمال المشابهة لهذه الدراسة [[1](#References) ,[2](#References)]

فمثلا في الدارسة [[2](#References)] قام المطور بتنظيف مجموعة البيانات ومعالجتها وتجريب عدة نماذج مثل Logistic Regression وSVC ليحصل على دقة 87.67

واستخدم النموذج SVM-kernel ليحصل على دقة 95.2 على مجموعة بيانات التجريب  
ونجد في الدراسة [[1](file:///C:\Users\Mosab\Desktop\Ml%20Paper.docx#References)] أنا المطور قام باستخدام RandomForestCalssifier ليحصل على دقة 0.93 على مجموعة بيانات التجريب

وبمقارنة نتائج النماذج نجد أن الـ RandomForestCalssifier أدى أفضل نتائج

**2 -1 Dummy Classifier:**

مصنف يقوم بالتنبؤ باستخدام ابسط القواعد، هذا المصنف مفيد باستخدامه كمصنف مبدئي لمقارنة باقي المصنفات معه.

**2 -2 Logistic Regression:**

تقنية الـ Logistic Regression طورت من قبل D.R Cox [[3](#References)] عام 1958 وبعد هذا الاختبار أصبح الـ Logistic Regression جزء من تعلم الآلة.

**2 -3 SVC:**

ما يميز هذا النموذج من التعلم التلقائي هو فعاليته في التعامل مع الأحجام الكبيرة للبيانات كما أنه متعدد الاستخدام حيث يمكن تغيير النواة المستخدمة لإعطاء نتائج مختلفة

**2 -4 Decision Tree Classifier:**

الهدف من أشجار القرار هو بناء نموذج تعلم تلقائي للتصنيف حيث يقوم بالتصنيف عن طريق التعلم من قواعد القرار البسيطة المستنبطة من ميزات البيانات

**2 -5 Gaussian Naive Bayes:**

تقنية تعلم تلقائي تستخدم الاحتمالات كمنهجية للتصنيف، تفترض هذه التقنية استقلالية سمات مجموعة البيانات المستخدمة في التصنيف، في هذا النموذج نوجد احتمال ظهور الصف وفقا لقيم سمات معينة، وبالتالي الصف الذي سيعطي أكبر احتمال سيعرف على انه الصف الهدف

**2 -6 Random Forest:**

هذا النموذج يقوم بإنشاء مجموعة كبيرة من الأشجار العشوائية وهو ما يستغرق وقت طويل ولكنه يعطي دقة في النتائج أكثر من باقي النماذج،

**2 -7 Scaler:**

Minmax: إزاحة البيانات لتستقر بين 0, 1

Standard: كل سمة من سمات البيانات يكون المتوسط 0 والتباين 1, كل السمات يكون لها نفس المطال.

Normalize: يقوم بأسقاط كل عينة على الدائرة نصف قطرها 1 وتستخدم هذه الطرية عندما يكون اتجاه البيانات يهم.

RobustScaller: هذا النوع شبيه ب minmax لكنه يقوم باستخدام العدد الأوسط (median) بدل من المتوسط مما يجعله يتجاهل العينات الشاذة.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| الاسم | Precision | Recall | F1 score | Accuracy |
| Dummy stratified | 0.49 | 0.50 | 0.49 | 0.50 |
| Logistic regression | 0.93 | 0.93 | 0.93 | 0.932 |
| Naïve Bayes Classifier | 0.89 | 0.81 | 0.85 | 0.862 |
| Decision Tree Classifier | 0.95 | 0.93 | 0.94 | 0.94 |
| Random Forest | 0.94 | 0.93 | 0.93 | 0.93 |
| Scaler | 0.95 | 0.92 | 0.93 | 0.94 |
| selectFromModel | 0.94 | 0.92 | 0.93 | 0.94 |

مقارنة نتائج النماذج المستخدمة

**3 – مصادر وسمات البيانات:**

مجموعة البيانات تم جمعها من خلال استبيانات المسافرين عبر خطوط الطيران الأمريكية، حيث تم جمعها على قسمين ومن ثم قمنا بدمج قسمي البيانات لتصبح مجموعة بيانات واحدة ومتوازنة،

تحوي مجموعة البيانات على السمات التالية:

• **الجنس**: جنس المسافر (ذكر، أنثى)

• **نوع المسافر**: أنواع المسافرين مخلص (loyal)، غير مخلص (disloyal)

• ا**لعمر:** العمر الحقيقي للمسافر

• **نوع الرحلة**: نوع الرحلة يكون (رحلة عمل، رحلة شخصية)

• **فئة الحجز**: فئة الحجز على متن الطائرة للمسافر (رجال الأعمال، ايكو، ايكو بلس)

• **مسافة الطيران**: مسافة طيران الرحلة

• **خدمة شبكة الانترنت على متن الطائرة**: درجة رضى المسافر عن هذه الخدمة

• **وقت المغادرة/ الوصول المناسب**: درجة رضى المسافر عن وقت الوصول والمغادرة

• **سهولة الحجز بالأنترنت**: درجة رضى المسافر عن سهولة الحجز عبر الانترنت

• **موقع البوابة**: درجة رضى المسافر عن موقع البوابة

• **الطعام والشراب**: درجة رضى المسافر عن الطعام والشراب

• **الولوج للأنترنت**: درجة رضا الزبون عنها.

• **راحة المجلس**: درجة رضى الزبون عن مكان الجلوس

• **الترفيه على متن الطائرة**: درجة رضى الزبون عن الترفيه

• **الخدمات عند الصعود على متن الطائرة**: درجة رضى المسافر الخدمات عند الصعود على متن الطائرة

• **مساحة الساقين**: درجة رضى المسافرين عن المساحة الممنوحة للحركة

• **حامل الحقائب**: درجة رضى المسافر عن حاملي الحقائب

• **خدمة الاشتراك**: درجة رضى المسافر عن خدمة الاشتراك

• **الخدمة على متن الطائرة**: درجة رضى المسافر عن الخدمة على متن الطائرة

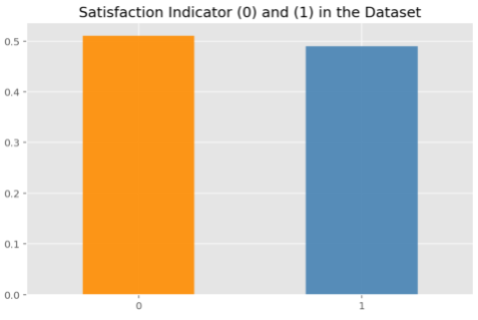
• ا**لنظافة:** درجة رضى المسافر عن النظافة

• **تأخر مغادرة الطائرة بالدقائق**: دقائق تأخر المغادرة

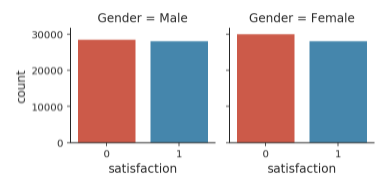
• **تأخر الوصول بالدقائق**: دقائق تأخر الوصول

• **الرضى:** درجة رضى المسافر عن الرحلة

**4 – تفاصيل البيانات والعلاقات بينها:** قمنا في هذا العمل باستعمال مجموعتين منفصلتين من البيانات، قمنا بدمجهما لنحصل على مجموعة واحدة من البيانات ومتوازنة [[3](#balance)] , بعد ذلك قمنا ببعض عمليات التنظيف والمعالجة لتصبح جاهزة للتصنيف، في هذه الصورة توضح عملية الموازنة في مجموعة البيانات راضِ أو غير راضٍ.

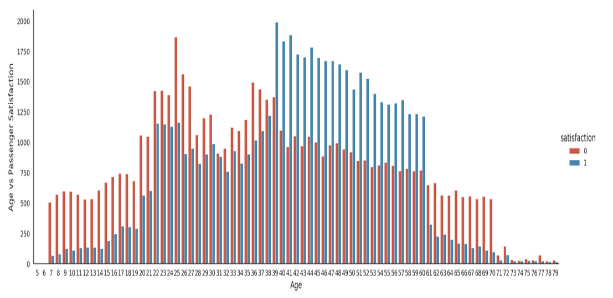


**4-1**- **القليل** **من** **تحليل مجموعة البيانات:**

1. 

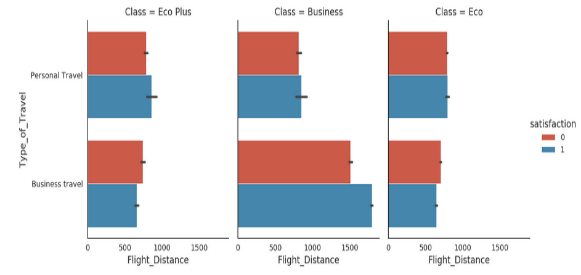
نستنتج أن توزع مجموعة البيانات حسب الجنس هو نفسه تماما بين الصفين.

2.



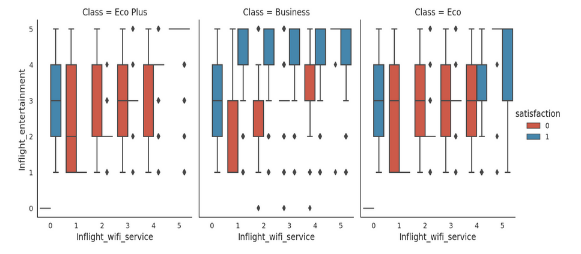
**هنا نستنتج أن من العمر 7 إلى 38 ومن 61 إلى 79 , أن فئة الغير راضٍ أو طبيعي هي الفئة الأعلى، و من العمر 39 إلى 60, إن فئة الراضٍ هي الفئة الأعلى.**

**3.**



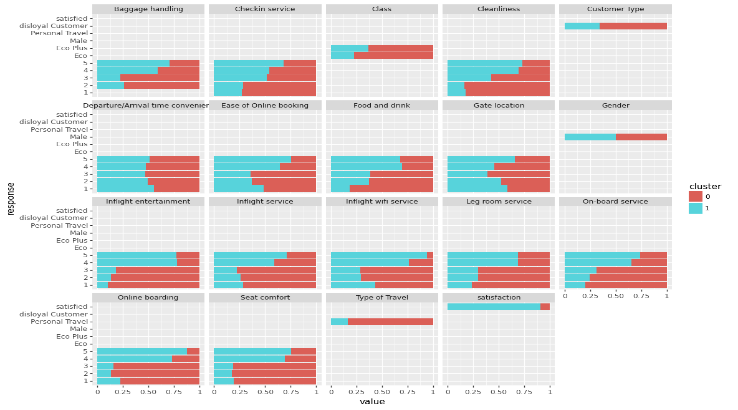
هنا نستنج أن فئة المسافرين ذات الرحلات العمل وأصحاب الحجوزات من الصنف الأعلى على متن الطائرة هم الأكثر رضا على المسافات الطويلة.

**4.**



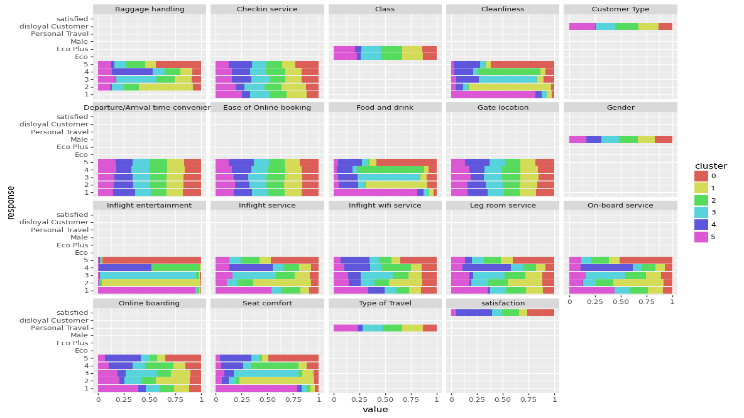
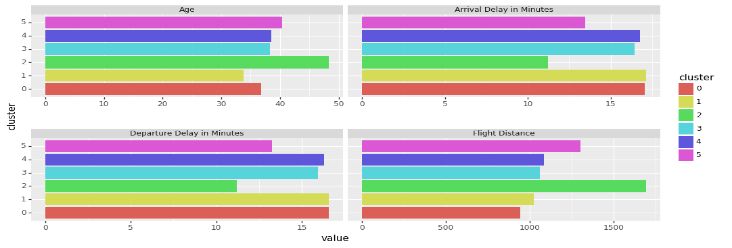
هنا نستنتج أن المسافرون من فئة eco plus هم راضون عن الرحلة من دون خدمة الأنترنت على متن الطائرة والمسافرون من فئة business فقط **أعلى مستوى من التسلية على متن الطائرة يجعلهم راضون، المسافرون من فئة Eco فقط أعلى مستوى من خدمة الأنترنت على متن الطائرة والمستوى العالي من التسلية على متن الطائرة (3-5) يجعلهم راضون.**

**5.**

**باستخدام clustering** 

**من المخطط السابق نستنتج أن الفئة 0 (غير راضية) هي الفئة العليا من التقييمات (rating 1,2) والفئة 1 (راضية) هي الفئة العليا من التقييمات (rating 4,5)**

**6.باستدام clustering وتحديد k=6**

**من المخططات السابقة نستنتج التالي:**

**الفئة 0:** هي فئة راضية وفئة سعيدة جدا لأنها تقيمها عالي جدا بالسمات الثلاثة التالية

food and drinks, Cleanliness, on-board servicing

**الفئة 1:** هي فئة غير سعيدة وأصغر فئة من حيث العمر وأكثر فئة غير سعيدة بالسمة inflight service تقييمها (2 من 5)

**الفئة 2:** هو أكبر مجموعة من حيث العمر الأكبر ومن أكثر الفئات سعيدة بتقييمها للسمات التالية:

Cleanliness, food and drinks تقيمهم لهم (4 من 5)

**الفئة 3:** هذه الفئة فئة متوسطة تقيمها متسوط غالبا" (3 من 5)

**الفئة 4:**هذه الفئة على الرغم أنها من أكثر الفئات تقيمها على أنهم متأخرين لكنهم سعيدين.

**الفئة 5:**هذه الفئة غير سعيدة بأكثر الرحلات تقيمها لأكثر السمات (1 من 5).

**4-2 – معالجة البيانات:**

قمنا بأخذ نسخة من مجموعة البيانات الأصلية وطبقنا عليها التالي:

* حذف الأعمدة الغير مهمة مثل الـ id
* معالجة القيم الشاذة (outlier)
* معالجة القيم الفارغة: قمنا باستبدال القيمة الفارغة في عامود السمة أ بمتوسط قيم عامود السمة هذا بالنسبة للقيم الرقمية أما بالنسبة لقيم الأصناف (category) نستخدم القيمة الأكثر تكرارا" (mode).
* معالجة السمات التي تحوي على أصناف نصية (String categories) باستخدام (one hot encoding) تم الاستبعاد عن (label encoder) لأن السمات تحوي أكثر من صنفين في حالة label سيكون لصنف ما قيمة أعلى من الصنف الأخر.

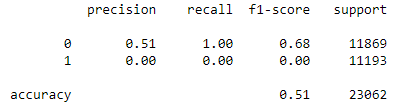
**5 - التنبؤ والتصنيف:**

في هذا القسم سنقوم باستخدام المكتبات الموجودة في البايثون مثل NumPy , scikit-learn ...الخ , ونلاحظ أن مكتبة scikit-learn تعطي نتائج ممتازة في بناء ومقارنة نماذج التصنيف والتنبؤ مثل random forest , Logistic Regression , Gaussian Naïve Bayes

بعد أن قمنا بتحميل مجموعة البيانات ومعالجتها أصبح حجم مجموعة البيانات

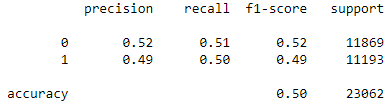
115307 × 23 وبعد عملية encoding أصبح 115307 × 79

**5-1-1 Dummy Classifier:**   
في التجربة الأولى قمنا بأنشاء Dummy Classifier مع استراتيجية most frequent فكانت النتائج كالتالي:

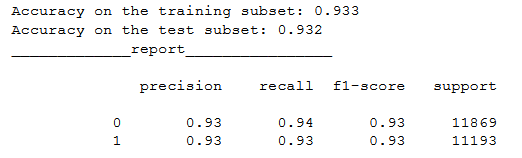


نلاحظ هنا أن قيم ال recall عند الصفر هي 1.00 وعند الواحد (راضٍ) هي 0.00 لأن مجموعة البيانات فيها القليل من الانزياح للصفر (غير راضٍ).

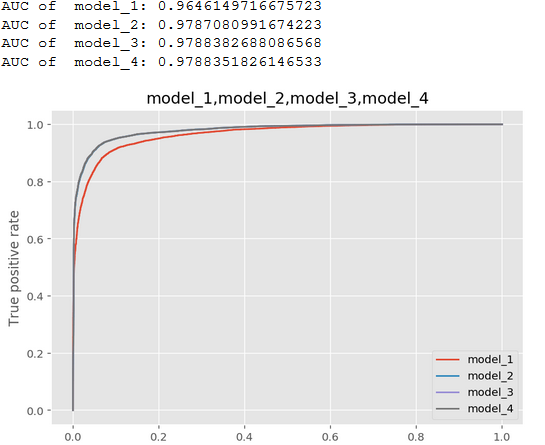
بعدها قمنا باستخدام نفس النموذج لكن مع استراتيجية stratified فكانت النتائج أفضل:



**5-1-2 Logistic Regression:**

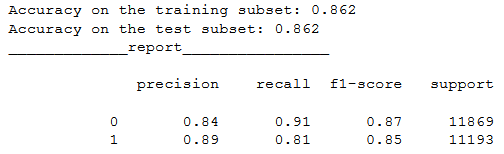
قمنا باختبار هذه النموذج مع أربع قيم دخل للبارامتر C وهي 1000 ,0.001,1,100وأعطى أفضل نتيحه عند القيمة 1000,100: 

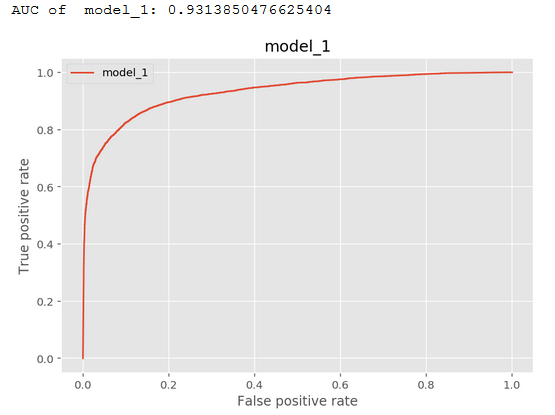
كما هو موضح بالصورة التالية النموذج الأفضل هو رقم 3,4 , الأفضل كلما زادت قيمة True positive وانخفضت قيمة False Positive.

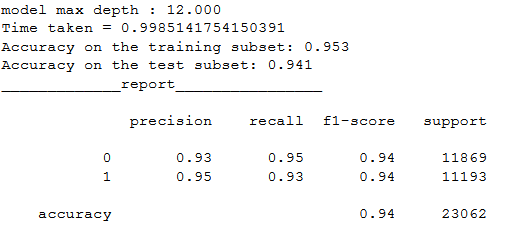


**5-1-3 Naïve Bayes Classifier:**

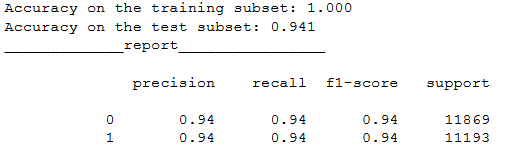
هذا النموذج أعطى النتيجة التالية من ثم نستنتج أن النموذج السابق أفضل من حيث Recall , precision , AUC:



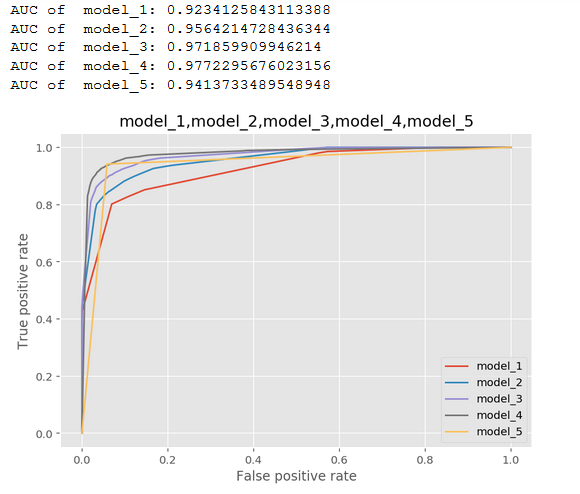
**5-1-4 Decision Tree:**

قمنا باختبار هذا النموذج على ثلاث قيم عمق مختلفة 2,6,8,12 وأعطى أفضل نتيجة عند العمق 12 :   


وعند اختبار الحالة الافتراضية (النموذج الغير مقيد (حصلنا على حالة over fitting تعرفنا على هذه الحالة من خلال الحصول على نسبة 100% على قسم التدريب وأيضا من خلال وجود تباين عالي بين قيم تقيم الموديل على قسم التدريب والاختبار كما هو موضح في الصورة التالية:



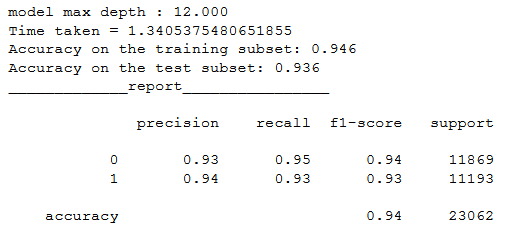
الرسم البياني التالي يوضح قيمة AUC والنموذج الأفضل هو النموذج 4:



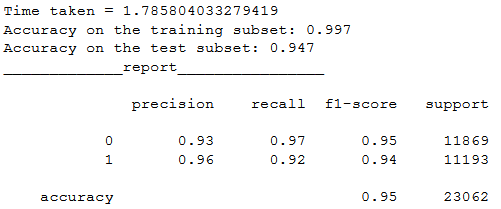
من السابق نستنتج أن النموذج هنا أفضل من Dummy، Logistic Regression, Naive Bayes

**5-1-5 Random Forest:**

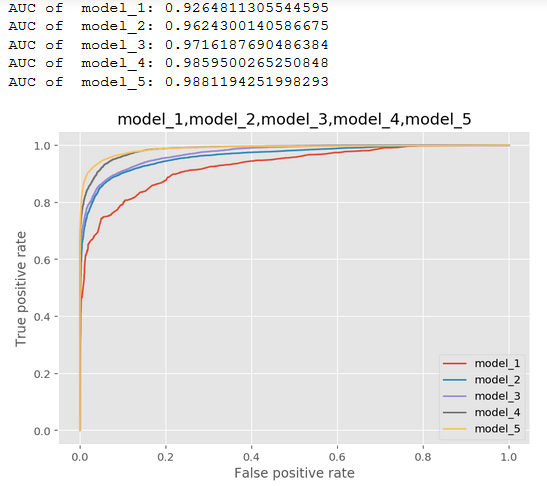
أيضا قمنا باختبار هذا النموذج على ثلاث قيم عمق أعظمي مختلفة ,2,6,8,12 بالإضافة الى القيمة الافتراضية للعمق (النموذج الغير مقيد) حيث قام بإعطاء النطيحة الأفضل عند العمق 12:



وعند اختبار الحالة الافتراضية (النموذج الغير مقيد (حصلنا على حالة over fitting تعرفنا على هذه الحالة من خلال الحصول على نسبة 99.7% على قسم التدريب وأيضا من خلال وجود تباين عالي بين قيم تقيم الموديل على قسم التدريب والاختبار كما هو موضح في الصورة التالية:



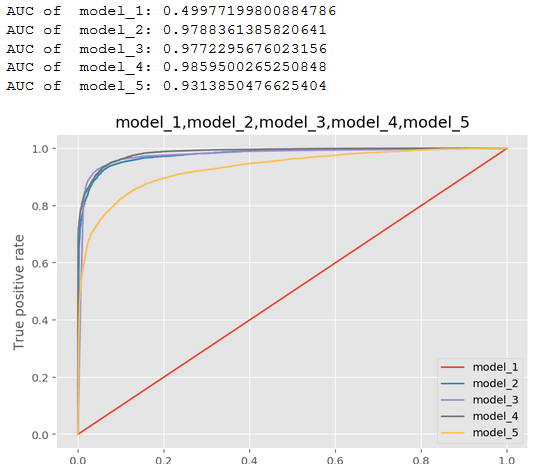
الرسم البياني التالي يوضح قيمة AUC والنموذج الأفضل هو النموذج 4 تم استبعاد النموذج 5 لوجود حالة ال over Fitting:



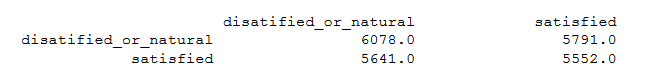
* ***نتيجة*:**

بمقارنة أفضل النماذج بالنسبة لكل نموذج أي أن تمت مقارنة أفضل نموذج من Dummy and Logistic regression and Naive bayes and Decision tree and Random Forest

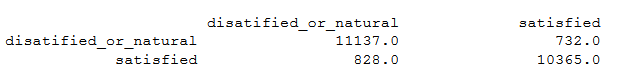
**تم الحصول النموذج الأفضل وهو Decision Tree عند العمق 12.**



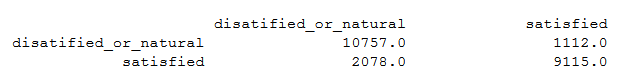
**5-2 نتائج الـ confusion matrix:**

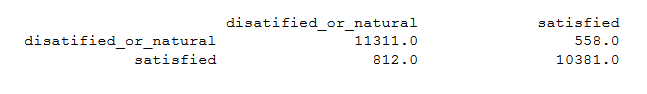


Confusion Matrix of Dummy Classifier

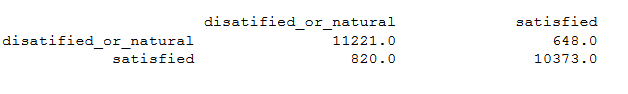


Confusion Matrix of Logistic Regression

Confusion Matrix of Naïve Bayes Classifier



Confusion Matrix of Decision Tree Classifier



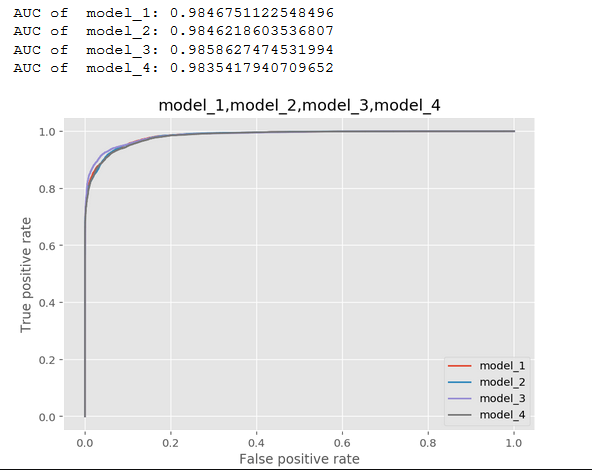
Confusion Matrix of Random Forest

Figure 13. Confusion matrix

**5-3 Scaling:**

قمنا بعمل Scaling بجميع أنواعه المذكورة سابقا لكن لم يتم تحسين أي نموذج

من النماذج السابقة تم توضيح ذلك في اللاب يرجى الاطلاع عليه.



أفضل نموذج هو النموذج رقم 3 وتم استخدام Minmax scaler

تم الحصول على ذات النتيجة تقريبا" في Decision Tree without scaling

حتى الان سيتم الاعتماد على النموذج Decision Tree with out scaling

**5-4 Feature selection:**

**SelectPercentile:**

تحذف كل السمات ماعدا التي يحددها المستخدم النتيجة الأعلى.

SelectKBest:

تحذف الكل لكن السمات k ذات النتيجة الأعلى.

selectFromModel:

هو محول ميتا يمكن استخدامه مع أي مقدر له سمة هامة

بعد التدريب. تعتبر الميزات غير مهمة وتتم إزالتها.

**5-1-4 SelectPercentile:**

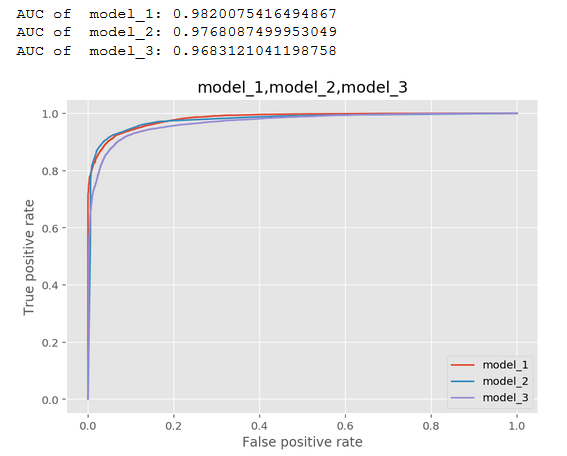
هنا تم استخدام percentile =50 أي تم إزالة 50% من السمات

كان لدينا 78 سمة أصبح 39 سمة وتم تطبيق النماذج الثلاث التالية

Random, logistic regression, decision tree

كما وضح بالشكل التالي وحسب confusion matrixالنموذج الأفضل هو Random forest

عند العمق 12.



النموذج قريب جدا من أفضل نموذج تم أيجاده مسبقا"

**without feature selection**

AUC Decision Tree =97.7

Accuracy on the training subset: 0.956

Accuracy on the test subset: 0.946

**feature selection(percentile)**

AUC Random Forest=98.1

Accuracy on the training subset: 0.945

Accuracy on the test subset: 0.938

حتى الان النموذج الأفضل حسب المعطيات و confusion matrix هو

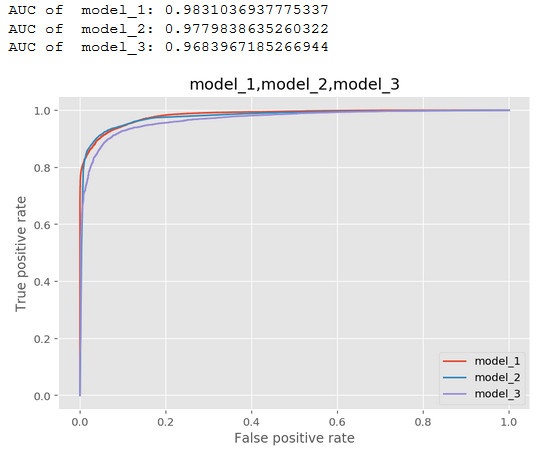
Decision Tree without feature selection or data scaling **5-2-4 selectKBest:**

**عند اختيار k=39 وتم** تطبيق النماذج الثلاث التالية

Random, logistic regression, decision tree

كما وضح بالشكل التالي وحسب confusion matrix النموذج الأفضل هو Random forest

عند العمق 12.



النموذج قريب جدا من أفضل نموذج تم أيجاده مسبقا"

**without feature selection**

AUC Decision Tree =97.7

Accuracy on the training subset: 0.956

Accuracy on the test subset: 0.946

**feature selection(KBest)**

AUC Decision Tree=98.2

Accuracy on the training subset: 0.930

Accuracy on the test subset: 0.933

حتى الان النموذج الأفضل حسب المعطيات و confusion matrix هو

Decision Tree without feature selection or data scaling

أو يمكن اختيار النموذج الجديد كونه ينافسه بشدة وعدد سمات أقل.

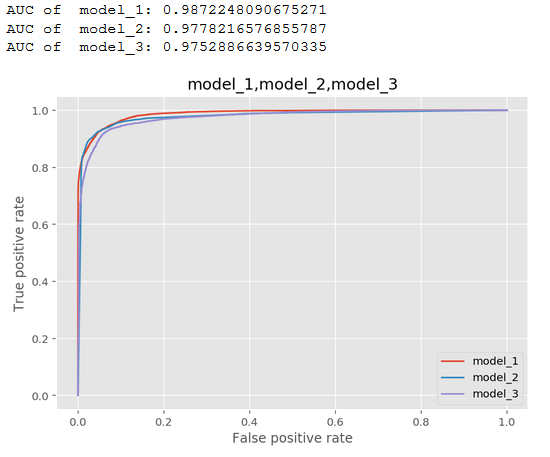
**3-4-5 selectFromModel:**

**عند اختيار threshold=median وتم** تطبيق النماذج الثلاث التالية

Random, logistic regression, decision tree

كما وضح بالشكل التالي وحسب confusion matrixالنموذج الأفضل هو Random Forest

عند العمق 12.



النموذج قريب جدا من أفضل نموذج تم أيجاده مسبقا" ويمكننا الان القول

أنه أفضل من السابق عند أخذ بعين الاعتبار عدد سمات أقل

وحسب النتائج وحسب confusion matrix.

**without feature selection**

AUC Decision Tree =97.7

Accuracy on the training subset: 0.956

Accuracy on the test subset: 0.946

**feature selection(Random Forest)**

AUC Decision Tree=98.7

Accuracy on the training subset: 0.947

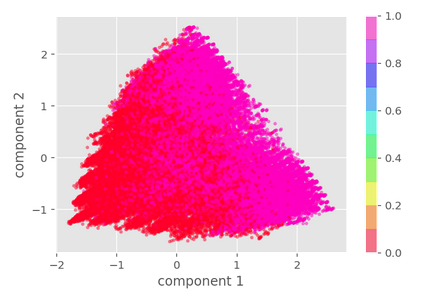
Accuracy on the test subset: 0.938

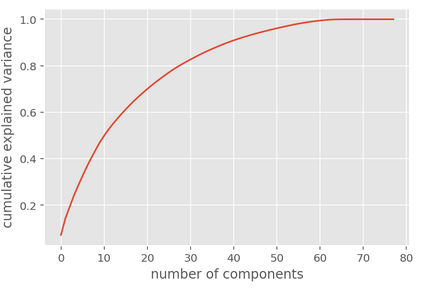
**5-5principal component analysis :**

في تقليص الأبعاد وعرض مجموعة البينات Pca تمت الاستفادة من

عند اختيار n\_component=2 سيتم تقليص مجموعة البيانات إلى

N\_sample\*2



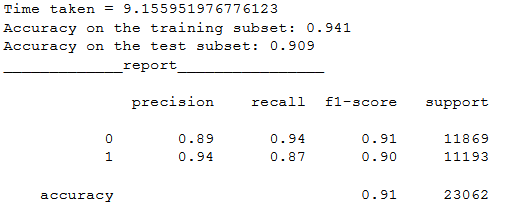
5-5-1 choosing the number of component

هنا يمكننا اختيار عدد السمات 60 وما فوق يكون فيها التباين عالي

يقوم بحذف السمات ذات التباين المنخفض والحفاظ على ذات التباين العالي.

وعند اختيار n\_component=62 وتدريب نموذج Random forest عند

العمق 12 حصلنا على النتائج التالية:



وبمقارنة النتائج مع النموذج السابق الأفضل (Random forest with SelectFromModel)

وعدد السمات أصبحت عند استخدامه n\_sample\*39 نجد أنه أفضل من هذا النموذج pca.

* **المراجع**

[1] <https://www.kaggle.com/codesagnik/airline-passenger-satisfaction-prediction-95>

[2] <https://www.kaggle.com/himanshupadhiar/flight-customer-satisfaction-eda-95-accuracy>