Dédicace

Je dédie ce travail à :

Mes chers parents
Pour tous leurs sacrifices, leurs bienveillance à mon succès, et leur soutien moral.
Pour leur aide, durant tout la période de mes études.
Que ce travail soit la preuve de ma reconnaissance éternelle, amour et respect.

Mes Sœurs
Hajar et Nada, en témoignage de l'affection et de tout l'amour qui nous unit. Pour leur soutient moral et leur aide tout le long de mes études, et qu'elles trouvent dans ce travail l'expression de ma profonde gratitude.

Ma famille et mes amis
Pour leur soutien, leur reconnaissance et leur affection.

Mes encadrants
Pour leur aide et leur accompagnement qui m'ont permis de mener à bien mon projet.

Mes respectables professeurs
Qui m'ont tant formé pour être à la hauteur.

Remerciement

En préambule à ce mémoire, il m'est agréable de m'acquitter d'une dette de reconnaissance auprès de toutes les personnes dont l'intervention au long du projet a favorisé son succès et aboutissement.
Je tiens à exprimer ma profonde gratitude à mon établissement et son corps pédagogique et à mon encadrant pédagogique Mr. Imad HAFIDI, Professeur et chef de département Informatique et Télécoms, pour son suivit, ses conseils et ses recommandations qu'il m'a prodigué tout le long du stage.

Je remercie mon encadrant au sein de *Méditel,* Mr. Hicham JADIR, Senior manager pricing et Business plans au bureau de Planification et pilotage de performance et madame Sahar HASSANINE, expert analyste business planning à la direction Centrale Commerciale pour leur encadrement, disponibilité, conseils et leur accompagnement tout le long de la réalisation de ce travail. Ainsi que tout le corps actif de *Méditel*.
Mes remerciements s'adressent également au corps professoral de l'École Nationale des Sciences Appliquées de Khouribga, en particuliers les enseignants du département génie Informatique d'avoir contribué à ma formation.

Chers membres du jury, c'est un honneur pour moi que vous avez accepté de juger mon travail. Je tiens à vous adresser mon respect et mes sincères remerciements.
Enfin, que toute personne ayant contribué de près ou de loin à la réussite de ce travail trouve ici l'expression de ma reconnaissance.

Résumé

Dans le but d'optimiser et d'augmenter la performance du data mining et du business intelligence au sein de *Méditel,* l'opérateur cherche à passer au traitement des données en utilisant le langage R pour remplacer la solution courante et déjà opérationnelle fournie par SAS. En utilisant R *Méditel* cherche à intégrer une opportunité de développement et d'optimisation de ses procédures qui se reposent sur l'exploitation et l'analyse de données.

Dans ce cadre, s'inscrit ce projet de fin d'étude au sein de *Méditel* ayant pour objectif: Tester les performance de R et savoir son apport vis à vis de SAS à *Méditel* et proposer un scénario d'intégration de cet outil d'analyse.

Pour aboutir à cet objectif, la démarche suivie consiste, au premier temps, à faire une étude sur l'utilisation de SAS et son environnement au sein du département Pricing et Business intelligence pour en tirer le maximum de cas d'utilisation de cet outil. La deuxième étape consiste à se documenter et se familiariser avec le langage R et avoir des notions d'analyse de données volumineuses ainsi que les solutions entreprise de R pour optimiser le temps de calcul et augmenter la productivité du département ainsi proposer une intégration de cette solution. Puis une étude de cas sera donnée comme exemple d'application de R pour générer des modèles prédictifs sur des données de taille large en appliquant plusieurs modèles de prédiction d'un phénomène courant dans l'activité des clients de *Méditel* qui est la migration entre formules pour les clients mobile. Le travail sur ce sujet s'est appuyé sur plusieurs analyses de performance, scalabilité et de productivité. Avec une illustration à travers un modèle prédictif de migration comme étant produit de ce sujet.

Ce modèle prédictif illustrera la simulation de plusieurs méthodes de classification pour la prédiction et les équations de régression logistique sous R avec plusieurs bibliothèques et des testes de fiabilité et d'erreurs lors des testes de prédictions.
En comparaison avec les papiers blancs produits lors d'un modèle de prédiction fait par les consultants SAS pour le phénomène *Churn* (résiliation ou suspension de ligne) chez *Méditel,* la validité et la fiabilité du modèle sera estimée et programmée pour une mise à jour du modèle automatique.

Mots clé :

Business intelligence, data mining, SAS, open source R, Revolution R entreprise, Modèles de prédiction.

Abstract

In order to optimize and increase the performance of data mining and business intelligence in *Méditel*, the provider wants to manage its data processing using R language instead of the actual solution already operational provided by SAS. Using R *Méditel* seeks to integrate an opportunity for development and optimization of its procedures which are based on the exploitation and the analysis of data.
In this context, this project aims at: Testing the performance of R and benchmark it with SAS and suggest an integration scenario for this analysis tool.

To achieve this, the approach consists first on a study about how SAS and its environment are used within the department of Pricing and Business Intelligence to get the maximum use cases. The second step is to document and become familiar with the R language and have large data analysis concepts . Then a case study is given as an R application to generate predictive models on wide size data by applying several prediction models of a common phenomenon in the activity of *Méditel* customers who migrate between formulas and offers. The work on this subject was based on several analyzes of performance, scalability and productivity. With an illustration through a predictive model of migration as product of this.

This predictive model will illustrate the simulation of several classification methods for prediction and logistic regression equations in R with several libraries and tested reliability and error in predictions of testes compared to the white papers produced in a prediction model made by SAS consultants for the phenomenon Churn (termination or line suspension) in *Méditel*, validity and reliability of the model will be estimated and planned for an automatic update of the model.

ملخص

من أجل تحسين وزيادة أداء استخراج البيانات و حساب المؤشرات في ميديتل، تحاول الشركة إستبدال التكنولوجيا المعتمدة في معالجة البيانات بلغة بدلا من التكنولوجيا الحالية التي تقدمها. باستخدام تسعى ميديتل إلى خلق فرصة لتنمية وتعظيم مجال أنشطتها عن طريق تطوير إجراءاتها والتي تقوم على استغلال وتحليل البيانات بتقنية مفتوحة المصدر.في هذا السياق، يهدف هذا المشروع إلى: اختبار أداء وقياس مع ساس ويقترح سيناريو التكامل لهذه الأداة تحليل.

ولتحقيق ذلك، سنطوي أولا على دراسة كيفية استخدام SAS وبيئته داخل قسم التسعيرو تحليل البيانات للحصول على أكبر عدد من سيناريوهات الإستخدام. والخطوة الثانية توثيق والعمل بلغة R لتحليل البيانات الكبيرة. ثم يتم إعطاء دراسة حالة كتطبيق R لتوليد نماذج تنبؤية على حجم البيانات واسعة من خلال تطبيق عدة نماذج التنبؤ لتتبع ظاهرة شائعة في نشاط عملاء ميديتل الذين يبغيرون بين الصيغ والعروض. واستند هذا العمل على عدة تحليلات الأداء والإنتاجية.

النموذج التنبؤي سيمثل محاكاة عدة طرق للتصنيف و معادلات التنبؤ والانحدار اللوجستي في R بالعديد من الطرق واختبار الموثوقية والخطأ في التوقعات كمثال لعملية تحليلية دخل الشركة تمت بواسطة R.

Table des Abréviations

Méditel	Deuxième opérateur de télécommunication au Maroc
SGBD	Systèmes de gestion de base de données
BI	Business intelligence
CRM	Client relationnal management
SAS	Statistical Analysis System
BO	Business Object
DSI	Division des systèmes d'information
MMPR	Montant Moyen Par Recharge
CA	Chiffre d'affaire
IB	Inbound
OB	Outbound
RevoR	Revolution R
KKPV	K Plus proches voisins
SVM	Séparateurs à vaste marge
ROC	Receiver Operating Characteristics
ScalerR	Partie qui gères les lots de données de RevoR
RRO	Revolution R Open
RRE	Revolution R Entreprise
IDE	Integrated Developement Environment
GUI	Graphical User interface
GPL	General Public Licence
RDBMS	Relational database management system
CRAN	Répositoire des bibliothèques R vérifiées
ETL	Extraction, Transformation, Chargement

Sommaire

Liste des Figures

Introduction

La Business Intelligence (BI), également "intelligence d'affaires" ou "informatique décisionnelle", englobe les solutions informatiques apportant une aide à la décision avec, en

bout de chaîne, rapports et tableaux de bord de suivi à la fois analytiques et prospectifs. Le but est de consolider les informations disponibles au sein des bases de données de votre entreprise. Toutes les activités de celle-ci sont concernées par les tableaux de bord et en sont des utilisateurs potentiels. La BI représente pour vous une opportunité d'optimiser le pilotage de vos activités et d'anticiper les évolutions marché ou du comportement des clients.
Pour pouvoir déployer les données de telle façon à y appliquer de la business intelligence, la discipline du data mining prend place comme étant un nouveau champ situé au croisement de la statistique et des technologies de l'information (bases de données, intelligence artificielle, apprentissage, prédiction etc.) dont le but est de découvrir les structures dans de vastes ensembles de données et d'en tirer des modèles ou des éléments d'aide à la décision.
La métaphore du data mining signifie qu'il y a des trésors ou pépites d'or cachés spis des montagnes de données que l'on peut découvrir avec des outils spécialisés. Ainsi l'analyse des données recueillies à d'autres fins – c'est une analyse secondaire on peut dire de bases de données – souvent conçues pour la gestion de données individuelles.
La notion de Business Intelligence est apparue à la fin des années 1970 avec les premiers data-centers. Des systèmes qui envoyaient des requêtes directement sur les serveurs de production, ce qui se révélait plutôt dangereux pour ces derniers. Dans les années 1980, l'arrivée des bases relationnelles et du mode client/serveur a permis d'isoler l'informatique de production des dispositifs décisionnels. Dans la foulée, des acteurs spécialisés se sont lancés dans la définition de couches d'analyse « métier », dans le but de masquer la complexité des structures de données. Depuis, la BI n'est plus l'apanage des équipes techniques, elle est directement accessible aux responsables opérationnels. Tant qu'au data mining, il est né de l'évolution des SGBD vers l'informatique décisionnelle avec les entrepôts de données (Data Warehouse), de l'explosion phénoménale de la taille des données, du développement de la gestion de la relation client (marketing client au lieu de marketing produit) et les recherches en intelligence artificielle, apprentissage et extraction de connaissances.

Les outils de data mining destinés entreprise doivent répondre à leurs utilités d'un fiabilité et productivité extrêmes. Utiliser un outil open source présentera une porte ouverte pour développer des solutions dédiée à des cas spéciaux et des configurations illimités depuis l'outil lui même. Présentera des solutions à moindre coût vis à vis la formation et la maîtrise de l'outil. C'est le cas de R, un outil de statistique à la base développé ensuite pour le data mining pour être un langage vectoriel en évolution pour traiter les même les bases de données big data.
C'est dans ce cadre que ce projet est proposé pour intégrer R dans le processus du data mining à Méditel pour remplacer les solutions existante. Pour faire cela plusieurs étapes de benchmarking validerons la façon dont cet outil sera intégré. Un exemple de prédiction sera évoqué comme scénario d'utilisation de R Analysis.
La distribution du parc des lignes mobiles est divisé en deux formules d'offres, des offres prépayées et autres post-payées. Un client peut prendre la décision de changer sa formule à tout temps à cause d'une raison qui peut être déduite de son comportement avant la migration. L'importance de ce phénomène impacte en gros le chiffre d'affaire déduit des deux populations du parc. Prédire ce phénomène pourra donner une vision futur de la structuration du parc, le CRM pourra exploiter ces données prédis pour cibler les lignes pour leurs offres futurs dans la stratégie marketing intelligente.

Le présent rapport synthétise tout le travail effectué dans cette perspective. Il est organisé comme suit :

Partie A : < Présentation d'états des lieux>

Dans ce chapitre, l'organisme d'accueil Méditélecom sera présenté avec quelques détails du secteur d'activité ainsi que l'état des lieux et de la méthodologie du fonctionnement du service Business intelligence au sein de l'opérateur.

Partie B : <Data mining et reporting chez Méditel>

Ce chapitre a pour but d'extraire les différents cas d'utilisation d'utilisation de SAS Entreprise Guide et les différentes opérations auxquelles son soumises les données de l'opérateur.

Partie C : <Comparaison entre les solutions Open source de mining>

Ce chapitre a pour but de justifier le choix de R comme étant la solution Open Source proposée pour la migration de technologie du service pour faire du data mining en le comparant avec trois autres outils open source de data mining.

Partie D : <Intégration de R chez *Méditel*>

Pour bénéficier d'une transition fluide lors du changement de technologie, une étude des possibilité d'intégrations de R sous différentes implémentations et d'en choisir celle qui sera la plus optimale pour le service BI de *Méditel* comparé à l'outil SAS.

Partie E : <Problématique de classification sous R>

La partie C présente un cas d'application de R pour traiter les données de Méditel et en tirer un modèle de prédiction basé sur les méthodes de classification sous R. Le contexte de l'application est de prédire la migration des lignes prépayées vers le post-payé.

Partie F : <Conclusions et discussions>

Ce rapport est clôturé par ce chapitre qui décrit les ajouts que le projet a pu apporter ainsi que les perspectives à envisager pour améliorer la performance du service.

Partie A : < Présentation d'états des lieux>

Dans ce chapitre, l'organisme d'accueil Méditélecom sera présenté avec quelques détails du secteur d'activité ainsi que l'état des lieux et de la méthodologie du fonctionnement du service Business intelligence au sein de l'opérateur.

A- Présentation d'état des lieux

A.1- Présentation de l'entreprise d'accueil

Méditel, anciennement Méditelecom est une entreprise marocaine de télécommunications. Créée en 1999, elle est le deuxième opérateur de téléphonie mobile au Maroc. L'entreprise emploie plus de 2 000 salariés et génère plus de 20 000 emplois indirects. Son chiffre d'affaires s'est élevé à 5,3 milliards de dirham.

A.1.1- Historique

Meditelecom fut créée en 1999 à la suite d'un partenariat entre des investisseurs marocains et les groupes Telefonica et Portugal Telecom qui en détenaient 32,18 % chacun. Actuellement Méditel est devenue en quelques années une entreprise marocaine de référence. Retour sur quelques dates clés d'une histoire riche en événements :

2000
- Arrivée de Méditel permettant une dé-monopolisation du marché du Télecoms

2001
- Mise en place des services révolutionnaires (factures et forfaits plafonnées,...)

2002
- Déjà 700 collaborateurs ont rejoint la grande famille Méditel

2003
- 89% de couverture atteinte avec 1616 BTS en service

2004
- Près de 3 000 000 de clients, une évolution de +42% par rapport à 2003

2005
- 1er exercice excédentaire après 5 ans d'existence

2006
- Méditel restructure sa dette. L'opération est la première du genre en finance.

2007
- Méditel dépasse le seuil de 6 millions de clients et lance la 3G

2008
- Méditel lance un programme d'investissement de 4,2 milliard de dirham

2009
- 10 ans après sa création, montée en puissance des groupes marocains actionnaires

2010
- Méditel franchit le cap des 10 millions de clients

2011
- Conclusion de partenariat avec le groupe France Telecom

2013
- Méditel adopte la signature <dima rahtek>

2014
- Diversification du parc mobile avec des forfait sans engagement

2015
- Méditel est le premier opérateur à lancer la technologie 4G

2016
- Méditel lance Hany net, l'offre la plus généreuse en volume de navigation.

Fig 1 : Historique de l'opérateur

A.1.2- Chiffres clés

Jour après jour, Méditel mobilise plus de 2 000 collaborateurs et un réseau près de 20.000 ambassadeurs dans plus de 13.000 poins de contact à travers le Royaume.

Fig 2 : Chiffres clefs de l'opérateur

A.1.3- Valeurs

Cette réussite ce construit, également, sur la base des engagements et valeurs que Méditel tient vis-à-vis de son environnement. Méditel, c'est des ambitions, de l'investissement, du savoir-faire, mais surtout de valeurs fortes :

- **Simplicité**

Méditel est simple parce qu'elle place la technologie à la portée de tous, parce qu'elle utilise un langage direct et propose à ses clients des offres claires et des services faciles à utiliser.

- **Transparence**

Méditel est transparente parce qu'elle privilège l'échange dans sa relation avec ses partenaires et ses clients et adopte un discours en adéquation avec les offres qu'elle propose, sans mauvaise surprise.

- **Générosité**

Méditel est généreuse parce qu'elle conçoit des offres et des services accessibles à tous, avec le souci constant de satisfaire ses clients.

- **Proximité**

Méditel est proche parce qu'elle est toujours à l'écoute de ses clients et attentive à répondre à chacun de leurs besoins.

- **Attention portée au client**

Méditel est attentionnée parce qu'elle facilite la vie de chacun et agit chaque jour pour le confort de tout.

A.2- Actionnariat et Conseil d'administration

A sa création en 1999, Méditel est le fruit d'une alliance entre les majors des télécoms en méditerranée et de solides groupes financiers et industriels marocains.
En décembre 2010, le groupe France Telecom (Orange) a signé son entrée définitive dans le capital Méditel.

A.2.1- Finance Com

Groupe marocain privé aux ambitions régionales et internationales affirmées, Finance Com est présent notamment dans les services financiers (Groupe BMCE bank), les assurances (RMA Wataniya) et les télécommunications (Méditelecom).

A.2.2- La caisse des dépôts et de gestion

Le groupe CDG est un groupe financier dédié au développement économique et social du Maroc. Avec plus de 5 000 collaborateurs et une quarantaine de filiales métiers, le groupe CDG est un acteur de référence dans la gestion de fonds institutionnels, les métiers bancaires et financiers et le développement territorial.

A.2.3- Orange, Groupe France Telecom

Orange est la marque phare de France Telécom, un des principaux opérateurs européens du mobile et l'accès internet ADSL et l'un des leaders mondiaux des services de télécommunications aux entreprises multinationales, sous la marque Orange Business Services.

A.3- Organigramme

Méditel organise son travail d'une façon claire affin d'assurer une efficacité optimale. Ainsi, six directions centrales structurent l'entreprise ; il s'agit d'un pole financier, de la direction centrale service, de direction centrale et ressource, de la direction centrale commerciale, de la direction centrale entreprise et la direction centrale réseau.

Fig 3 : Organigramme de l'entreprise

- La direction Centrale Finance et Ressources

Elle intègre toutes les activités liées à la finance, nécessaires pour une bonne gestion de l'entreprise et à l'origine de décisions stratégiques en collaboration avec la direction commerciale.

- La direction Centrale Service

Elle revoit, quand à elle, au savoir-faire technologique de Méditel et intègre la dimension qualité des produits et services par la mise en œuvre d'un système d'ingénieure des systèmes d'information de haut niveau

- La direction Centrale Commerciale

Elle est répartie en unités d'affaires, chacune renvoyant à une catégorie de produits et services offerts par l'entreprise qui travaille en collaboration avec des unités de support. C'est à ce niveau que se dessine la stratégie Marketing adoptée par Méditel et qui témoigne de

l'expertise et de l'expérience dont est munie l'entreprise. Mon stage s'est déroulé dans cette entité.

- La direction Centrale Entreprise

Elle gère un segment potentiel et particulier des clients Méditel à savoir le segment des entreprises.

- La direction Centrale Réseau

Elle intègre toutes les activités liées au déploiement du réseau, nécessaires pour une couverture optimale du territoire ainsi que l'acheminement des appels téléphoniques. Cette organisation a été mise ne place pour assurer :

- Une meilleure efficacité opérationnelle.
- Un positionnement des métiers dans des directions centrales cohérent et homogène.
- Une amélioration du processus de prise de décision favorisant un niveau de délégation plus important et un renforcement et la culture de la performance.
-

A.4- Présentation du service d'accueil

Avant de présenter le service Business Intelligence, le service ou le stage a été effectué, il serait nécessaire de connaître la structure de la direction centrale commerciale et les différentes unités la composant.

A.4.1- Département Pricing & Business Intelligence

D'une façon générale, le département Pricing et Business Intelligence fournit à l'entreprise des informations à caractères économique, financier, stratégique et politique pour permettre le développement de stratégies d'entreprises pertinentes.

Savoir-faire méthodologique :

- Repérer, analyser et synthétiser des informations (rendre claire et simple une matière complexe)
- Élaboration des business Case (deuxième étape du parcours de produit).
- Mesure et suivi des revenus des nouveaux produits.

A.4.2- Service Business Intelligence

La politique suivie par l'opérateur Télécom Méditel met le service business intelligence à la disposition des autres services concernés par ses analyses et ses extractions des bases, les autres services se bénéficient de ces analyses et extractions pour prendre des décisions.

Fig 4 : Organigramme de la direction commerciale

A.4.3- État des lieux

Le service Business Intelligence utilise des outils informatiques pour traiter les différentes données de l'entreprise et les analyser, mais cela était fait d'une manière manuelle à travers SAS EG et SAP Business Objects qu'on utilise pour créer les différents rapports dépendamment du besoin de l'entreprise.

Changer l'outil de traitement et d'analyse de données vers un outil de source libre créera un champs très vaste de développement d'outils et de procédures personnalisées au BI et au data mining à Méditel.

Pour cela et dans le cadre de mon stage, on m'a confié la tache de benchmarker deux solutions qui servent d'outils d'analyse et traitement de données, SAS et R pour préparer le champs à la migration totale des procédures du service vers la nouvelle technologie open source.

A.4.4- Méthodologie et outils du Service Business Intelligence

A.4.4.1- Outils du service BI

Le service BI utilise un ensemble d'outils informatiques pour extraire les données, traiter, analyser et présenter ; on cite ici :

- **Business Objects :** ce logiciel est utilisé lorsque les données nécessaires à la demande sont très récentes (analyse sur les derniers jours). Une description du logiciel est disponible à l'annexe.

- **SAS Enterprise Guide :** ce logiciel est utilisé lorsque les données nécessaires à la demande ne sont pas très récentes (analyse sur un ou plusieurs mois), il est fréquemment utilisé que BO puisqu'il permet des bases de grande taille. Une description du logiciel est disponible à l'annexe.

- **Excel, PowerPoint et Think-Cell :** Après extractions et analyses assurées par les deux derniers logiciels, les deux logiciels Excel et PowerPoint sont utilisés pour présenter les données et restituer les évolutions des différents indicateurs et paramètres, ils permettent d'assurer la standardisation des fichiers circulant entre directions ; ThinkCell est un add-on à installer sur office. Il a pour but de faire le lien entre Excel, blindé de données, de formules et tableaux divers et un PowerPoint, qui est la synthèse graphique avec juste quelques explications et quelques graphiques pour nos décideurs. Une description de ces outils est disponible à l'annexe.

A.4.4.2- Processus des demandes adressés au services

Le service reçoit chaque jour des demandes de la part des différentes directions de l'entreprise, ces demandes sont diversifiées, citons quelques exemples :

- Suivi des KPIs commerciaux d'usage et de revenu.
- Élaboration des rapports concernant la performance de l'entreprise.
- Réalisation des analyses prédictives.
- Extraction des bases de clients suivant les critères liés au besoin.
- Analyse avant et après les actions commerciales.
- Suivi journalier, hebdomadaire et mensuel des activités des entreprises.
- Élaboration des business Case (deuxième étape du parcours de produit).

Après réception d'une demande, le service Business Intelligence la traite en quatre étapes avant la livraison de l'Output :

Fig 5 : Processus de traitement des demandes

- **Analyse du besoin :** le besoin doit être analysé pour déterminer le degré d'importance du besoin et les informations et données répondant au besoin.
- **Extraction des données :** dans cette phase, l'opérateur commence à extraire les données en utilisant les outils BI cités ci-dessus ; pour les données récentes (avant actualisation des DATAMART), le service utilise le logiciel Business Objects pour accéder directement au data warehouse.
- **Traitement et analyse :** dans la plupart des cas, le besoin consiste en une analyse explicative ou prédictive, ce qui conduit le service à utiliser les techniques Data Mining pour faire ces analyses.
- **Présentation des résultats :** Avant de livrer les résultats au service concerné, ils seront restitués et présentés dans des fichiers Excel ou sur PowerPoint avec des graphes et commentaires explicatifs.

A.5- Présentation du projet

A.5.1- Problématique

Ce projet s'inscrit dans le cadre d'une étude post migration de l'outil de traitement et d'analyse de données, permettant d'intégrer R comme solution open source au lieu de SAS au sein du département Pricing et business intelligence.

La solution actuelle de traitement et de data mining SAS réside un outils professionnel et un progiciel de renom, mais penser à utiliser un outil de source libre permettra d'élargir le champs de fonctionnalités et rendre l'outil plus configurable aux analyses faites à Méditel.
Le département cherche à analyser le travail sous R pour traiter tous les cas d'utilisation auquel le département Pricing et Business intelligence aura à rencontrer dans son travail de

data mining et reporting pour aboutir à une idée globale de la couverture de R comme outil au sein du services.

C'est dans ce cadre que Méditel a confié la mission de comparer les deux solutions, de trouver un scénario d'intégration de l'outil R et de créer les modèles de prédiction d'un phénomène courant dans son parc de lignes actives comme exemple d'un cas d'utilisation orienté data mining. Le phénomène en question est la migration des lignes d'une offre Prépayée dans laquelle le client recharge son numéro avant de commencer à utiliser et consommer son solde en appels, SMS et/ou en data, vers une offre post-payée qui inclus les formules d'abonnement classique et les forfait nouveaux. Ce module en question permettra d'anticiper cette migration pour cibler les potentiels migrants dans la publicité et estimer l'évolution du chiffre d'affaire après cette migration.

A.5.2- Objectif du projet

Dans le cadre de la vision du service Data mining du département Pricing & Business Intelligence, l'opérateur Télécom Méditel souhaite intégrer un outil de data mining open source pour remplacer la solution SAS Entreprise Guide pour faire le traitement et l'analyse des données. Le service cherche à savoir l'apport en valeur ajoutée de l'outil ainsi que savoir la qualité de ses performance vis à vis le travail quotidien fait au service sur de grandes tailles de données. Ensuite illustrer le travail sous R, en modélisant la prédiction des migrations prépayé vers post-payé du parc mobile de Méditel.
Les grands points touchés par ce projet :

- Comparer R aux autres outils open source de data mining,
- Comparer R et SAS,
- Intégrer R au sein de Méditel,
- Déployer R pour créer un modèle prédictif.

A.5.3- Conduite du projet

A.5.3.1- Cycle de vie d'un projet de migration de technologie

La réalisation de tout projet décisionnel qui porte sur une migration de technologie commence par la phase d'analyse et comparaison des sujets de cette migration, Ainsi que les deux outils R et SAS serons mis sous microscope.
La phase d'analyse débutera par extraire les cas d'utilisation de SAS au niveau du département et faire le suivi de quelques scénarios dont le service en aura besoin. Ensuite chercher ces fonctionnalités sous R et la limite en productivité de ses fonctionnalités dans un milieu entreprise. Et pour avoir un bilan complet de la comparaison, deux benchmarks technique et de performance ont fait sujet de recherche pour tester le temps de réponse de chaque outil ainsi qu'un cas d'utilisation en workflow pour concrétiser, sa limite de traitement et les ressources demandées. Enfin, pour finaliser l'analyse avant migration il va falloir choisir

la méthode d'intégration la plus optimale, optimale dans la définition d'une facilité d'utilisation et de fluidité de transaction question ressources techniques et humaines pour garder ou augmenter la taux de productivité du département.

A.5.3.2- Implémentation du cycle dans ce projet

Selon le cycle de vie décrit dans la partie précédente, la planification est une étape préliminaire qui s'impose avant d'entamer tout projet décisionnel. Elle consiste à prévoir le déroulement du projet tout au long des phases du cycle.

Pour amener le projet à bonne fin, ce diagramme de GANTT décrit les prévisions concernant les différentes phases du projet :

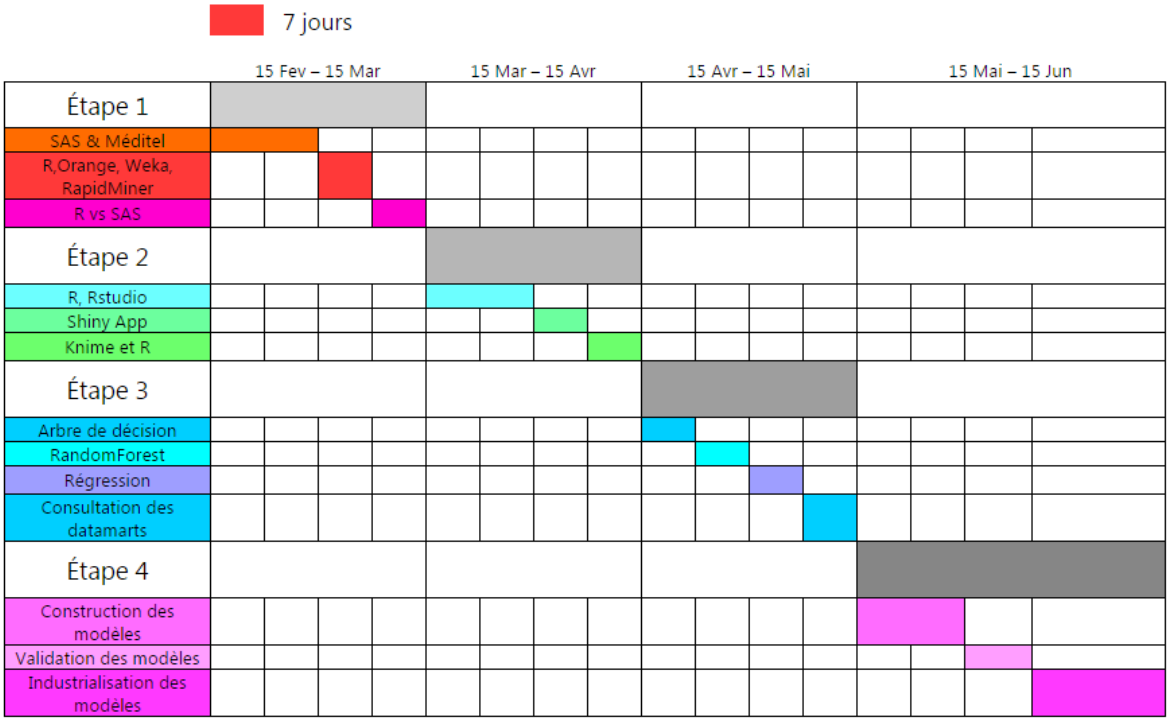


Fig 6 : Diagramme de GANTT

Partie B : <Data mining et reporting chez Méditel>

Ce chapitre a pour but d'extraire les différents cas d'utilisation d'utilisation de SAS Entreprise Guide et les différentes opérations auxquelles son soumises les données de l'opérateur.

B- Data mining et reporting chez Méditel

B.1- Architecture d'implémentation des outils

Le Division des Systèmes d'Information a mis en place la plate-forme sur laquelle se base le département Pricing et Business Intelligence pour subvenir à ses besoins en Data mining ainsi qu'au reporting et garantir la persistance et la cohérence des outils avec les entrepôts de données, La DSI collectionne les données de plusieurs sources émergentes de l'activité des clients selon plusieurs perspectives (dimensions) architecturés en étoile arrivant des différents départements (Géolocalisation, Ventes, CRM, …) pour les stocker dans le data warehouse et les datamarts conçus chacun pour une utilisation précise (orientés reporting, orientés data mining, orientés extraction …) ainsi qu'installer la partie de traitement SAS server qui rendra l'utilisation de l'outil disponible via SAS Entreprise Guide sous forme de client, ou sur SAP Business Object.

La structure d'installation est la suivante :

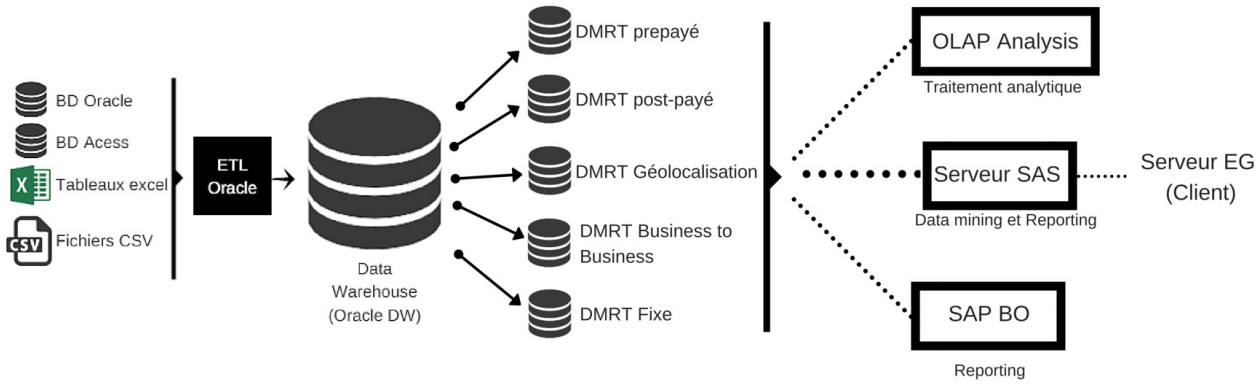


Fig 7 : Architecture des services BI

B.2- Cas d'utilisation

L'utilisation des deux outils SAS et BO est choisie chez *Méditel* selon l'ancienneté des données à traiter ça reste une convention chez le département de traiter les données récentes ou journalières via Business Object et laisser le travail sur les datamarts mensuels et demi annuels et le travail majeur du data mining sous SAS.

B.2.1- Service Data mining

Les opérations majeurs de data mining se font sous SAS pour traiter des modèles prédictifs de plusieurs phénomènes concernant le parc mobile, fixe et business to business ainsi que le travail statistique classique sur l'ensemble des données pour aboutir à réaliser les besoins suivants :

- Réalisation des analyses prédictives.

- Extraction des bases de clients suivant les critères liés au besoin.Analyse avant et après les actions commerciales.
- Suivi des KPIs commerciaux d'usage et de revenu.

Les cas d'utilisation de l'outil SAS pour le datamining peuvent être constitué en combinant les fonctions élémentaires suivantes :

Importation : Importer les datamarts soit avec une connexion RDBMS, avec des fichiers texte plat .csv ou des tableaux Excel de taille importante.

Jointures : De tables verticalement ou horizontalement. Soit en intersection, jointure droite, gauche ou union.

Colonnes calculés : Création de colonnes calculées pour désigner des variables présentant des mesures de performance ou des valeurs significatifs et plus représentatifs.

Requêtes : Requêtes sur les données permettant la sélection, le tri, le filtrage … etc.

***Statistiques* :** Statistiques classiques sur les variables des données.

Échantillonnage : Création de tout types d'échantillons pour réduire la taille des données à traiter dans l'étape suivante d'un procédé.

Machine learning : Création de modèles prédictifs sur différents phénomènes suivant plusieurs méthodes (Classification, Régression, Réseaux de neurones…)

Le produit final du data mining sert à aider à aider à la décision en offrant les données et les information concrètes sur lesquelles se baser.

B.2.2- Service Business Intelligence

Le service BI de Méditel assure le soutient des projets décisifs de l'entreprise en assurant les taches suivantes :

- Élaboration des business Case (deuxième étape du parcours de produit).
- Élaboration des rapports concernant la performance de l'entreprise.
- Suivi des KPIs commerciaux d'usage et de revenu.
- Suivi journalier, hebdomadaire et mensuel des activités des entreprises.

B.2.2.1- Les indicateurs clés de décision

Le reporting en réalité présente les éléments finaux d'aide à la décision pour les différentes taches faites par le département. L'exemple des KPIs réside l'un des majeurs applications du reporting car les indicateurs clés de performance (ICP), appelés le plus souvent KPI ("Key Performance Indicator"), sont des indicateurs d'aide à la décision dont le but est de générer des rapports détaillés sur l'évolution des facteurs clés de succès des activités d'une entreprise. Leur principale utilité consiste donc à évaluer les performances des actions qui ont été mises en place en fonction des objectifs définis.

Ils répondent au besoin de présenter des données techniques dans un langage compréhensible par tous les interlocuteurs. Un KPI est une information ou un ensemble d'informations permettant et facilitant l'appréciation, par un décideur, d'une situation donnée. C'est une mesure d'un aspect critique de la performance globale du projet.

On distingue généralement plusieurs types de KPIs :

- Les **indicateurs d'alerte**, définissant un seuil critique à ne pas franchir. Ils soulignent le niveau de normalité d'un système, d'un processus ou d'un projet.
- Les **indicateurs d'équilibre** : ils permettent de vérifier l'adéquation entre l'objet d'étude (système, processus ou projet) et l'objectif attendu. Ils informent sur l'état du système et mettent en évidence les dérives potentielles.
- Les **indicateurs d'anticipation** : ils permettent de mettre en exergue les éléments de prospective liés à un système, un processus ou un projet. Ce sont des indicateurs de tendance.
- Un **indicateur de performance** est un instrument lié à un critère d'évaluation de la performance. Il est la mesure d'un objectif à atteindre, d'une ressource mobilisée, d'une réalisation, ou d'un effet obtenu.

Et comme dans tous les domaines, le domaine Télécom a ses KPIs qui permettent de planifier et piloter la performance de l'opérateur. On va diviser les KPIs en deux familles :

KPIs de l'Utilisation :

Les indicateurs de l'utilisation concernent la performance de l'opérateur Télécom en fonction de l'utilisation des clients en termes de trafic voix, ces indicateurs permet d'avoir des idées sur les durées des appels, le nombre de clients qui font des appels, les augmentations et les baisses de ces durées et beaucoup d'autres choses, nous citons ci dessous les KPIs d'utilisation qui vont s'avérer utiles lors de la dernière partie du rapport lors de la construction des modèles :

- **Nombre total des appels :** Comme son nom l'indique, cet indicateur montre le nombre total des appels émis par les clients de l'opérateur quel que soit l'opérateur de la ligne destinataire.
- **Nombre des appelants :** cet indicateur montre le nombre total des appelants du mois, ce sont les clients de l'opérateur qui ont fait au moins une durée d'appel sortant

supérieure à zéro. C'est une information extraite en sommant le nombre des lignes distinctes qui ont effectué au moins un appel.

- **Fréquence des appels :** l'indicateur fréquence des appels montre une moyenne de nombre de fois qu'un client appelant effectue un appel (nombre de reproduction de l'acte appeler chez un client) ; il est calculé par le ratio de nombre d'appelants par le nombre des appels.
- **Durée totale des appels :** Comme son nom l'indique, cet indicateur montre la durée totale des appels émis par les clients de l'opérateur quel que soit l'opérateur de la ligne destinataire. C'est une information extraite en sommant les durées des appels des lignes clients.
- **MOU :** Minute of Use ou minutes utilisées par appelant (MUPA): Cet indicateur montre la moyenne d'utilisation totale en minutes par client de l'opérateur. On peut distinguer les clients prépayés et les clients post-payés. Il est obtenu par le ratio de la durée totale des appels par le nombre des appelants.
- **ACD :** Average Call Duration ou durée moyenne par appel (DMPA): Cet indicateur montre la durée moyenne mensuelle des appels effectués par un client quelconque. On peut aussi distinguer entre les types de clients. En comparant avec le mois précédent, on conclut si les clients utilisent plus ou moins leur carte … il est obtenu par le ratio de la durée totale des appels par le nombre des appels.
- **Durée Off-net :** cet indicateur montre la durée totale des appels émis par les clients de l'opérateur vers les clients d'un autre opérateur.
- **Durée internationale :** cet indicateur montre la durée totale des appels émis par les clients de l'opérateur vers les lignes internationales.
- **Pourcentage Off-net :** cet indicateur montre le pourcentage de la durée des appels off-net sur la durée totale des appels, il permet d'avoir une idée sur le coût d'interconnexion (un opérateur télécom paye un tarif sur les appels destinés aux lignes d'un autre opérateur, ce tarif est connu sous le nom : coût d'interconnexion).
- **Pourcentage international :** cet indicateur montre le pourcentage de la durée des appels internationaux sur la durée totale des appels.

KPIs de Revenu :

Les indicateurs de revenu concernent la performance de l'opérateur télécom en ce qui concerne son revenu. Ils permettent de faire des zooms sur les montants rechargés, le nombre de clients qui rechargent, le cout des appels on-net (entre nous clients) et off-net (vers un client d'un autre opérateur ou d'un opérateur vers le nôtre) et internationaux. Nous citons ci-dessous les KPIs de revenu contenus dans le rapport :

- **CA :** Chiffre d'Affaires : Cet indicateur montre la somme totale des montants rechargés par les clients prépayés de l'opérateur, il est calculé en TTC.
- **Nombre de Rechargeurs :** Comme son nom l'explique, cet indicateur montre le nombre total des clients qui ont fait au moins un acte de recharge.
- **Pourcentage de rechargeurs :** c'est le ratio des clients rechargeurs dans le mois par le nombre total des clients actifs ou un client actif selon la norme ANRT est le client qui a

fait au moins une activité (recharge, appel entrant ou sortant, SMS entrant ou sortant ou navigation) dans les trois derniers mois.

- **Nombre de Recharges :** cet indicateur montre le nombre des actes de recharge (carte ou recharge expresse) effectués dans un mois.
- **MMPR :** Montant Moyen Par Recharge : Comme son nom l'explique, cet indicateur montre le montant moyen des recharges effectuées par les rechargeurs, il est calculé par le ratio entre le revenu total du mois et le nombre de recharges.
- **Fréquence de la Recharge :** l'indicateur fréquence de la recharge montre une moyenne de nombre de fois qu'un client rechargeur effectue une recharge (nombre de reproduction de l'acte recharger chez un client) ; il est calculé par le ratio entre le nombre de rechargeurs et le nombre de recharges.
- **ARPU (MMR Parc):** Average Revenue Per User : Cet indicateur montre le revenu de l'opérateur télécom en moyenne par utilisateur. Cette moyenne est calculée par le ratio entre le revenu total du mois et le nombre de clients actifs.
- **MMR (MMR Rechargeur):** Montant Moyen Rechargé : Comme son nom l'explique, cet indicateur nous donne le montant moyen rechargé par les utilisateurs. Il est calculé par le ratio entre le montant total des recharges du mois sur le nombre de clients de cet opérateur ayant rechargés ce même mois.

B.2.2.2- Procédures de reporting

La solution existante actuellement pour réaliser les tableaux de bord repose sur l'utilisation de logiciels de raquetteurs SAS et BO ainsi que les outils bureautiques tel que Excel PowerPoint pour la partie présentation .

Chaque mois on est amené à faire la procédure suivante :

- **Extraire les données :** à l'aide des logiciels précédemment cités, la première étape pour l'élaboration des rapports est l'extraction des données à partir du datamart. Ces outils facilitent ces extractions à l'aide des interfaces utilisateur, en glissant les variables dans des fenêtres de requêtes, le logiciel génère une requête SQL et exporte le résultat dans une table juste après la fin de l'exécution. Le nombre d'extractions à faire varie d'un tableau de bord à un autre ce qui augmente le temps nécessaire à un rapport donné.

- **Faire les calculs :** une fois les données extraites et restituées dans des tables SAS, l'opérateur commence à calculer les différents indicateurs et paramètres suivant les formules de calcul. Le générateur de requêtes permet de faire ces calculs en lui spécifiant les méthodes de calculs.

- **Faire des vérifications :** toute erreur est possible, c'est pourquoi l'opérateur doit faire attention aux résultats et procède à des vérifications s'il pense qu'un chiffre peut être erroné ou inattendu.

- **Exporter et restituer les résultats :** une fois les indicateurs et paramètres sont calculés, l'opérateur fusionne les différentes tables pour aboutir à une table simple à exporter. Les outils BI du service cités précédemment permettent l'exportation des tables avec l'extension Excel (*.xls) à l'aide d'un bouton d'exportation.

On peut schématiser le processus comme suit :

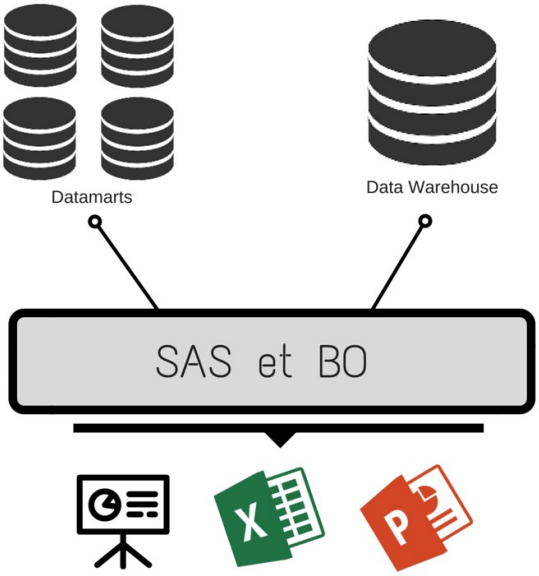


Fig 8: Processus de reporting

Avant de présenter les résultats, ils sont restitués dans des autres fichiers Excel, générés des graphes pour assurer une bonne lisibilité des informations livrées.

Partie C : <Comparaison entre les solutions Open source de mining>

Ce chapitre a pour but de justifier le choix de R comme étant la solution Open Source proposée pour la migration de technologie du service pour faire du data mining en le comparant avec trois autres outils open source de data mining.

C- Solutions Open Source de data mining

Pour créer une grande opportunité au développement de ses propres solutions et réduire le coût et la dépendance d'une seule gamme de produits sous licence propriétaire, le département Pricing et Business Intelligence cherche à intégrer une solution Open Source comme outil de base pour le travail du service Data mining. Une solution open source garantira l'optimalisation de ces procédures, une personnalisation orientée au besoins spécifiques du service et assurer une fluidité et créer une plate-forme de data mining qui peut être implémentée par d'autres opérateurs. Dans un premier temps comparant les outils data mining open source disponibles sur le marché, pour en déduire le choix de R parmi eux et présenter les solutions entreprise utilisant R pour des architectures plus performantes pour assurer une augmentation de productivité et une réponse rapide aux requêtes et créer ou déployer les modèles plus rapidement. Les paragraphes qui suivent présenterons quatre outils data mining basées sur différents concepts et concentrés sur différentes applications.

C.1- Orange

C.1.1- Présentation de l'outil

Orange est un outil gratuit et à source libre d'exploration de données. Il propose des fonctionnalités de modélisation à travers une interface visuelle, une grande variété de modalités de visualisation et des affichages variés dynamiques. Développé en Python, il existe pour les trois systèmes d'exploitation : Windows, Mac et Linux.
Créé par une Orange, une société créée en 1996 sous licence GPL et permet la visualisation et l'analyse de données en offrant plusieurs services :

- Programmation visuelle
- interface simple et interactive
- visualisation, interaction et analyse
- boite à outils
- interface de scripting
- extensible
- peut exécuter des scriptes en python
- valide pour les experts et les débutants

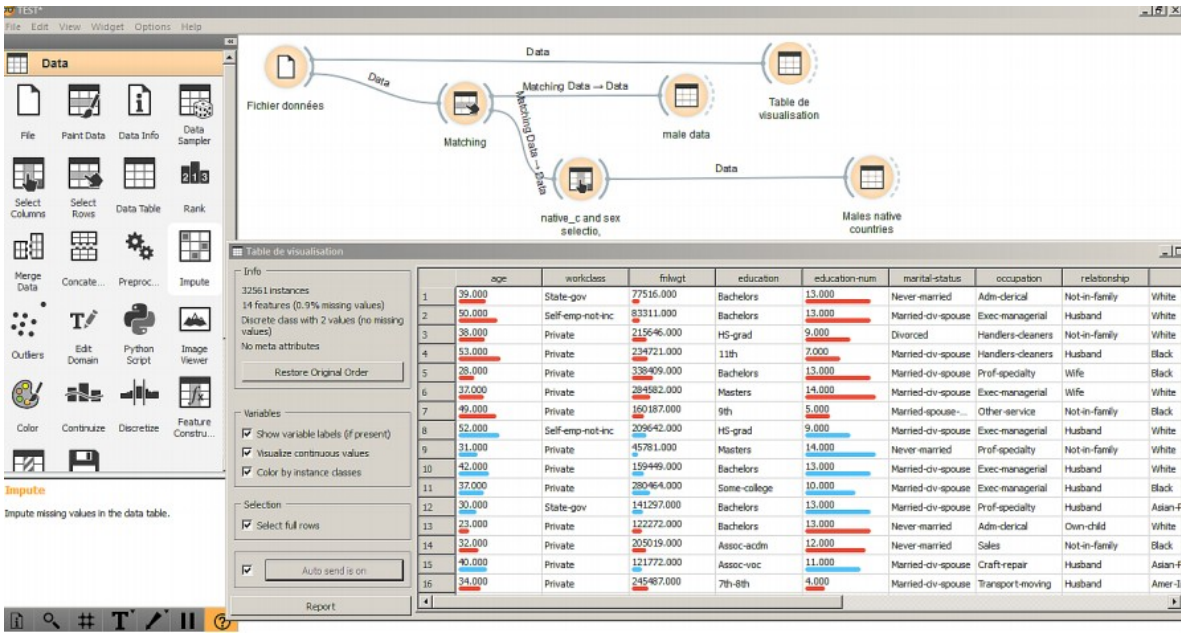


Fig 9 : interface d'orange

C.1.2- Puissance et limites

Puissance	Limites
• Supporte un langage haut niveau (python) • Scriptes réduits de data mining • Facile à utiliser • Interface Drag & Drop • Catégorisation et prédiction fluide • Debugger inclu	• Installation Volumineuse (Bibliothèques Qt) • Machine learning limité • Acune version serveur • Faible en statistique classique • Reporting faible

C.2- Open source R

C.2.1- Présentation de l'outil

R est un logiciel libre de traitement de données et d'analyse statistiques mettant en œuvre le langage S (langage de programmation très haut niveau pour l'analyse des données et des graphiques). C'est un logiciel libre distribué selon les termes de la licence GPL et disponible sous Linux, FreeBSD, NetBsd, Mac OS X et Windows. La première version a été publiée en 1993.

Il présente les services suivants :

- analyse des données massives
- visualisations graphiques
- analyse spatiale

- clustring
- text mining
- analyse de réseaux sociaux
- statistiques
- manipulation de données

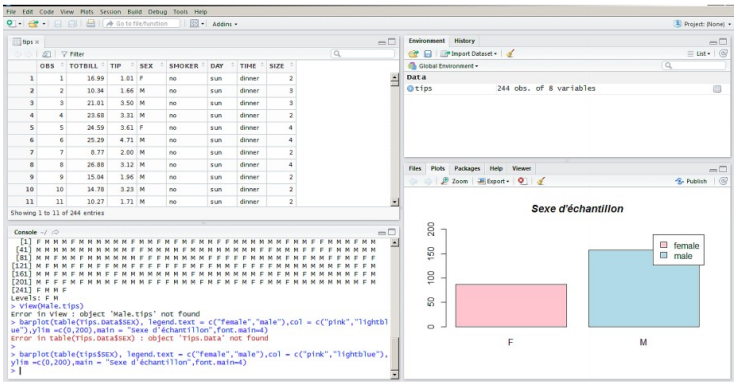


Fig 10 : Interface R

C.2.2- Puissance et limites

Puissance	Limites
• Gestion de données de volume énorme • Multitude de bibliothèques • technologie maîtrisée • Documentation riche • Puissant APL • Importation et exportation fluides • Haut niveau de programmation • Large communautés de développeurs	• Moins spécialisé en Data mining • Perte d'optimisation en programmation • nécessite des connaissances en APL

C.3- Weka

C.3.1- Présentation de l'outil

Weka (*Waikato Environement for knowledge analysis*) est une suite de logiciels d'apprentissage automatique. Écrit en Java, développé à l'université de *Waikato* en Nouvelle-Zélande. Weka est un logiciel libre à licence GPL. Le développement a débuté en 1993 avec du langage C mais après un échec, l'université a décidé de reprendre le projet avec du Java en 1997.

La suite de Weka présente les services suivants :

- visualisations des données et analyses
- GUI (interface graphique utilisateur)

- taches primitives de data mining
- exécution de plusieurs formats de fichiers de données
- connexion avec les bases de données
- gestion des connections via RDBMS

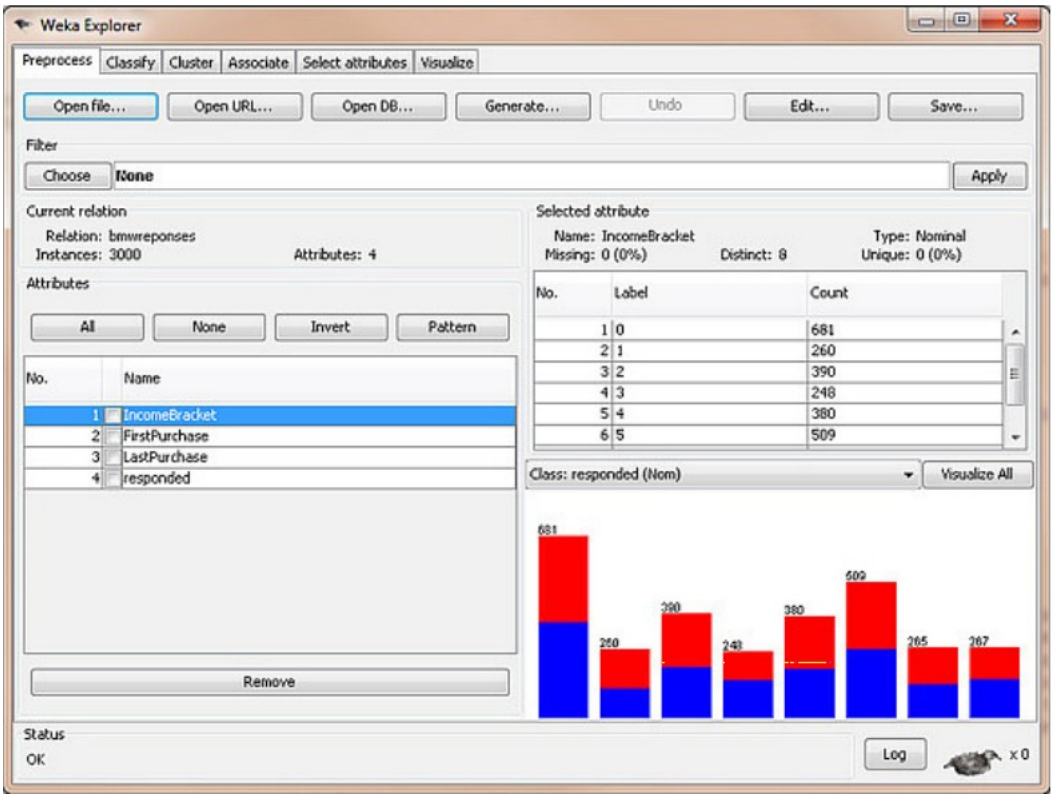


Fig 11 : interface Weka

C.3.2- Puissance et limites

Puissance	Limites
• Bonne connexion avec les bases de données • Intégrable avec d'autres technologies Java • parfait pour le machine learning et les règles d'association	• Faible documentation • Mauvais pour gérer Excel • Lecteur CSV non robuste • Faible en statistique classique

C.4- Rapid Miner

C.4.1- Présentation de l'outil

RapidMiner est développé par une société qui porte le même nom qui voulait présenter un outil groupant l'apprentissage automatiques des machines, data mining, text

mining, l'analyse prédictive et les analyses business. La version initiale est sortie en 2006 sous licence Affero GPL avec une documentation détaillée.
RapidMiner offre les services suivants :

- absence du code
- analyses prédictives
- chargement des données
- transformation de données
- visualisations multi-couches
- support de plusieurs sources de données

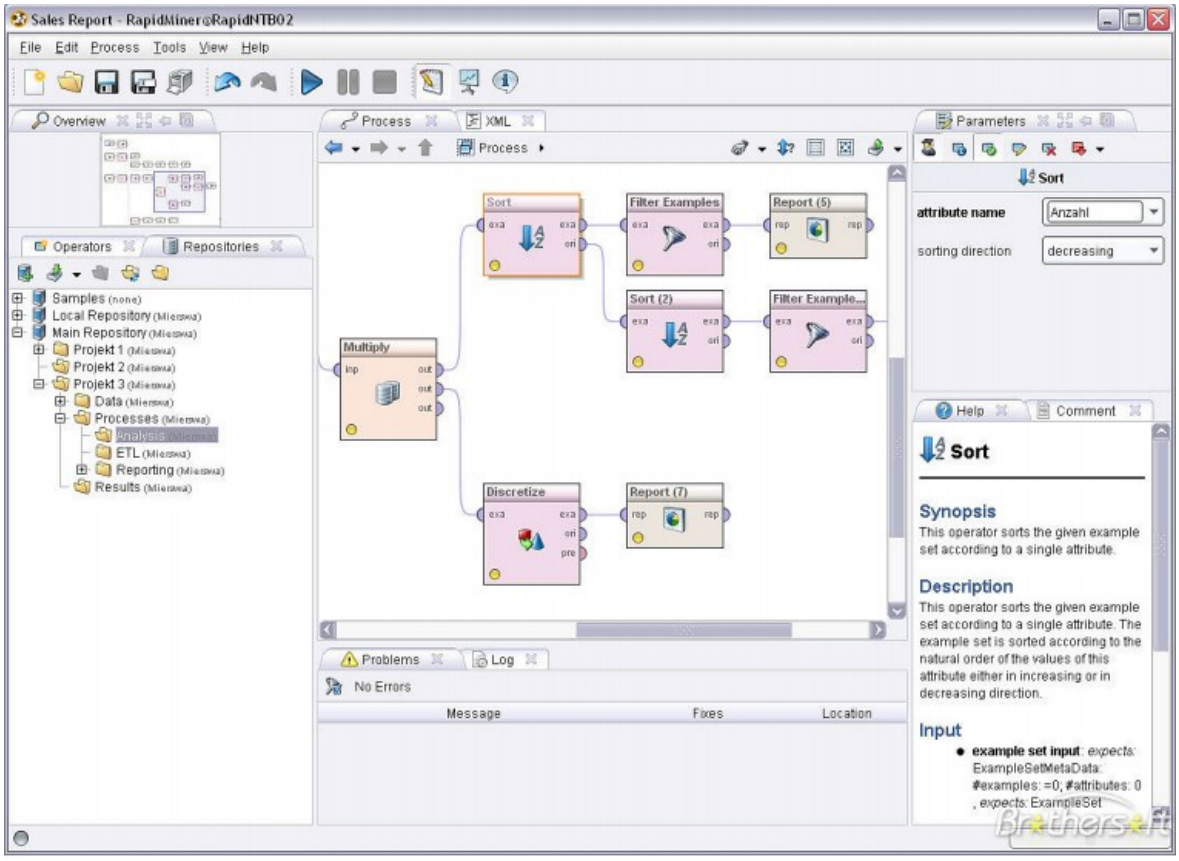


Fig 12 : interface RapidMiner

C.4.2- Puissance et limites

Puissance	Limites
• Plus de 1500 méthodes d'intégration, transformation, analyse, modélisation et visualisation • Multitude de procédures.	• Nécessite des connaissances en manipulation des fichiers de base de données et en SQL

C.5- Benchmarking

C.5.1- Benchmarking technique

La fiche technique présente la technologie sur laquelle est basée l'outil :

Outil	Lancement	Licences	Plateformes	Langages	Site web
Orange	2006	GNU GPL	Multi os	Python	orange.biolab.si
R	1997	GNU GPL	Multi os	R, C et Fortran	r-project.org
Weka	1993	GNU GPL	Multi os	Java	cs.waikato.ac.nz/ml/weka/
Rapid Miner	2006	GNU GPL	Multi os	Indépendant	rapidminer.com

Fig 13 : fiche technique des outils open source

C.5.2- Benchmarking de performance

Cette partie compare les quatre outils en cinq aspects d'utilisation qui sont la gestion des fichiers, les statistiques, le data mining, la business intelligence puis l'architectures de implémentation de l'outil avec des recommandations de plusieurs sociétés universelles.

Outil	Gestion de fichiers	Statistiques	Data mining	Business intelligence	Architecture	Recommandations
Orange	• Simple • Lente • non paramétrable	• Faible • Limité	• Machine learning limité • Catégorisation fluide • Scripts réduits	• Reporting faible • Interface interactive	• Client Desktop	★★☆☆☆ • Astra-Zeneca • Jožef Stefan Institute
R	• Rapide • Paramétrable • Distante • optimale	• Adapté • Tout types • puissant	• Moins spécialisé • Manipulations APL	• Distribué sur plusieurs packages • reporting non finalisé	• Client Desktop • Client / Serveur • RStudio	★★★★☆ • Bank of America • Facebook • google
Weka	• Simple • Faible pour .csv • faible pour Excel	• Faible • Limité	• Robuste pour le machine learning • Plusieurs méthodes disponibles et optimisées	• Trop faible à cause des problèmes d'importation de données	• Client Desktop • Client / Serveur • Web services	★★★☆☆ • Rechtsportal • Particuliers
Rapid Miner	• Robuste • Paramétrable • Intéractive	• puissant • adapté	• Optimisé • text mining • machine learning	• Donne de bon résultats prêts à l'exploitation	• Client Desktop • Client / Serveur	★★★★☆ • CISCO • Paypal

Fig 14 : benchmarking de performance des outils open source

C.5.3- Conclusion :

Ce benchmarking a permit de déduire la décision d'intégrer R pour le data mining au sein de Méditel parmi les outils open source décrits dans cette partie. R est purement statistiques, il est très développé et maîtrise, les solutions entreprise basés sur R sont à coût très réduit, il est bien adapté en architecture Client/Serveur et il est connu pour avoir une communauté actives pour l'assistance et les mises à jour.

C.6- Version orientée entreprise de R :

Parmi les solutions disponibles sur le marché basées sur R et orienté Entreprise et Big Data on trouve Revolution R Entreprise qui est un projet orienté data mining sur des architectures distribuées et gérant des lots immenses de données. Le projet fait partie d'un projet du fameux Révolution Analytics.

C.6.1- Revolution R et la vision entreprise :

RevoR acronyme de Revolution R est une distribution puissante de R Open Source publiée par Revolution Analytics comme étant la plate-forme la plus rapide et la plus efficace pour les données massives et l'intégration dans l'entreprise pour faire de grandes analyses. Supporte une variété de statistiques sur les grandes données , la modélisation et l'apprentissage de machine à capacités prédictives, Revolution R Enterprise est également 100% R.

Révolution R Entreprise (RRE) est la version entreprise de Revolution R Open (RRO). Ces deux solutions qui présentent une distribution de Open Source R (R brut) par la société Révolution Analytics.

C.6.2- Puissance de RevoR :

Revolution R Enterprise stimule et accélère R, l'exécution de scripts R avec une haute performance, une architecture parallèle qui prend en charge les systèmes distribués et clusters, y compris Hadoop et les entrepôts de données d'entreprise.
Revolution R Enterprise accélère l'analyse statistique traditionnelle en utilisant le calcul des données massives et des techniques de gestion des données. Avec Revolution R Enterprise, les utilisateurs R peuvent explorer, modéliser et prédire à haute échelle.

Cette distribution optimise et offre plusieurs outils existants et étend les outils de R en sa version basique. Et apporte les fonctionnalités suivantes :

- Dernière version de R pour le calcul statistique
- Haute performance avec le multi-threading

- Compatibilité avec tout les outils R
- Traitement de big data
- Méthodes plus puissantes replaçant celles de R
- Fluidité lors de l'importation de données

C.6.3- Versions disponibles :

	Revolution R Open	Revolution R Enterprise
R Language Engine	Included	Supported
Multi-core processing (Intel® Math Kernel Library)	Included	Supported
Reproducible R Toolkit	Included	Supported
ParallelR: Parallel Programming Toolkit		Supported
RHadoop: R interface to Hadoop MapReduce		Supported
DeployR Open: Web Services API		Supported
RRE DeployR - Scalable & Secure Deployment		Licensed & Supported
RRE ScaleR - Big Data Toolkit and PEMAs for R		Licensed & Supported
RRE DistributedR - EDW, Grids, Hadoop		Licensed & Supported
AdviseR Technical Support		Included

Fig 15 : versions RRO et RRE

Partie D : <Intégration de R chez *Méditel*>

Pour bénéficier d'une transition fluide lors du changement de technologie, une étude des possibilité d'intégrations de R sous différentes implémentations et d'en choisir celle qui sera la plus optimale pour le service BI de *Méditel*.

D- Intégration de R chez Méditel

D.1- SAS vs R (RRE):

D.1.1- Test de performance vis à vis l'architecture

Le centre de recherche et de planning *Allstate*, a voulu implémenté un modèle prédictif linéaire générique sur 150 millions d'enregistrements en utilisant 4 approches de différentes technologies pour en sortir avec une évaluation de performance pour chacun d'eux.

1ère Approche : 'proc GENMOD' de SAS
2ème Approche : 'Hadoop Cluster'
3ème Approche : 'Open Source R'
4ème Approche : 'Revolution R Entreprise'

Le test est fait de tel sorte à maximiser les besoins matériels et logiciels pour réussir l'implémentation dans chaque solution.

Les plat-formes de test :

Outils	**Plate-forme**
SAS	20 cœurs serveur SUN
Hadoop cluster	10-nœuds (8 cœurs/nœud)
Open Source R	Serveur 16 cœurs et 250Go de RAM
Revolution R Entreprise	5 Noeuds (4 cœurs/nœud) LSF cluster

*Les plat-formes de test sont différentes dû à l'architecture conseillé pour mieux tourner l'outil à ses performances maximales.

Sous SAS :
'proc GENMOD' présente le traitement des modèles linéaires génériques sous SAS. Pour implémenter le modèle du test SAS a pris 5 heures pour retourner le résultat du modèle de poisson avec 150 millions d'observations et 70% comme degré de liberté.Ce qui représente un taux de productivité faible surtout dans les taches à court délai.

Hadoop Cluster :
En installant un cluster Hadoop, plusieurs taches de programmation étaient nécessaires (les équations de matrices pour chaque itération, le map-reduce…) ce qui était coûteux en temps sans parler du temps qu'a pris chaque itération 1h30min pour chaque itération et 10itérations au total.

Open Source R :
Étant gourmand en RAM (tout le traitement se fait dans la mémoire vive) l'implémentation s'est faite sur un serveur de 250 Go de RAM. Le problème était lors du chargement des enregistrements qui a pris 3 jours. L'échantillonnage après était rapide mais c'est difficile d'accepter un modèle basé sur échantillonnage. Chargement de données médiocre mais traitement rapide.

Revolution R Entreprise :
Revolution R Entreprise (avec RevoR), le chargement était fait sur un cluster de 5 nœuds de 4 cœurs (20 cœurs au total). l'opération a pris 5,7 min en s'aidant de l'outil rxGlm disponible dans la distribution. C'est la plus rapide approche dans ce test. Car vis à vis l'architecture distribuée de la plat-forme l'optimisation de la distribution des calcules sur plusieurs cœurs (diviser le travail pour régner) rend le temps de calcul et de traitement très réduit.

On peut sortir avec les points suivants :
• SAS peu faire le traitement mais un peu lent.
• C'est possible de programmer le modèle sous Hadoop mais c'est plus long.
• Les données sont très larges pour Open Source R.
• RRE donne les mêmes résultats que SAS mais en un temps 5 fois plus rapide.
• Opter pour une architecture adéquate pour l'outil permet de bien bénéficier de ses performances.

D.1.2- Test de performance vis à vis les données à traiter

Un test publié par Inside Big Data sur la différence entre le traitement de données sous SAS et sous RRE, En testant plusieurs fonctionnalités et outils essentiels de statistique et de data mining sur une source de données immense de 1 000 000 d'observations sur la version 9.4 de SAS (HP Procs inclus) et 7 de RRE.

Les outils de chaque programme :

Tache	RRE 7	SAS 9.4
Statistique descriptive d'une seule variable numérique	rxSummary	PROC SURVEYMEANS
Median et déciles pour une variable numérique	rxQuantile	PROC SURVEYMEANS
Distribution de fréquence d'une variable textuelle	rxCube	PROC FREQ
Régression linéaire (1 réponse, 20 variables numériques)	rxLinMod	PROC REG PROC HPREG
Régression linéaire (1 réponse, 20 variables mixtes)	rxLinMod	PROC GENMOD
Régression linéaire escalier avec 100 variables numériques	rxLinMod	PROC REG
Régression logistique (1 réponse, 20 variables numériques)	rxLogit	PROC LOGISTIC
Régression généralisée (1 réponse, 20 variables numériques de distribution Gamma)	rxGlm	PROC GENMOD
Clustering K-means avec 20 variables actives	rxKmeans	PROC FASTCLUS

Tache	RRE 7	SAS 9.4
Clustering K-means avec 100 variables actives	rxKmeans	PROC FASTCLUS
Scoring du premier modèle linéaire, avec 10 fois la taille de données	rxPredict	PROC SCORE

Les scriptes du test sont disponibles sur: https://github.com/RevolutionAnalytics/Benchmark

Les scriptes sont lancés sous chaque outil pour assurer l'exécution indépendante de chaque tache et ça a donné les résultats suivants :
le temps d'exécution est en secondes,

Tache	RRE 7	SAS 9.4
Statistique descriptive d'une seule variable numérique	1,2	24,3
Médiane et déciles pour une variable numérique	1,4	24,7
Distribution de fréquence d'une variable textuelle	0,8	26,2
Régression linéaire (1 réponse, 20 variables numériques)	6,8	26,7
Régression linéaire (1 réponse, 20 variables mixtes)	7,3	26,9
Régression linéaire escalier avec 100 variables numériques	13,9	26,2
Régression logistique (1 réponse, 20 variables numériques)	16,9	98,0
Régression généralisée (1 réponse, 20 variables numériques de distribution Gamma)	32,7	57,6
Clustering K-means avec 20 variables actives	10,1	100,2
Clustering K-means avec 100 variables actives	32,5	107,5
Scoring du premier modèle linéaire, avec 10 fois la taille de données	137,6	519,4

On peut en tirer les résultats suivants :

•ScaleR de RRE fait le traitement et l'analyse 42 fois plus rapide que SAS
•ScaleR de RRE peut s'adapter aux données massives.
•SAS HP Procs augmente la performance de SAS d'une façon marginale.
•RRE présente un outil puissant pour les solutions d'analyse et data mining.

D.2- Scénarios d'intégration.

R en sa version open source s'exécute soit sous forme de scripts (.R) ou avec des commandes sur la R console. Ceci qui pose un problème pour les utilisateurs non habitués à travailler avec des langages vecteurs ou multidimensionnels (R,matlab..), et requière des connaissances basiques indispensables du langage R.
A *Méditel*, remplacer SAS par R nécessite une adaptation de toutes les situations d'utilisation du premier outils et l'intégrer avec R en gardant un aspect d'entreprise (Productivité, Respect des délais, Reproductibilités, Consistance, Disponibilités, Rapidité…). Alors comment intégrer R (RevoR) chez *Méditel* ?

D.2.1- Intégrer R avec un IDE

Déployer R sous console pour exécuter des commandes ou scripts sur l'outil R GUI (R graphical user interface), téléchargeable en version exécutable depuis le dépôt CRAN officiel de R : https://cran.r-project.org/mirrors.html

D.2.1.1-Implémentation

Parlant d'un déploiement local au premier lieu, après avoir installer R sur une machine il faut intégrer Revolution R Open comme première étape :
-cette version est téléchargeable sur : http://www.revolutionanalytics.com/revolution-r-open
-pour un traitement multi-thread il faut disposer d' Intel® Math Kernel Library
Après, vient l'intégration de RevoR Entreprise: http://www.revolutionanalytics.com/get-revolution-r

Pour s'assurer que Revolution R est bien intégré la console va afficher la version après les commandes de help() et demo() :

```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Revolution R Enterprise (Compute Node) version 7.5 (64-bit):
an enhanced distribution of R
Revolution Analytics packages Copyright (C) 2015 Revolution Analytics, Inc.

Type 'revo()' to visit www.revolutionanalytics.com for the latest
Revolution R news, 'forum()' for the community forum, or 'readme()'
for release notes.
```

Fig 16 : RRE console

Un IDE est un environnement de développement intégré son objectif est d'augmenter la productivité des programmeurs en automatisant une partie des activités et en simplifiant les opérations. Pour R l'ide recommandé est Rstudio téléchargeable en version desktop ou Serveur
sur : https://www.rstudio.com/products/rstudio2/
Rstudio va permettre d'exécuter des commandes, des scripts, d'importer des données, gérer les variables, gérer les bibliothèques…

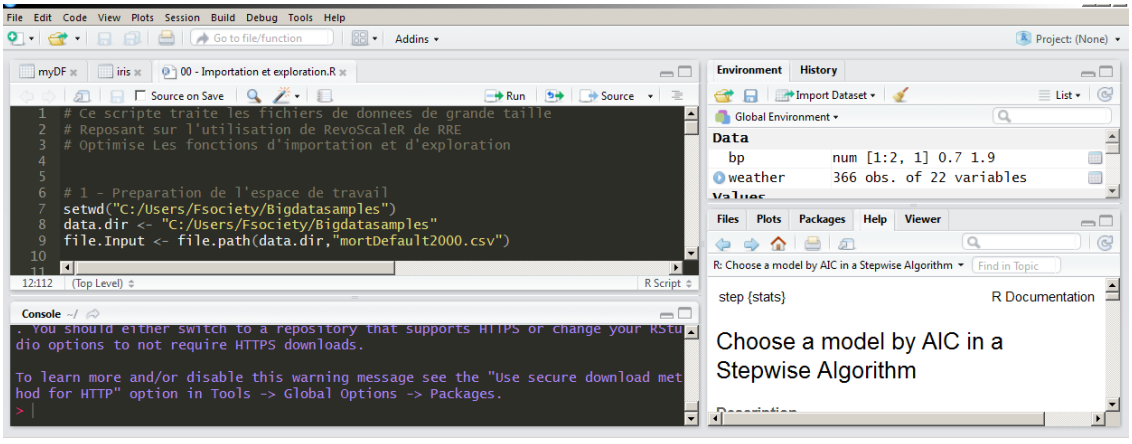


Fig 17 : Interface Rstudio RRE

D.2.1.2- Architecture

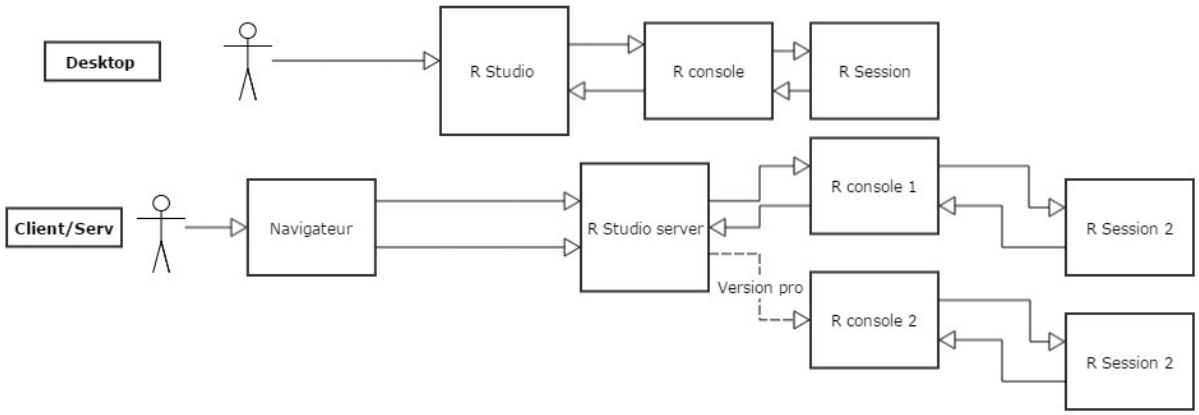


Fig 18 : architecture Rstudio (IDE)

D.2.1.3- Avantages et inconvénients

Avantages	Inconvénients
- Contrôle totale. - Transparence d'exécution. - Compréhension des erreurs. - Montée en expérience en R - Optimisation d'exécution. - Optimisation des ressources. - Paramétrage immense des processus.	- Nécessite une connaissance moyenne du langage. - Dépend du scripting. - Productivité ralentie par le code. - Faible interaction avec un non programmeur. - Documentation mono linguistique. - Dépendance de la documentation.

D.2.2- Développer une interface pour R

D.2.2.1- Principe et implémentation

Plusieurs langages de programmation proposent des bibliothèques, frameworks et APIs pour intégrer R dans leurs applications. Donnant l'exemple de Java avec la bibliothèque Rcaller pour exécuter du R au sein de JAVA, ou bien le framework (*structure logicielle,* un ensemble cohérent de composants logiciels structurels) Shiny proposé par Rstudio pour créer des applications interactives écrites en R et interprétées en Javascript et interfacés par du HTML5/CSS.

Shiny propose plusieurs fonctions en R pour une multitude d'éléments à intégrer pour créer une interface interactive en s'appuyant sur des éléments en html pour l'affichage des

sorties/entrées.

La programmation d'une ShinyApp (application shiny) s'exécute sur une machine cliente où sur un serveur (version serveur) sous la (les) session dédiée à Rstudio.

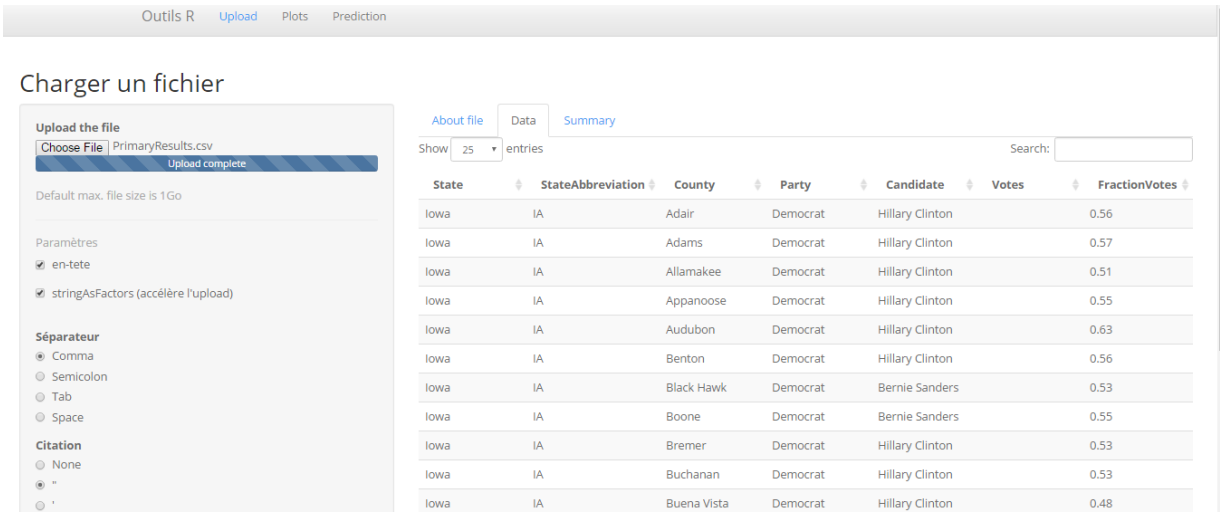


Fig 19 : application shiny avec R

D.2.2.2- Architecture

La structure de travail sur une application Shiny est structurée comme suit :

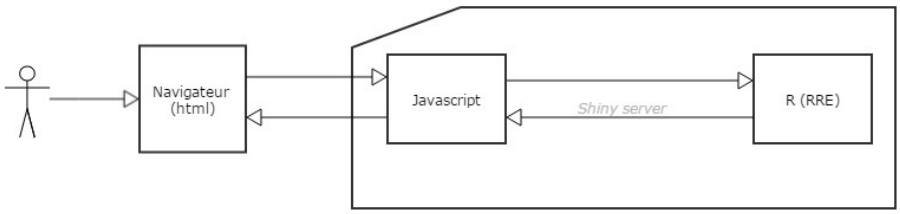


Fig 20 : architecture shinyApp

Quand au développement de l'application, il se fait sous deux scripts en R :

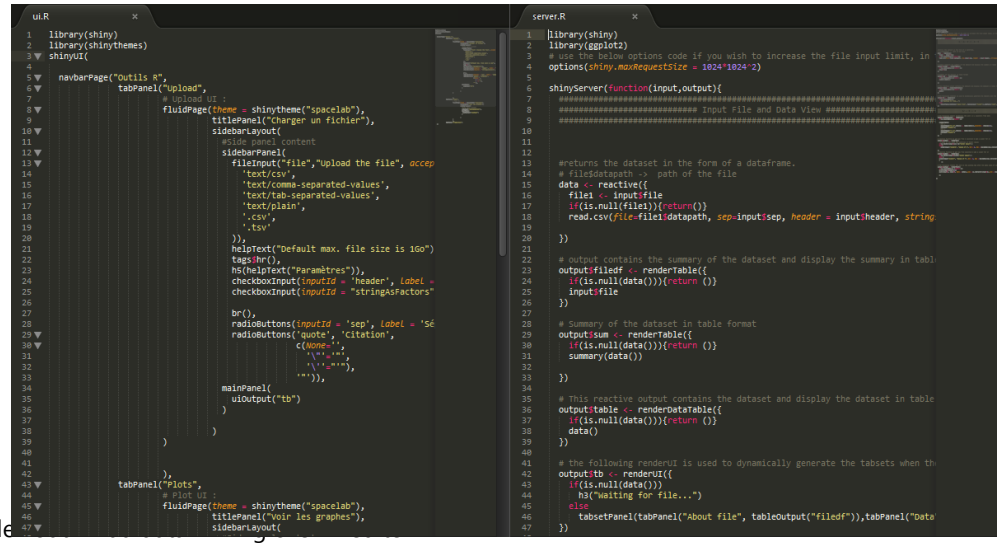


Fig 21 : Code pour l'application Shiny

ui.R : Script pour générer les éléments visuels (html) soit permanent soit dynamiques. Plusieurs fonctions pour les entrées et des arguments pour le style (css). L'identificateur de chaque élément sera chargé pour le traitement dans deux objets input et output selon le rôle.

server.R : Script pour le traitement niveau serveur. Où se fait le traitement des entrées avec du code R orienté traitement et analyse et associe les éléments de sortie de ui.R avec leurs valeurs à afficher. Toutes les fonctions sont réactives, ça veut dire qu'elles s'exécutent d'une façon continue tant que la page est ouverte, néanmoins pour exécuter une fonction à exécution unique il existe une fonction *isolate()*.

D.2.2.3- Avantages et inconvénients

Avantages	Inconvénients
- Présentation esthétique. - Interaction rapide. - Facilité de travail. - Intègre plusieurs fonctions javascript.	- Coût de développement très haut. - Lourd sur la mémoire vive. - Limité à des problèmes de moyenne taille. - Nécessite l'assistance d'un programmeur. - Déploiement pas si fiable.

D.2.3- Noeuds Knime.

D.2.3.1- Knime pour le data mining

KNIME (prononcé naɪm), est une plate-forme open source d'analyse de données, de rapports et d' intégration. KNIME intègre divers composants pour l' apprentissage des machines et l'exploration des données grâce à son concept modulaire de pipelining de données. Une interface graphique qui permet l'assemblage de nœuds pour le traitement des données (ETL: Extraction, Transformation, Chargement), pour la modélisation et l'analyse des données et la visualisation. Conçu et basé sur Java, Knime peut intégrer plusieurs outils de traitement de données externes tel que des bases de données en intégrants des extensions, et supporte plusieurs langages de script.

Knime est disponible en version gratuite ou payante sur : https://www.knime.org/server-products

D.2.3.2- R implémenté sur Knime.

Vu que Knime supporte des extensions développés pour étendre ses fonctionnalités (une version qui comprend toutes les extensions gratuite est aussi disponible : https://www.knime.org/downloads/overview) elle implémente une version interne de R, avec un set de plusieurs nœuds R :

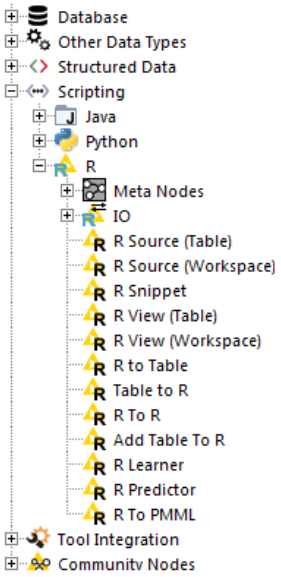


Fig 22 : Noeuds R sous Knime

Pour changer la version de R et utiliser une personnalisée (RRE dans notre cas) il faut changer l'emplacement de l'exécutable de R dans Knime.

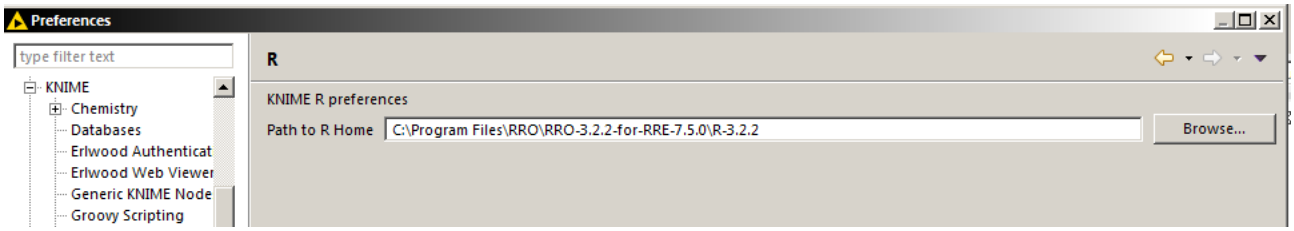
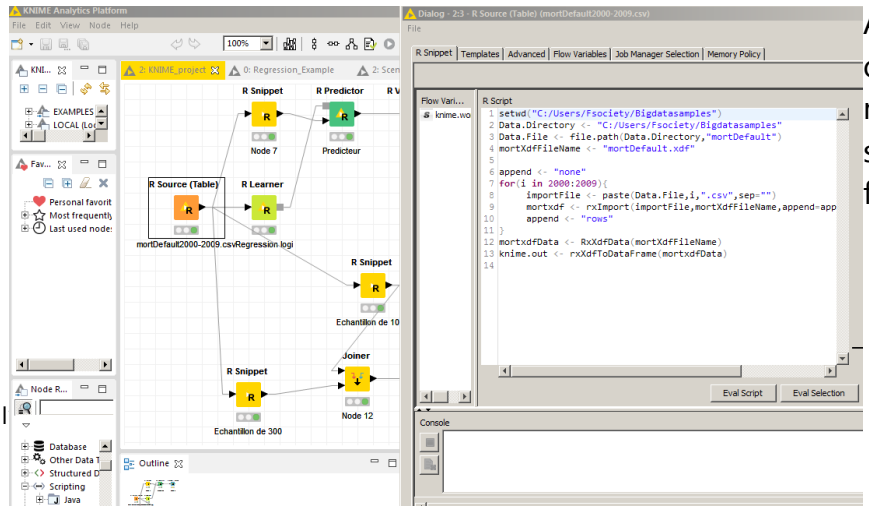


Fig 23 : configuration R sous Knime



Après avoir configuré ça, les nœuds R peuvent supporter des fonction de RevoR :

Fig 24 : nœud avec un script R

D.2.3.3- Les nœuds R

Les nœuds de Knime englobe des relations entre entées et sorties tant qu'elles sont compatible avec le type de données traités dedans. Le plus surprenant est que la sortie donnée de R peut être manipulée par d'autres nœuds propres à Knime comme pour la jointure par exemple. En gros les sorites Knim.out et les entrée Knim.in peuvent être soir des données, soit des workflow sous R. Et il existe des convertisseur pour changer la structure des données et des variables sous R pour les traiter avec Knime et vis versa.

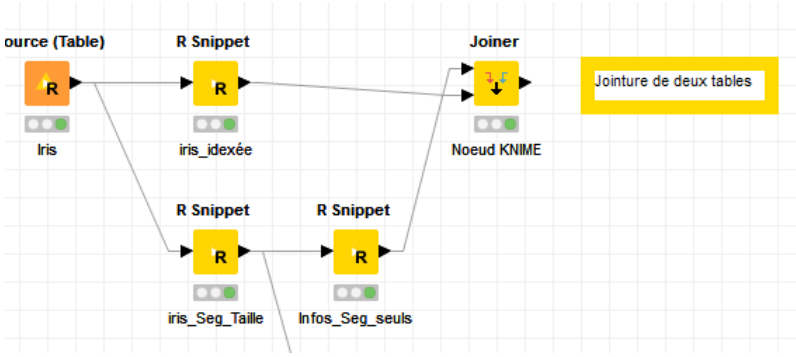


Fig 25 : workflow avec des nœuds R

D.2.3.4- Avantages et inconvénients

Avantages	Inconvénients
- Programme modulaire. - Workflow claire. - Haute productivité. - Stockage en mémoire au choix. - Intégration de plusieurs outils externes. - Configurable. - Déploiement facile.	- Nécessite une mémoire vive gigantesque. - Les fenêtres on besoin d'êtres rafraîchies de temps en temps. - Dépendance de Java pour des configurations bas niveau. - Besoin d'outils externes pour un reporting avancé.

- Possibilité de développer dessus.	

D.2.3.5- Conclusion et cas d'utilisation

En comparant l'utilisation de chaque solution d'intégration et ce que *Méditel* a l'habitude de travailler avec sous SAS, le choix tend vers Knime avec son interface et son workflow semblable à SAS. Pour intégration plus optimisée il faut considérer la version entreprise de R sous Knime pour avoir plus de services et de performance et de penser à faire le traitement dans des clusters et machines séparée pour rendre le temps de traitement trop court et gagner en productivité.
Comment intégrer R server sous Knime : https://goo.gl/11iIiv

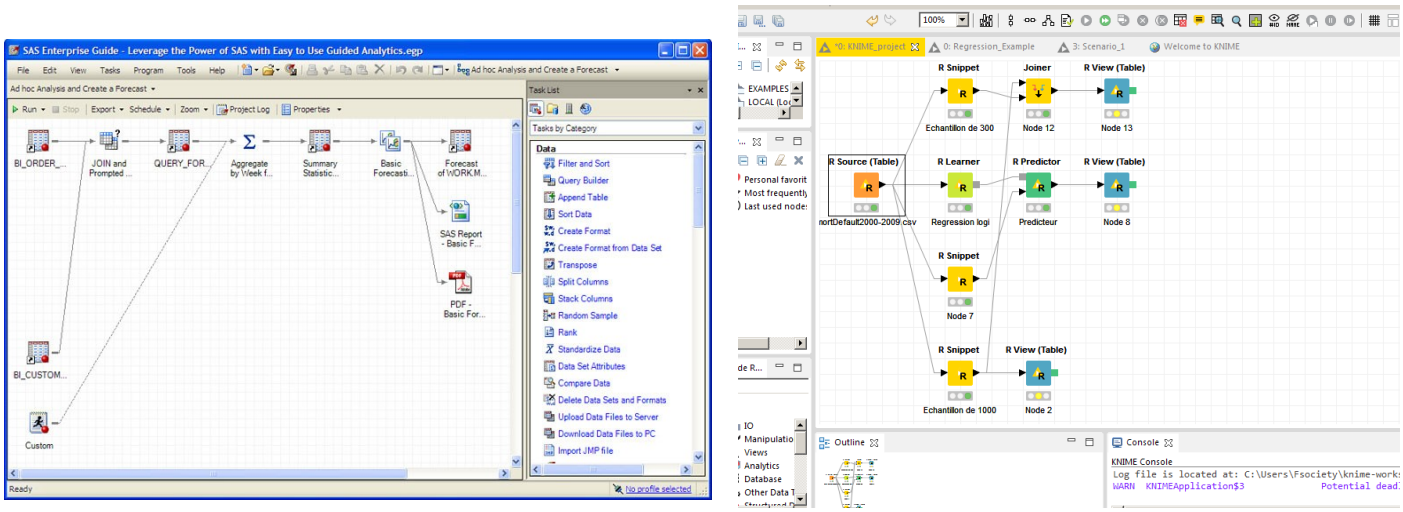


Fig 26 : interface de SAS (gauche) et Knime (droite)

Prenant un cas d'utilisation qu'on traitera avec les deux solutions, Knime et SAS EG pour créer le même workflow présentant une importation, une jointure, des colonnes calculées, une segmentation, un échantillonnage stratifié et deux reporting d'effectif segmenté.
Il s'agit de traiter deux datamarts de lignes mixtes (appartenant à plusieurs offres) qu'on aura à calculer quelques KPIs, segmenter les offres avec leurs libellés plus représentatifs (au lieux des codes d'offre) et en tirer les quatre catégories d'offres.
Comme résultat reporting on aura à donner l'effectif de chaque catégorie d'offre comme résultante de la segmentation de ces lignes. Ainsi qu'un échantillonnage stratifié qui garantira que la population de chaque catégorie est la même sur tout l'échantillon :

a - Workflow SAS Entreprise Guide :

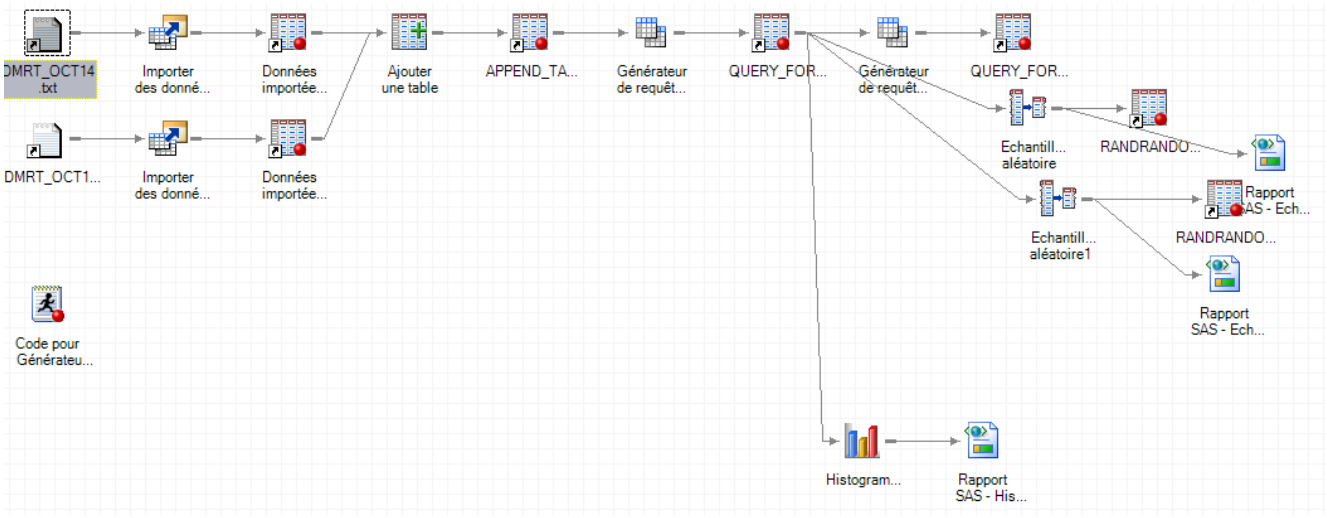


Fig 27 : Workflow SAS (cas d'utilisation)

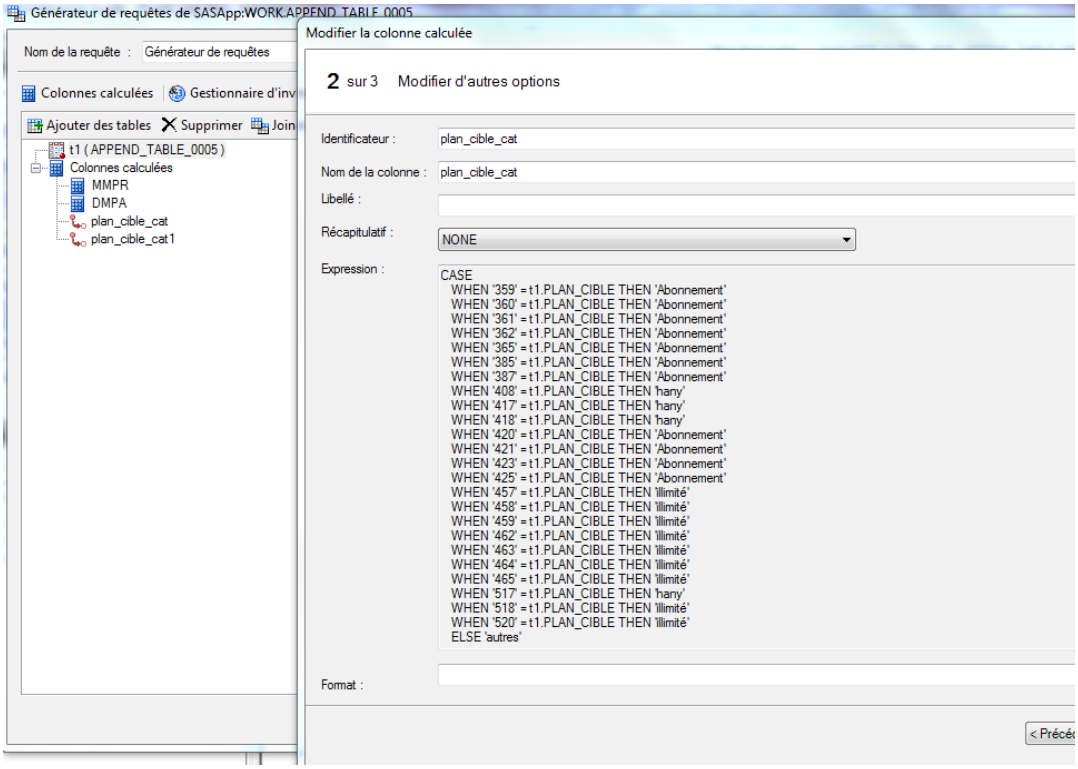


Fig 28 : Partie Segmentation

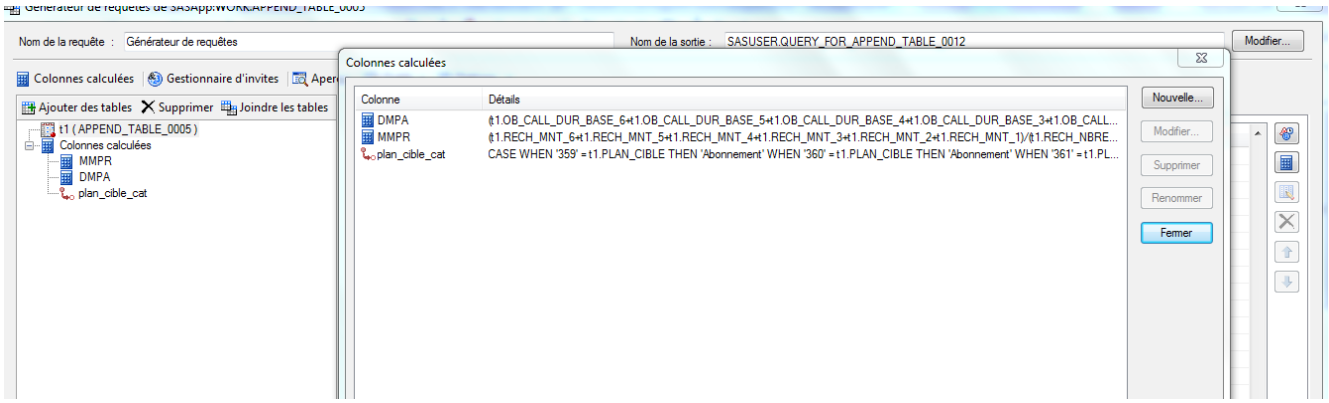


Fig 29 : Partie Colonnes calculées

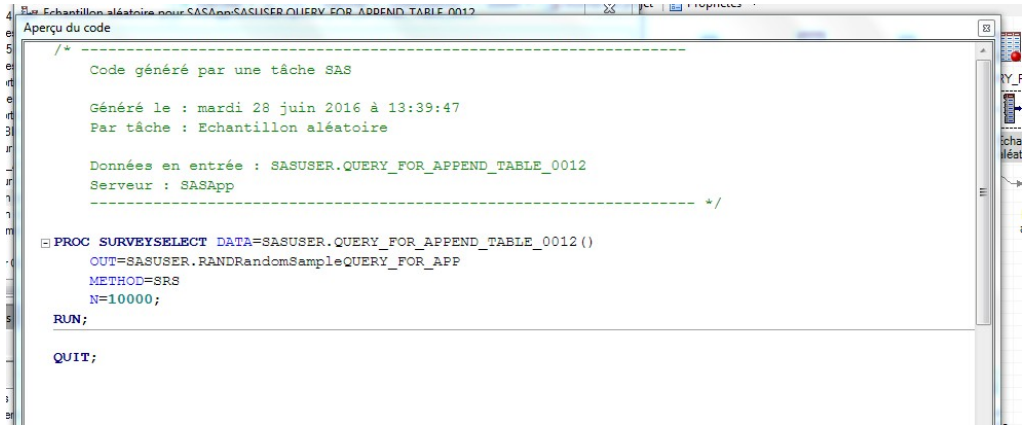


Fig 30 : Partie échantillonnage stratifié

Reporting de segmentation :

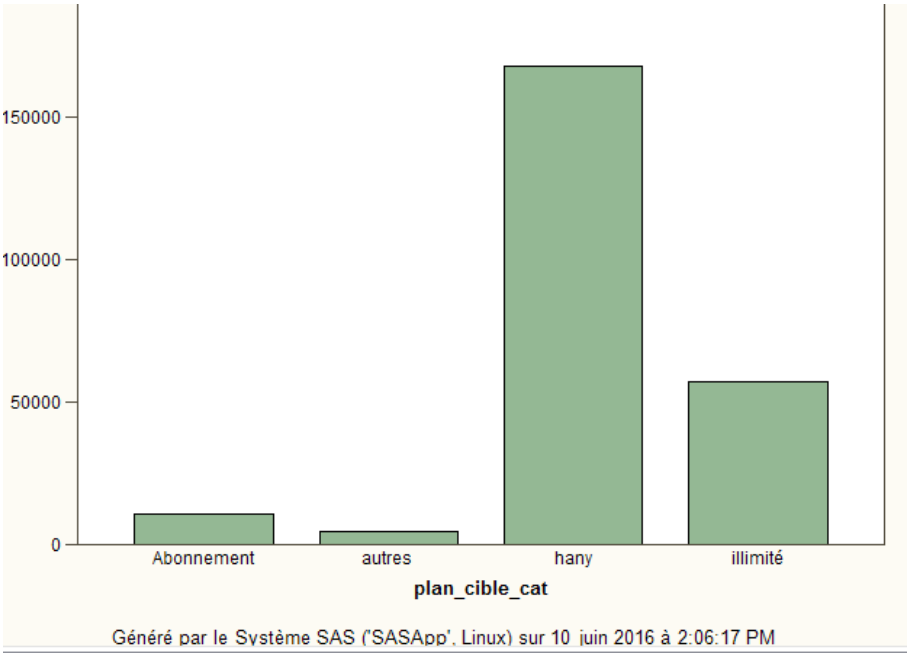


Fig 31 : Reporting de segmentation

b - Knime avec R inclus :

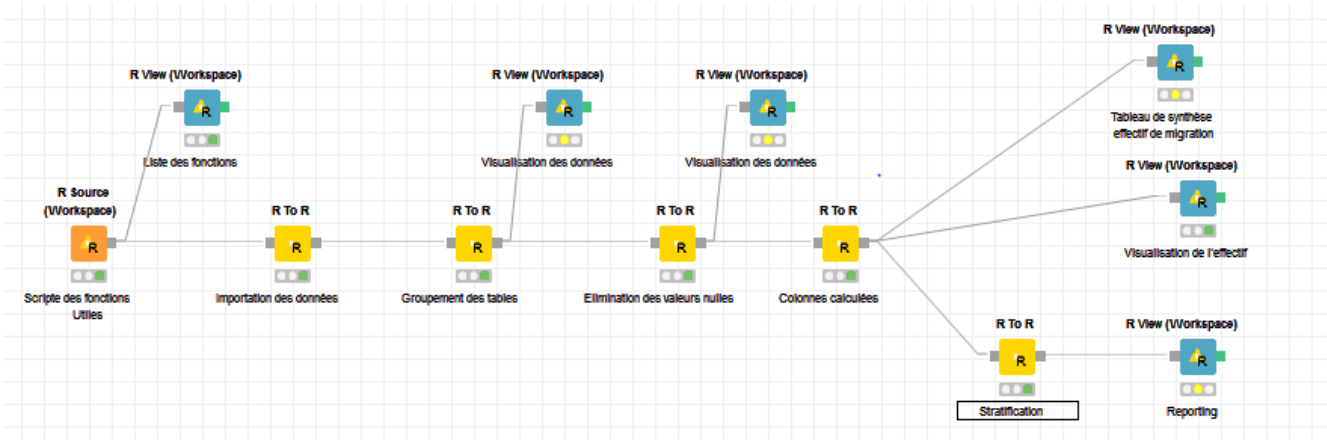


Fig 32 : Workflow Knime (cas d'utilisation)

Les étapes du workflow sont arrangées dans le scripte : Préparation de données.R arrangé pour le projet. Et on fait appel à chaque fonction dans les nœuds si dessous.

Reporting de segmentation :

	Var1	Freq
1	delete	0
2	Meditel Abonnement	10376
3	Pack Hany	165144
4	Pack Illimite	56054

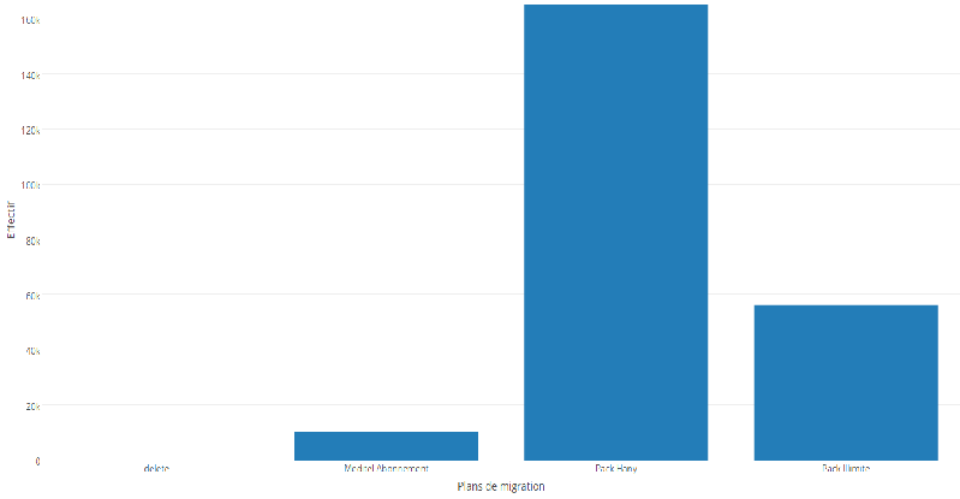


Fig 33 : Reporting Knime

c – Comparaison des résultats :

Les deux technologies donnent le même résultats sous le même workflow avec des vitesses différentes du à la différence entre les deux machines. Pour un exemple de data mining sous R, un problème de classification sera traité dans la partie suivante.

Partie E : <Problématique de classification sous R>

La partie E présente un cas d'application de R pour traiter les données de Méditel et en tirer un modèle de prédiction basé sur les méthodes de classification sous R. Le contexte de l'application est de prédire la migration des lignes prépayées vers le post-payé.

E- Problématique de classification sous R :

E.1- Problématique de classification

La classification désigne la répartition systématique en classes, en catégories, d'êtres, de choses ou de notions ayant des caractéristiques communs notamment afin d'en faciliter l'étude ou bien pour désigner un comportement ou changement de comportement. La classification réside une opération majeur dans le data mining. Et plusieurs méthodes de résolution de ce type de problématique font plusieurs champs d'études.

Dans une problématique de classification supervisée on peut formuler l'étude en une association des n éléments de la population (x1,x2…..,xn) à une des k classes prédéfinies et connues, tandis que la classification non-supervisée vise à ranger les éléments en k ensembles homogènes.
Parmi les méthodes de résolution des problèmes de classification on site les algorithmes suivants :

- Classificateur de Bayes
- Analyse discriminante et classificateur de Fisher
- kppv
- Arbres de décision
- Random forest
- Réseaux de neurones
- Séparateurs à vaste marge (SVM)

Au niveau de l'étude du modèle pris comme exemple dans ce qui suit deux méthodes seront prises en charge lors de la classification : l'arbre de décision et Random forest.

E.1.1- Les arbres de décision

Les arbres de classification et de régression (parfois aussi appelés arbres de segmentation ou de décision) sont des méthodes qui permettent d'obtenir des modèles à la fois explicatifs et prédictifs. Parmi leurs avantages on notera d'une part leur simplicité du fait de la visualisation sous forme d'arbres, d'autre part la possibilité d'obtenir des règles en langage naturel.

On distingue notamment deux cas d'utilisation de ces modèles :

- on utilise les arbres de classification pour expliquer et/ou prédire l'appartenance d'objets (observations, individus) à une classe (ou modalité ou catégorie) d'une variable qualitative, sur la base de variables explicatives quantitatives et/ou qualitatives.

➢on utilise les arbres de régression pour expliquer et/ou prédire les valeurs prise par une variable dépendante quantitative, en fonction de variables explicatives quantitatives et/ou qualitatives.

L'arbre de décision est un modèle qui repose sur l'utilisation d'un arbre comme modèle prédictif dans le but d'évaluer la valeur d'une caractéristiques d'un système depuis l'observation d'autres caractéristiques du même système.
La structure de l'arbre est faite de tel sorte que les feuilles sont les valeurs de la variable-cible et les embranchements correspondent à des combinaisons de variables d'entrées qui mènent à cette valeur.
C'est une technique supervisée car on utilise un ensemble de données pour lequel on connaît la valeur de la variable cible affin de construire l'arbre :

1. L'arbre est construit en séparant les données en des sous ensembles en fonction de la valeur d'une caractéristique d'entée d'une manière récursive.
2. La récursion prend fin jusqu'à ce que tout les sous ensembles ont la même valeur de caractéristique cible ou s'il n'y a pas d'amélioration de prédiction. Cet algorithme glouton est nommé l'Induction descendante d'arbres de décision.
3. Une variable d'entrée est sélectionnée à chaque nœud intérieur de l'arbre selon une méthode qui dépend de l'algorithme.
4. Chaque arête vers un nœud fils correspond à un ensemble de valeurs d'une variable d'entrée.
5. Chaque feuille représente soit une valeur cible soit une distribution de probabilité des valeurs cibles possibles.
6. La combinaison des valeurs d'entrées est représentée par le chemin de la racine vers la cible (la feuille).

Parlant des notions sur lesquels reposent le traitement lors de la création de l'arbre :

E.1.1.1- Notion d'entropie :

L'entropie d'un système désigne le niveau de désordre de ses éléments. L'entropie au niveau de l'arbre de décision est présenter au tant qu'une valeur entre 0 et 1 de la façon suivante :

- Les observations du système appartient à une même classe (entropie = 0)
- Si les classes prédéfinies sont uniformément distribuées dans les observations du système (entropie = 1)
- Si le système contient des observations distribuée aléatoirement sur les classes, là l'entropie est entre 0 et 1

Une entropie qui est égale à 1 signifie que le système a besoin d'autres informations pour le bien classifié et diminuer son entropie et quand au 0 on dit que le système est homogène et qu'on a aboutit à une seule classe à toutes les observations du système.

L'entropie H d'un ensemble d'observations D avec un ensemble de k classes (C) se calcule de la façon suivante :

$$H(D) = -\Sigma P(C_i/D).\log_k(P(C_i/D))$$

où P :

$$P(C_i/D)=\frac{i}{n}$$

i : nombre d'observations d'une classe Ci

n : nombre d'observations totale dans D

On introduit ensuite la notion de l'entropie pondérée qui est calculée à chaque décision (division) :

$$H(D_L \vee D_R)=\frac{D_L}{D}H(D_L)+\frac{D_R}{D}H(D_R)$$

L : à gauche R : à droite

E.1.1.2- Notion de gain d'information :

Le gain d'information qui désigne la réduction de l'entropie après avoir pris une décision ou division est calculé comme suit :

$$IG(D,D_L,D_R)=H(D)-H(D_L \vee D_R)$$

Cette méthode supporte les variables quantitatives et même qualitatives (selon le type) et caractérisée par sa simplicité de compréhension vis à vis des réseaux de neurones, pas de normalisation de données (dans la phase de préparation), validation du modèle par des testes statistiques et peut être performant sur de grand jeux de données en économisant les ressources.

Néanmoins l'apprentissage de l'arbre est un problème NP-Complet et se base sur des heuristiques avec des concepts difficiles à exprimer dans le cas du sur-apprentissage.

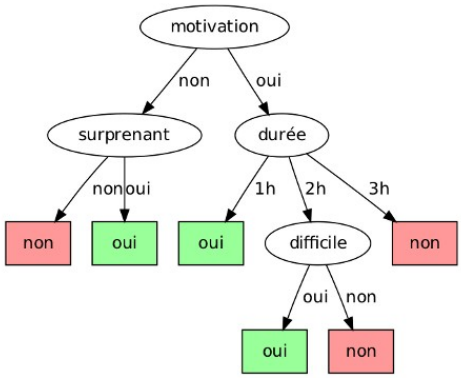


Fig 34 : Exemple Arbre de décision

E.1.1.3- Arbre de décision sous R :

E.1.1.3.1- Rpart

Les données seront stockées sous format *data.frame* (la classe standard d'un set de données sous Open Source R) et la bibliothèque qui gère les arbres de décision est '*rapart*' téléchargeable sur https://goo.gl/Gc5fdj qui gère la création des arbre de décision en en implémentant plusieurs algorithmes de création d'arbres.

Le programme de Rpart construit des modèles des classification et de régression d'une structure générale sur deux étapes; le modèle résultat peut être représenté sous forme d'un arbre binaire. Le type d'arbre que la bibliothèque traite peut être soit dédié à la décision (Oui/Non), variables continues (densité minérale), modèle de poisson ou des modèles de survie (temps jusqu'à la mort/dernière apparition). Rpart contient des outils pour modéliser, représenter et résumer problématiques de classification et de régression.

Utilisation de Rpart :

étape 1 : importer la bibliothèque
> library(rpart)

étape 2 : décider le type du modèle sortie voulue :
par catégorie ⇒ method = 'class'
continue ⇒ method = 'anova'
modèle poisson ⇒ method = 'poisson'
survie ⇒ method = 'exp'

étape 3 :
créer le modèle avec l'appel à rpart() en pensant à choisir les bonnes valeurs pour les arguments de contrôle.

étape 4 :
évaluer le modèle et passer à la prédiction.

Les options de Rpart :

la fonction principale dans Rpart est rpart() qui prend en considération la création des modèles en précisant différents arguments tel que :

formula : la formule de prédiction sous la forme Y ~ X1 + X2 +X3…
***data* :** le set d'apprentissage
***method* :** détermine le type de la règle de répartition (classification, anova, poisson ou exponentielle)
***params* :** les paramètres de la méthode choisie sous forme de liste.
***na.action* :** comment la construction du modèle va gérer les valeurs nulles.
***control* :** liste des paramètres de contrôle de la classe rpart.control qui contient les arguments suivants

- ***minsplit* :** le nombre minimum d'observations qui doivent exister dans un nœud pour en faire un split.
- ***Minbucket* :** le nombre minimum d'observations dans une feuille. Si l'un des deux (minsplit et minbucket) est spécifié on l'autre calculé de tel sorte que minsplit ← minbucket * 3.
- ***cp* :** Paramètre de complexité, tout split qui ne diminue pas l'ajustement par un facteur cp n'est pas tenté. Son rôle est d'économiser le temps de calcul en taillant les splits qui valent pas le coup.
- ***maxcompete* :** le nombre de fractionnements concurrents retenus dans la sortie, utile pour savoir qu'elle décision a été choisie et qu'elle variable a été choisie après.
- ***maxsurrogate* :** le nombre de divisions des substitutions retenues à la sortie.
- ***usesurrogate* :** comment utiliser les substitut dans le processus de fractionnement.
  0 pour afficher seulement une valeur manquante pour la régler de séparation primaire n'est pas envoyée en bas de l'arbre.
  1 signifie utiliser les substitut pour diviser les sujets manquants la variable primaire, si tous les substituts sont absents de l'observation elle ne sera pas divisée.
  2 si tous les substituts sont manquante envoyer l'observation dans la majorité.
- ***xval* :** nombre de validations croisées (méthode d'estimation de fiabilité).

E.1.1.3.2- Party

Introduit par *Hothorn* en 2006 comme amélioration de l'approche de rpart lors de la construction des arbres de décision en proposant des arbres Qui adressent l'overfitting et les variables biais de rpart en faisant appel à des variables statistiques.

Pour créer un arbre d'inférence conditionnelle sous party on fait appel à la directive :

```
> library(party)
> target <- Var_cible ~ . # variable cible en rapport avec toutes les autres
> cdt <- ctree(target, data) # data indique le nom du dataframe
> plot(cdt,type="simple")
```

E.1.1.4- Revolution R Entreprise et la classification

La bibliothèque de RevoScaleR Revolution Analytics fournit une rapide, évolutive, distribuable analyse toutes les fonctionnalités données prédictives. La fonction rxDTree inclus permet d'estimer les arbres de décision efficace sur de très grands ensembles de données. Les arbres de décision (Breiman, Friedman, Olshen, & Stone, 1984) fournissent relativement des modèles facile à interpréter, et sont largement utilisés dans une variété de disciplines. Par exemple,

- Prédire les caractéristiques des patients qui sont associés à un risque élevé, par exemple, une crise cardiaque.
- Décider ou non d'offrir un prêt à une personne sur la base de caractéristiques individuelles.
- Prédire le taux de rendement des différentes stratégies d'investissement.

La fonction rxDTree correspond aux modèles d'arbres à l'aide d'un algorithme de partitionnement récursif basé mining. Le modèle résultant est similaire à celle produite par

l'ensemble de rpart recommandé (Therneau & Atkinson, 1997). Les deux arbres de type de classification et d'arbres de type de régression sont pris en charge,

L'algorithme de rxDTree :

Les arbres de décision sont des algorithmes efficaces largement utilisés pour la classification et de régression. Les algorithmes classiques pour la construction d'un arbre de décision trient toutes les variables continues afin de décider où couper les données. Cette étape de tri devient gourmandes en temps et en mémoire lorsqu'ils traitent de grandes données. Diverses techniques ont été proposées pour surmonter l'obstacle de tri, ce qui peut être grossièrement classés en deux groupes: traitant des données pré-triées ou à l'aide des variables statistiques approximatives des données. Bien que les techniques de pré-tri suivent des algorithmes d'arbres de décision classiques de plus près, ils ne peuvent pas accueillir de très grands ensembles de données. Ces grands arbres de décision de données sont normalement parallélisés de diverses manières pour permettre un grand apprentissage à l'échelle: le parallélisme ré-partitionne les données horizontalement ou verticalement afin que les différents processeurs travaillent sur différentes observations ou différentes variables et le parallélisme des tâches construit différents nœuds d'arbres sur différents processeurs.

L'algorithme de rxDTree est un algorithme d'arbre de décision approximative avec un parallélisme de données horizontale, spécialement conçue pour la manipulation de très grands ensembles de données. Il calcule des histogrammes pour créer des fonctions de distribution empiriques des données et construit l'arbre de décision dans un mode de largeur d'abord.
L'algorithme peut être exécuté en parallèle des paramètres comme une machine multiconducteur ou un groupe (cluster ou grille) environnement distribué. Chaque travailleur reçoit seulement un sous-ensemble des observations des données, mais a une vue de l'arbre complet construit jusqu'à présent. Il construit un histogramme des observations qu'il voit, ce qui comprime essentiellement les données à une quantité fixe de mémoire. Cette description approximative des données est ensuite envoyé à un maître avec une faible constante de complexité de communication indépendante de la taille de l'ensemble de données. Le maître intègre les informations reçues de chacun des travailleurs et détermine les nœuds d'arbre terminaux pour diviser et comment. tant que l'histogramme est construit en parallèle, il peut être rapidement construit, même pour de très grands ensembles de données.

Créer un arbre simple avec rxDtree() :

```
rxArbre <- rxDTree(Cible ~ Var1 + Var2 + Var3, data = donnes, cp=0.01)
```

L'objet retourné est un objet de rxDTree de classe. La classe rxDTree, bien sûr, a des composants similaires à un objet rpart : frame, cptable, splits, etc. En fait, la fonction rxAddInheritance ajoute l'héritage rpart aux objets rxDTree.

Créer un arbre pour des données larges :

```
airlineTree <- rxDTree(CIble ~ Var1 + Var2 + Var3, data=Grandes_Donnees,
                                    blocksPerRead=30, maxDepth=5, cp=1e-5)
```

Le cp par défaut de 0 produit un très grand nombre de divisions; spécifiant cp = 1e-5 produit un ensemble plus gérable de scissions dans ce modèle.

E.1.2- Foret de décision

L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents. Le but c'est de corriger quelques inconvénients qui concernent la méthode précédente tel que le choix de l'importance des caractéristiques du système à modéliser une par rapport au autres en créant un ensemble d'arbres partiellement indépendants.
On peut exprimer le processus ainsi :
1- Créer *Q* nouveaux ensembles d'apprentissage par double échantillonnage :

- Bootstrap: sur les observations, en utilisant un tirage avec remise d'un nombre *N* d'observations identique à celui des données d'origine,
-Sur les *p* prédicateurs, en n'en retenant qu'un échantillon de cardinal $m<\sqrt{p}$ avec remise.

2- On entraîne un arbre de décision selon une des techniques connues en limitant sa croissance par validation croisée (méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage)
3- On stocke les Q prédictions de la variable cible pour chaque observation d'origine.

Ce qui rend cette méthode un simple vote majoritaire. Le principal revers de cette méthode est que l'on perd l'aspect visuel des arbres de décision uniques.
Construire un arbre de décision unique fournit un modèle simple , mais il est souvent trop simple ou trop spécifique.De nombreux modèles qui travaillent ensemble sont souvent mieux qu'un modèle qui fait tout.
Dans la construction d'un arbre de décision unique, il y a souvent de différence dans le choix entre les variables alternatives.Deux ou plusieurs variables pourraient ne pas se distinguer en termes de leur capacité à partitionner les données en des ensembles plus homogènes de données.L'algorithme de forêt aléatoire construit tous aussi de bons arbres et puis les combine en un seul modèle, ce qui entraîne une meilleure modèle globale .
Créer un modèle randomForest :

```
> library(randomForest)
> target <- Cible ~ var1 + var2 + var3
> rf <- randomForest(target, data=donnees, ntree=1000, proximity=TRUE)
```

E.1.3- Régression :

La régression est de construire une fonction des variables indépendantes (également connu en tant que prédicateurs) pour prévoir une variable dépendante (également appelée

réponse). Par exemple, les banques évaluent le risque des candidats à domicile prêt en fonction de leur âge, le revenu, les dépenses, l'occupation, le nombre de personnes à charge, la limite de crédit total, etc.
Cette partie présentera les concepts de base et présentera des exemples de diverses techniques de régression.
Dans un premier temps, il montre un exemple sur la construction d'un modèle de régression linéaire pour prédire les données. Après ça,il introduit la régression logistique. Le linéaire généralisé.

E.1.3.1- Régression linéaire :

La régression linéaire est de prédire la réponse à une fonction linéaire de valeurs prédites comme suit:

$$Y = c_0 + c_1 x_1 + c_2 x_2 + ... + c_k x_k$$

où $x_1, x_2, ..., x_k$ présentent les prédicateurs et Y est la variable à prédire.

Le modèle de régression linéaire est construit avec la fonction lm() sur des données avec des variables x1,x2, x3 et y comme suit:

```
> attach(donnees)
> reg_Mod <- lm(Y ~ x1 + x2 + x3)
```

Le modèle contient les coefficients sous forme de table reg_Mod$coefficients[[i]] pour un indice i.

E.1.3.2- Régression logistique :

La régression logistique est utilisée pour prédire la probabilité d'occurrence d'un événement en obtenant des données pour une courbe logistique. Un modèle de régression logistique est construit comme l'équation suivante:

$$logit(y) = \ln\left(\frac{y}{1-y}\right) = c_0 + c_1 x_1 + c_2 x_2 + ... + c_k x_k$$

La régression logistique est construit avec la fonction glm () en réglant l'argument family = binomial (link = "logit"):

```
> glm(formula = Y ~ x1 + x2 + x3 ,family = binomial(link = "logit"), data = donnees)
```

E.1.3.3- Régression linéaire générale :

Le modèle linéaire généralisé (GLM) généralise la régression linéaire, en permettant le modèle linéaire à être lié à la variable de réponse par l'intermédiaire d'une fonction de liaison et permettant l'amplitude de la variance de chaque mesure qu'elle soit fonction de la valeur prédite. elle unie divers autres modèles statistiques, y compris la régression linéaire, la régression logistique et la régression de Poisson. Fonction glm () est utilisée pour t modèles linéaires généralisés, en donnant une description symbolique du prédicteur linéaire et une description de la distribution des erreurs.

```
> data("bodyfat", package="mboost")
> myFormula <- DEXfat ~ age + waistcirc + hipcirc + elbowbreadth + kneebreadth
> bodyfat.glm <- glm(myFormula, family = gaussian("log"), data = bodyfat)
> summary(bodyfat.glm)
```

E.2- Modèle d'appétence : Prédire les migrations prépayé -> post payé

E.2.1- Problématique

Les offres mobiles *Méditel* sont classés en deux classes : des formules pré-payées et des formules post-payées. Le client (la ligne) peut changer de formule à un certain temps t pour une raison ou plusieurs. L'univers indéterministe des raisons de migration d'une formule à une autre fait appelle à la un modèle d'appétence pour pouvoir modéliser ce phénomène vis à vis du comportement des lignes pendant leur service. Une migration détermine à ce niveau le changement de la formule prépayée d'une ligne vers une autre post-payée sans pour une ligne active. Un ligne prépayée recharge son solde avant d'avoir droit à émettre des appels, des SMS et du data tandis qu'une ligne post-payé contient les formules classiques d'abonnement ainsi que les nouveaux forfaits.

Les techniques de modélisation utilisées sont des méthodes supervisées, basées sur un historique de 12 mois d'activité de lignes qui ont ou n'ont pas migré. L'apprentissage se fera sur un échantillon stratifié de la population, et se basera sur le comportement des lignes en prenant plusieurs variables en considération (montant mensuel rechargé, Trafic moyen national, Trafic moyen international, Data, Passes, Ancienneté, Nombre d'appelants, nombre d'appelés, Suspensions..).

Définition de la variable cible :

La cible présente la variable que l'on souhaite modéliser. Cette variable correspond à un comportement. Dans le cas de la migration, le comportement que l'on souhaite modéliser est le changement d'offre d'une ligne passant d'une formule prépayée à une formule post-payée représentée par une des 4 classes disponibles dans le parc *Méditel* :
Nom de la variable : Plan_Cible_Catego

Valeurs : ''Meditel Abonnement'', ''Pack Hany'', ''Pack Illimité'' ou ''Non migrant'' dans le cas de non migration.

Modèle	Prédiction migration Pré-payé→ Post-payé
Principes du modèle	• Arbre de décision • Régression • Foret (Random Forest)
Historique d'apprentissage	12 mois d'activité : Du 15 Mar 2014 au 16 Fev 2015
Variable à prédire	Plan_Cible_Catego • ''Meditel Abonnement'' • ''Pack Hany'' • ''Pack Illimité'' • ''Non migrant''
Variables macros de comportement	• MMPR (montant mensuel rechargé) • Ancienneté • Trafic_national_out • Trafic_intern_out • Data • Passes • Nombre d'appelants • Nombre d'appelés

E.2.2- Préparation des données

L'apprentissage du modèle se fera sur un échantillon stratifié de la population qui représente les lignes actifs sur la période entre le 15 mars 2014 et le 16 février 2015. Stratifié de tel sorte que le pourcentage de la migration dans cet échantillon sera le même pour toutes les classes de la variable cible. Les étapes de la préparation des données sont comme suit :

1. Extraction et préparation des données depuis toutes les bases concernées.
2. Création d'une table groupant les informations par numéro de la ligne.
3. Choix et création des variables d'analyse.
4. Calcule de la variable de décision de toutes la population.
5. Échantillonnage de la population.
6. Préservation des enregistrements dédiés à l'apprentissage.

Extraction et préparation des données depuis toutes les bases concernées :

Cette phase consiste à consulter le data warehouse et les datamarts concernant les informations des lignes prépayées en prenant en compte les variables d'analyse cherchées. Ces informations doivent être groupées dans une table affin de passer à la transformation des données pour les préparer à la phase analytique.
Les tables concernées : DMRT_OCT14, DMRT_OCT15, DMRT_FEV16, DMRT_AVRT15 ...

La table cible : - Migrated_Model_DATA

Création d'une table groupant les données des autres tables

La table prépayée étant prête il s'agit maintenant de voir les lignes qui ont déjà migré vers le post payés pendant les 12 mois d'analyse. Ces lignes qui seront identifiées comme étant l'intersection de cette table avec la table des lignes migrantes de la même période.
Cette étape se fera en appelant le scripte 'préparation de données.R' qui contient des fonctions d'importations qui gère le format du data warehouse et le transforme en format exploitable sous R.
Ainsi les tables en format texte plat serons pris par R comme étant des objets Data.frame avec des colonnes interprétables en R (convention des dates, lignes …). Des jointures entre les tables pour avoir les lignes de la période entière dans une seule table.

Choix et création de variables d'analyse

Les données disponibles dans les dataframes importés depuis les datamarts sont exploitables dans l'analyse en calculant des valeurs statistiques pour représenter un comportement d'une ligne.
La fonction *data_model()* du scripte 'Preparation des données.R' veille à calculer les variables suivantes :

$$MMPR = \frac{\sum_{1}^{6} RECH\ MNT\ i}{\sum_{1}^{6} RECH\ NBR\ i}$$

Montant moyen par recharge

$$OB\,CALL\ BASE\ PER\,CALL = \frac{\sum_{1}^{6} OB\,CALL\ DUR\ BASE\ i}{\sum_{1}^{6} OB\,CALL\,CNT\ BASE\ i}$$

durée moyenne d'un appel national

$$OB\,CALL\ INTR\ PER\,CALL = \frac{\sum_{1}^{6} OB\,CALL\,DUR\ INTR\ i}{\sum_{1}^{6} OB\,CALL\,CNT\ INTR\ i}$$

durée moyenne d'un appel international

$$AVG\ DATA = \frac{\sum_{1}^{6} VOLUME\,TOTAL\,NAVIGATION\ i}{6}$$

Volume moyen de data consommée

$$AVG\ APPELANTS = \frac{\sum_{1}^{6} NBRE\ DISTINCT\ APPELANTS\, i}{6}$$

Nombre moyen d'appelants

$$AVG\ APPELES = \frac{\sum_{1}^{6} NBRE\ DISTINCT\ APPELES\, i}{6}$$

Nombre moyen d'appelés

$$AVG\ PASS3 = \frac{\sum_{1}^{6} NBRE\ PASSES\, 3}{6}$$

Nombre moyen de passes *3

$$AVG\ PASS1 = \frac{\sum_{1}^{6} NBRE\ PASSES\, 1}{6}$$

Nombre moyen de passe *1

Calcule de la variable de décision de toutes la population

Pour obtenir la variable de décision il faut juste segmenter les offres cibles de migration de tel sorte à avoir les classes: 'Pack Hany', 'Meditel Abonnement' et le 'Pack Illimité'. Et pour les lignes qui n'ont pas migré on aura la classe 'Non Migrant'.
Les datamarts contient une indication aux plans cibles avec des codes qui peuvent être groupés de sorte :

```
hany<- c("417","418","408","517")
Illimite <- c("462","464","463","465","520","518","457","458","459")
Meditel_Abon <- c("359","421","425","420","360","385","361","423","362","387","365")
```

Avec ces vecteurs on segmente les lignes dans ces trois catégories et on associe 'Non Migrant' aux autres.

Échantillonnage de la population

La phase d'échantillonnage a plusieurs objectifs :
- Réduire la volumétrie pour la phase de modélisation, car le travail sur le datamart en entier nécessiterait un temps de traitement très important et des ressources énormes.
- Stratifier sur la variable cible car elle est sous représentée dans la population (25% Hany, 25% Illimité, 25% Meditel abonnement et 25% non migrants) . Et effectuer un échantillonnage non stratifié sur la cible implique que l'on aura un volume de données plus faible mais que la partition sera la même. Une manière d'avoir un bon set de données pour l'apprentissage c'est

appliquer le sur échantillonnage (over-sampling) qui consiste à sur représenter la variable cible dans l'échantillon à avoir le même pourcentage de ces deux valeurs dans l'échantillon.
- Créer un échantillon d'apprentissage et un échantillon de validation. L'échantillon d'apprentissage va permettre de construire le modèle d'appétence. Ce modèle sera appliqué pour prédire sur l'échantillon de validation affin de tester les performances du modèle dans le cadre d'une validation croisée du modèle.

Préservation des enregistrements dédiés à l'apprentissage

Pour un bon apprentissage on doit préserver les données complet pour en créer un bon modèle. Aucune valeur manquante (Na) ne devrait se trouver dans une variable dans le set d'apprentissage. La fonction complete.cases() fait une restriction sur les valeurs manquantes et ne garde que les valeurs complètes.

E.2.3- Création et déploiement des modèles

E.2.3.1- Liste des scriptes :

Preparation des donnees.R : qui groupe plusieurs fonctions pour importer et transformer les données issue des datamarts.

Les fonctions du scripte:

Fonction	Entrées / Sorties		Description
importDMRT_MMXX()	Entrées	Nom du fichier plat du datamart	Cette fonction importe le fichier plat correspondant aux données d'un datamart en modifiant les formats de lecture des colonnes pour les rendre exploitables en R R.
	Sorties	Data.frame du datamart	
categories_Migration()	Entrées	Data.frame du datamart	Cette fonction retourne les plans de migration existants dans le datamart
	Sorties	Vecteur de plans de migration	
data_model()	Entrées	Data.frame du datamart	Cette fonction calcule les colonnes dans un nouveau data.frame pour préparer les données pour la phase de la création du modèle
	Sorties	Data.frame du modèle	
remove_NA()	Entrées	Data.frame	Retourne un data.frame sans des valeurs nulles
	Sorties	Data.frame	
NonDuplicate()	Entrées	Data.frame	Retourne un data.frame qui ne contient que des lignes uniques
	Sorties	Data.frame	
nombreToLigne()	Entrées	Vecteur de lignes parsées en	Les lignes devrons être prises

Fonction	Entrées / Sorties		Description
		nombres	comme étant des chaînes de caractères.
	Sorties	Vecteur de lignes comme étant des chaînes de char	

Scripte passes1.R : Ce scripte fait preuve de fusion entre les données collectés et calculées avec les nouvelles valeurs de passes *1 et *3 issues des deux datamarts de passes.

Arbres de decisions.R : qui gère les étapes de la création des arbres de décisions et les données auxquelles elle serons basée sur. Ce scripte présente 3 solutions d'arbres de différentes bibliothèques :

- rpart() de la bibliothèque Rpart, arbre de décision classique sous Open source R.
- ctree() de la bibliothèque Party, arbre d'inférence conditionnelle.
- rxDtree() de la bibliothèque ScaleR, arbre crée par calcule multi-cœur et traitement par bloques.

Regression.R : ce scripte gère le calcule des régressions des classes en les traitant une par une avec une règle de l'un contre les autres (one vs all), qui va calculer la régression logistique (entre 0 et 1) de l'affectation d'une ligne vers une classe 1 ou les autres restantes 0 pour chacune des classes existantes.

RandomForest.R : ce scripte gère la création du random forest pour prédire la migration de la population en utilisant la bibliothèque RandomFrorest.

E.2.3.2- Scoring des modèles :

La partie scoring tourne sur le teste de la validité de ces modèles et savoir les performances, calculer la fiabilité du modèles et finaliser leur paramétrages pour enfin produire la version finale et l'introduire dans le milieu de production. Le scoring des modèles suivra des modèles de teste mathématiques et aussi d'autres variables décisionnelles prisent au sain de l'entreprise pour prévoir le déploiement, la mise à jour et l'expiration des modèles.

E.2.3.2.1- Tests de validité :

E.2.3.2.1.a- Arbres de décision :

Les arbres de décision créés pour représenter le modèle seront analysé pour obtenir la précision maximale concernant la prédiction. Le set de données de validation représente 30% des données stratifiés du modèle (dont 70% était pour faire l'apprentissage). Le paramétrage des arbres portera sur les paramètres de création tel que le facteur de complexité, la taille de l'arbre la méthode de création vis à vis l'erreur de prédiction dans le processus de la validation croisée.

Arbre Rpart :

Le premier arbre de décision fait partie de la bibliothèque rpart mentionnée précédemment avec des paramètres de contrôle qui fixent la profondeur à 20 niveau, et le paramètre de complexité à 0,006 (pour avoir une analyse de la totalité des variables) :

```
Arbre_rpart <- rpart( target, data=trainData, method = "class",
                                    control = rpart.control(maxdepth = 20,cp = 0.006),
                                    parms = list(split = 'information'))
```

	CP	nsplit	rel error	xerror	xstd
1	0.2042203	0	1.00000	1.03255	0.010750
2	0.0432761	1	0.79578	0.79614	0.011572
3	0.0162732	2	0.75250	0.75608	0.011592
4	0.0064378	4	0.71996	0.72389	0.011584
5	0.0060000	5	0.71352	0.72210	0.011583

Ce tableau présente les données de validation croisée à chaque Split ainsi que la valeur de l'erreur de la validation croisée (xerror). On peut voir la dégradation de cette erreur par rapport à la complexité et la taille de l'arbre avec la fonction plotcp() qui prend en argument le modèle en question :

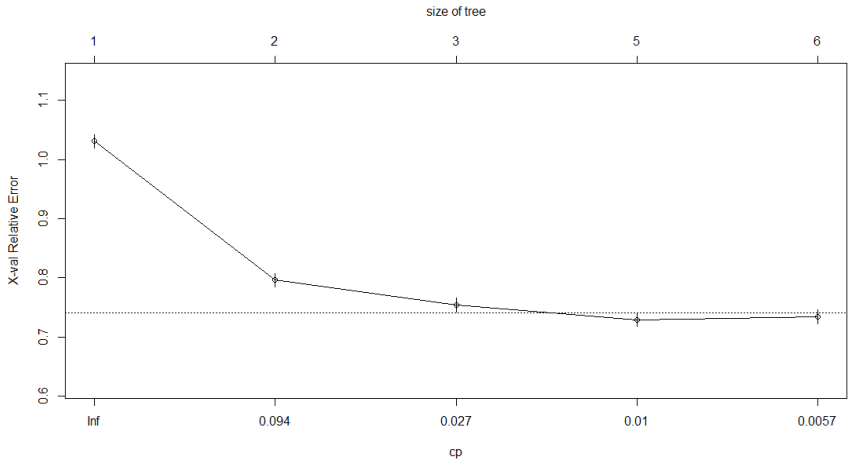


Fig 35 : évolution de l'erreur avec la complexité (rpart)

La dégradation de l'erreur avec cette valeur de complexité n'est pas si satisfaisante on doit alors changer de complexité et pour avoir un retour mesurable calculons l'erreur de prédiction.

On va prédire l'orientation de migration des lignes dans le set de données de test, ce set comprend 2882 lignes distribuées en réalité comme suit :

Méditel Abonnement	Pack Hany	Pack Illimité	Non Migrant
603	601	594	1084

La prédiction donne le résultat suivant :

Méditel Abonnement	Pack Hany	Pack Illimité	Non Migrant
631	571	676	1004

Si on croise les deux tables :

prédite \ réelle	Méditel Abonnement	Pack Hany	Pack Illimité	Non Migrant
Méditel Abonnement	363	113	135	79
Pack Hany	149	241	131	165
Pack Illimité	91	247	328	123
Non Migrant	120	127	154	316

L'erreur de prédiction : 1 – somme(diagonale)/somme_tableau = 0,48 (48 % d'erreur)

cette erreur est très marquante ce qui rend le modèle inexploitable.

→ ce qu'il faut faire c'est essayer de varier la complexité d'un dixième de sa valeur actuelle et de recalculer l'erreur de prédiction par la même formule :

CP	Erreur de prédiction
0,005	47 %
0,001	34 %
0,0001	28 %
0,00001	28 %
0,000001	28 %

du coup on changera la valeur de complexité de ce modèle vers 0,0001 comme valeur adéquate avec le minimum d'erreur.

E.2.3.2.1.b-Régression :

Le scoring des régressions portera sur l'analyse des coefficients de chaque modèle propre à sa catégorie (Illimité, Hany, Abonnement, Nom Migrant) comme l'a prescrit la méthode de l'une contre les autres. Le déploiement de ces modèles est de tel sorte à présenter le résultat de la prédiction sous forme de 4 nouvelles colonnes qui décrivent chacune la probabilité de la migration de la ligne vers l'offre indiquée.

b.1 - Modèle RxLogit :

Classe concernée	Variable indépendante	Coefficient
Méditel Abonnement	ANCIENNETE	0,0001355489
	MMPR	-0,0044841246
	AVG_OB_INT	0,0007583653
	AVG_OB_BASE	0,0010809342
	AVG_DATA_Go	0,1686313480
	AVG_APPELANTS	-0,0023171501
	AVG_APPELES	0,0220431006
	AVG_PASSE1_MENS	-0,0560314201
	(Intercept)	-3,5929257182
Méditel Illimité	ANCIENNETE	-0,0001052195
	MMPR	0,0038416727
	AVG_OB_INT	0,0003222862
	AVG_OB_BASE	-0,0005119443
	AVG_DATA_Go	0,1322823341
	AVG_APPELANTS	0,0227380520
	AVG_APPELES	-0,0119253114
	AVG_PASSE1_MENS	0,0111035928
	(Intercept)	-1,7137937768
Méditel Hany	ANCIENNETE	0,000065189
	MMPR	-0,002767489
	AVG_OB_INT	-0,000428853
	AVG_OB_BASE	0,0001416093
	AVG_DATA_Go	-0,1685297
	AVG_APPELANTS	-0,01986760
	AVG_APPELES	0,004231080
	AVG_PASSE1_MENS	-0,0001982734
	(Intercept)	1,589154
Non Migrant	ANCIENNETE	0,00020860
	MMPR	-0,00230624
	AVG_OB_INT	0,000321639
	AVG_OB_BASE	0,000138771
	AVG_DATA_Go	-0,13482376
	AVG_APPELANTS	-0,04569548
	AVG_APPELES	0,005077296

Classe concernée	Variable indépendante	Coefficient
	AVG_PASSE1_MENS	0,000134825
	(Intercept)	1,1918655

b.1 – Modèle linéaire généralisé :

Classe concernée	Variable indépendante	Coefficient
Méditel Abonnement	ANCIENNETE	0.0001355
	MMPR	-0.0044841
	AVG_OB_INT	0.0007584
	AVG_OB_BASE	0.0010809
	AVG_DATA_Go	0.1686313
	AVG_APPELANTS	-0.0023172
	AVG_APPELES	0.0220431
	AVG_PASSE1_MENS	-0.0560314
	(Intercept)	-3.5929257
Méditel Illimité	ANCIENNETE	-0.0001052
	MMPR	0.0038417
	AVG_OB_INT	0.0003223
	AVG_OB_BASE	-0.0005119
	AVG_DATA_Go	0.1322823
	AVG_APPELANTS	0.0227381
	AVG_APPELES	-0.0119253
	AVG_PASSE1_MENS	0.0111036
	(Intercept)	-1.7137938
Méditel Hany	ANCIENNETE	6.519e-05
	MMPR	-2.767e-03
	AVG_OB_INT	-4.829e-04
	AVG_OB_BASE	1.416e-04
	AVG_DATA_Go	-1.685e-01
	AVG_APPELANTS	-1.987e-02
	AVG_APPELES	4.231e-03
	AVG_PASSE1_MENS	-1.983e-04
	(Intercept)	1.589e+00
Non Migrant	ANCIENNETE	0,00027860
	MMPR	-0,00230924
	AVG_OB_INT	0,000221639
	AVG_OB_BASE	0,000158771
	AVG_DATA_Go	-0,13452376
	AVG_APPELANTS	-0,03569548
	AVG_APPELES	0,005457296
	AVG_PASSE1_MENS	0,000136825
	(Intercept)	1,1458655

Intercept désigne la constante de régression, la variable la plus signifiante pour chaque régression c'est elle qui a la plus grande valeur. Si le coefficient est négatif, ça veut dire que si la valeur de cet indicateur(variable) diminue la ligne aura plus de chance à être classifiée selon le plan.

E.2.3.2.1.c- RandomForest :

La foret de décision pour ce modèle comportera 500 arbres comme paramètre de lancement, le nombre de variable prise pour chaque arbre sera le plus proche de la racine carrée du nombre totale des variable par défaut, et c'est la valeur configurable dans le cas d'une correction du modèle. Le modèle prototype comporte :

nombre d'arbres : 500 nombre de variables pour chaque arbre : 2

La validation croisée à la création donne le taux d'erreur avec 6231 lignes:

prédite \ réelle	Méditel Abonnement	Pack Hany	Pack Illimité	Non Migrant
Méditel Abonnement	900	278	219	224
Pack Hany	323	673	402	403
Pack Illimité	257	423	726	124
Non Migrant	192	227	245	615

Erreur globale de 53,23 %

Tant que le modèle ne dépend que du choix des variable qui est aléatoire avec remise à chaque fois que la foret doit héberger un nouvel arbre. Le modèle suggère chercher une valeur près de la racine carrée du nombre totale des variables dans notre cas c'est 3 on essayera d'augmenter ce nombre pour atteindre un taux minimum d'erreurs.

Nombre de variables	Erreur
2	53,23%
4	40,15%
5	20,02%
6	22,2%

Une valeur optimale de 5 variables par arbre, et on remarque que l'algorithme en comparaison avec celui qui propose un seul arbre marque un taux d'erreurs plus réduit.

E.2.3.2.2- Rafraîchissement des modèles :

L'évolution du marché de télécommunications est très rapide. Le comportement des clients est très volatile. Cela a un impact direct sur la stabilité dans le temps des modèles prédictifs dans le temps.

Il est recommandé de re modéliser environ tous les 3 ou 4 mois sous plusieurs contraintes (apparition de nouvelles offres, apparition de nouvelles formules ...)

La validité du modèle pourra être calculée efficacement en évaluant l'écart entre les lignes prédites migrante avec leurs état au futur proche (impliquer un nouveau modèle de prédiction de temps t après un événement).

E.2.3.2.3- Industrialisation :

L'industrialisation est à la charge de la DSI, qui aura la tache de déployer les modèles et de créer les datamarts correspondants.
Les utilisateurs accéderont aux datamarts à distance soit pour exploiter en R soit avec SAS Entreprise Guide. Le datamart de migration sera nommé : **DMRT_Mig_MMAA**, désignant le mois et l'année où le score a été créé, par exemple **DMRT_Mig_AOU16** désignera le score basée sur le datamart des lignes prépayée de août 2016.
Les datamarts peuvent au choix être déployés pour chaque mois ou grouper l'analyse sur 3 mois.

Partie F : <Conclusions et discussions>

Ce rapport est clôturé par ce chapitre qui décrit les ajouts que le projet a pu apporter ainsi que les perspectives à envisager pour améliorer la performance du service.

F- Conclusion et Discutions

Ce projet de fin d'études a été pour moi non seulement une étape à franchir pour achever mon parcours académique, mais aussi une grande opportunité de prendre part à un projet réel qui s'inscrit dans le cadre de l'informatique décisionnelle. Cette dernière a pour objectif d'offrir un meilleur usage du flot de données afin d'aider les décideurs à la prise de décision.

Nous rappelons que le stage qui s'est déroulé au sein de *Méditel* consistait en l'étude de la migration potentielle de technologie remplaçant SAS avec R comme étant outil Open Source de datamining. Dans l'approche analytique et au niveau du benchmarking j'étais mené à comparer les outils open source avec R inclus, ensuite comparer le travail avec R contre celui de SAS pour en tirer les points en communs qui faciliterons la migration de technologie. Ensuite énumérer plusieurs cas qui permettrons une migration fluide à coût réduit en prenant en considération plusieurs aspects fonctionnels et architecturales afin d'en tirer le scénario d'intégration le plus convenable pour être adapté.

Pour illustrer le travail du datamining sous R, un modèle d'appétence est mis en application suivant différentes méthodes de prédiction tel que les arbres de décisions, les régressions et les foret de décision avec un comparatif de crédibilité et de stabilité de chaque modèle vis à vis de l'autre, pour illustrer un phénomène répondu chez les clients de *Méditel* qui est la migration des lignes prépayées vers le post-payé qui réside un phénomène décisif sur l'opérateur pour gérer et mieux se rapprocher des ses clients.

En se basant sur ce projet, il est possible de formuler quelques recommandations pour mieux gérer la transition entre les deux outils. La première recommandations est de bien se renseigner sur les notions basiques de la programmation sous R pour bien gérer les taches minimales. Deuxième recommandation réside à penser à la création d'une solution basée sur R propre à *Méditel* personnalisée au travail dédiée aux demandes de l'opérateur, cette idée peut se rendre rentable si l'outil développé peut être encapsulé pour être commercialisé comme étant outil de datamining propre aux opérateurs. La troisième est d'adapter les ressources techniques en renouvelant l'arsenal informatique au niveau des serveurs pour gagner du temps que prennent les opérations datamining intenses.

D'un point de vue personnel, j'ai pu découvrir le domaine du business intelligence ainsi que le métier Télécom. Par ailleurs, j'ai eu l'occasion d'observer et de m'impliquer dans la mise en place d'un système prédictif et sentir son impact décisif sur l'entreprise.

Bibliographies :

Business Intelligence in Telecoms Industry
Tanko Ishaya and Musiliudeen Folarin
The University of Hull, Scarborough Campus
United Kindom

R and Datamining: Examples and Case studies
Yanchang Zhao
published by Elservier

RevoScaleR 7.0 User Guide
Revolution analytics

International Journal of Advanced Research in Comuter Science and Software Engineering
issue #6 :Comparative Study of Data Mining Tools
Kalpana Rangra and Dr. K.L Bansal

Split Selection Methods For Classification Trees
Wei-Yin Loh and Yu-Shan Shih
University of Wisconsin-Madison and National Chun Cheng University

Big Data Decision Trees with R
Richard Calaway, Lee Edlefsen, and Lixin Gong
Revolution analytics

An Introduction to Recursive Partitioning Using Rpart Routines
E.J Atkinson, T.M Theneau and Mayo Foundation

Webographies :

Site officiel de Méditel : http://www.meditel.ma
Site officiel de l'ANRT : http://www.anrt.ma
Blog : http://www.r-bloggers.com/

Annexes

Outils du Service BI

Le service BI utilise un ensemble d'outils informatique pour extraire les données, traiter, analyser et présenter ; On cite ici :

Business Objects :

Business Objects est un éditeur de logiciels ou progiciels dans le domaine de l'intelligence économique, comme le benchmarking,le reporting, les entrepôts de données, l'ETL et le data mining.
C'est un éditeur international de logiciels d'informatique décisionnelle (ou business intelligence) principalement connu pour son outil de construction de requêtes et de rapports d'analyse ou tableaux de bord qui utilise des univers, des vues métier sur les données des entrepôts.
BO, consiste à outil de base dans le service, il permet d'accéder à la base de données entreprise à travers des univers et des vues prédéfinis pour effectuer des requêtes et extraire les données nécessaires.

SAS entreprise Guide :

SAS Enterprise Guide est une interface graphique pour Windows permettant de gérer le développement et l'exécution de programmes SAS, en en demandant l'exécution sur un serveur SAS qui peut être :

Soit local, c'est-à-dire, la station PC où SAS Enterprise Guide s'exécute. Il faut dans ce cas avoir installé SAS 9.1.3 SP4 ou SAS 9.3 sur son PC.

Soit sur un serveur (c'est l'utilisation présentée ci-dessous). Il faut dans ce cas disposer d'un login sur ce serveur, mais l'installation du logiciel SAS n'est pas nécessaire sur votre ordinateur. Les résultats d'exécution s'afficheront dans une fenêtre Windows, y compris les résultats graphiques.

L'utilisation de cet application au service business intelligence, consiste a traiter les données de grande taille et de faire des jointures qui ne sont pas possibles sur d'autres applications comme Excel.

Microsoft Excel :

Microsoft Excel est un logiciel tableur de la suite bureautique Microsoft Office,développée et distribuée par l'éditeur Microsoft. La version la plus récente est Excel 2013.

Cette application est un tableur; autrement dit, elle se présente sous forme de tableaux structurés en lignes et colonnes dans des onglets séparés avec, pour chaque cellule qui compose chaque feuille, des caractéristiques particulières pour les calculs, des outils de génération de graphiques, des outils d'analyse croisée dynamique et un module de programmation par macro ou en développement direct avec le langage Visual Basic pour Application (VBA).
Excel est utilisé au service pour rédiger et présenter des rapports journaliers, hebdomadaires ou mensuels.

Microsoft PowerPoint :

Microsoft PowerPoint est un logiciel de présentation édité par Microsoft. Il fait partie de la suite Microsoft Office. Microsoft PowerPoint fonctionne sous Windows et Mac OS.

Dans PowerPoint, ainsi que dans la plupart des logiciels de présentation, les textes, images,vidéos et autres objets sont positionnés sur des pages individuelles, les « slides » (on parle aussi de diapositives, de diapos ou de planches).

Power point consiste un outil de base pour le service business intelligence, voir pour Méditel, il est utilisé pour présenter le rapport de manière efficace et conviviale.

Think-Cell :

Think-Cell, est un add-on à installer sur office. Il a pour but de faire le lien entre Excel, blindé de données, de formules et tableaux divers et un PowerPoint, qui est la synthèse graphique avec juste quelques explications et quelques graphiques pour nos décideurs.
Les fonctions avancées de l'outil permettent de créer des présentations dynamiques (reliées à Excel) et des présentations impressionnantes.