

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
BUSINESS INTELLIGENCE MASTER

RICARDO LEMOS CAMPOS

**APLICAÇÃO DE DATA MINING PARA AUMENTO DE INVESTIMENTOS EM
PLANOS DE PREVIDÊNCIA COMPLEMENTAR**

Rio de Janeiro

2021

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
BUSINESS INTELLIGENCE MASTER

RICARDO LEMOS CAMPOS

**APLICAÇÃO DE DATA MINING PARA AUMENTO DE INVESTIMENTOS EM
PLANOS DE PREVIDÊNCIA COMPLEMENTAR**

Monografia apresentada ao Departamento de Engenharia Elétrica da PUC-Rio como parte dos requisitos para a obtenção do título de Pós-Graduação no Curso *Business Intelligence Master*.

ORIENTADORES:

FELIPE BORGES

MANOELA KOHLER

Rio de Janeiro

2021

Resumo

O presente trabalho consiste na realização de um estudo, com aplicação de técnicas de *Data Mining*, voltado para o aumento de investimentos dos participantes de planos de benefícios de previdência complementar. Com o aumento da concorrência no mercado de previdência com as entidades abertas e fechadas, as entidades de previdência fechada constantemente vêm buscando alternativas para redução de custos e aumento de receitas. Seja através de maior eficiência operacional, aumento do número de participantes ou aumento do valor de arrecadação, o mercado de previdência complementar está inovando e saindo da inércia antes associada ao conservadorismo. Com o recente lançamento de seu primeiro plano instituído, a Valia está buscando novos participantes e o aumento de investimentos nos planos que administra, e uma alternativa mais assertiva é a utilização dos dados para atingir este objetivo.

SUMÁRIO

1. Introdução	5
1.1 Conhecendo a Valia	5
1.2 Nossos números	5
1.3 Objetivo do Trabalho	6
1.4 Descrição do Trabalho	6
2.Entendendo o Problema/Necessidade	6
2.1 Aumento de Investimentos	7
3.Coleta de dados	8
3.1 Atributos	8
4.Tratamento de dados	10
5.Clusterização 1	10
5.1 Análise Exploratória e Seleção Atributos	11
5.2 Modelagem de dados	13
5.3 Resultados	14
6.Clusterização 2	17
6.1 Análise Exploratória e Seleção Atributos	17
6.2 Modelagem de dados	19
6.3 Resultados	20
7.Clusterização 3	23
7.1 Análise Exploratória e Seleção Atributos	23
7.2 Modelagem de dados	26
7.3 Resultados	26
8.Glossário	29
9.Referências Bibliográficas	32

1. Introdução

1.1 Conhecendo a Valia

A Fundação Vale do Rio Doce de Seguridade Social – VALIA, cuja Instituidora e Patrocinadora é a Vale S.A., denominada Instituidora, é, nos termos da Lei Complementar nº 109, de 29 de maio de 2001, uma entidade fechada de previdência complementar, sem fins lucrativos, com autonomia administrativa e financeira e iniciou suas atividades em setembro de 1973.

A VALIA tem por objeto instituir e executar planos de benefícios de caráter previdenciário e privado, concedendo benefícios suplementares ou assemelhados aos da Previdência Social, pecúlios ou rendas.

A VALIA é uma entidade multipatrocinada, com multiplano, administrando planos de benefícios com independência patrimonial.

1.2 Nossos números:

8 planos de benefícios, sendo 4 disponíveis para adesão e 4 fechados para novas inscrições.

51 patrocinadores

99.287 participantes ativos

25.589 participantes assistidos

209 empregados

R\$ 25,4 bi patrimônio

Base: abril/2021

1.3 Objetivo do Trabalho

O Trabalho tem como objetivo entender o problema/necessidade identificado pelas áreas de negócios da Valia e propor uma solução baseada em um projeto de Data Mining.

1.4 Descrição do Trabalho

O presente trabalho foi dividido em três modelos de Clusterização, uma para cada tipo de aumento de investimento a ser estudado.

Foram desenvolvidas as 06 fases de um projeto de Data Mining, as fases de 1. Entendimento do Problema, 2. Coleta de Dados e 3. Tratamento de Dados, foram desenvolvidas em conjunto para os três modelos. Já as demais fases: 4. Análise Exploratória, 5. Modelagem de Dados e 6. Resultados, foram desenvolvidas individualmente dentro de cada modelo de Clusterização.

As ferramentas utilizadas para desenvolvimento do trabalho foram o RapidMiner e o Microsoft Excel, ambas com suas respectivas licenças de estudante fornecida pela PUC-RIO.

2. Entendendo o Problema / Necessidade

Como o Propósito da Valia consiste em ajudar as pessoas na construção de um futuro mais digno e sustentável, temos a necessidade de fomentar a participação dos empregados das empresas patrocinadoras em seus respectivos planos de benefícios e uma vez inscritos nos planos, **aumentar seus investimentos** o que irá garantir uma maior renda futura para cada participante no momento de sua aposentadoria.

Como o percentual de adesão aos planos patrocinados da Valia supera 98% de forma histórica, atualmente o desafio relacionado à adesão de novos participantes se dá para o Plano Família (Prevaler), criado em outubro de 2020, para atender o desejo dos próprios participantes em ter um plano da Valia onde seus familiares pudessem aderir.

Além da adesão de novos participantes ao Plano Prevaler, a Valia também busca aumento de investimentos nos planos patrocinados, através do aumento dos valores de contribuições, que pode ser via aumento do percentual de contribuição ordinária ou contribuição esporádica.

2.1 Aumento de Investimentos

Atualmente existem três formas de o participante aumentar seu investimento na Valia, são elas:

- a) **Aumentar o percentual de contribuição normal ordinária mensal:** A contribuição normal ordinária mensal do participante contribuinte ativo e do autopatrocinado será de 1% (um por cento) da parcela do salário-de-participação, até o limite de 10 (dez) UR, acrescida de no mínimo 1% (um por cento) da parcela do salário-de-participação que exceder 10 (dez) UR. A contribuição do patrocinador não poderá exceder a aplicação do percentual de 9% (nove por cento) sobre a parcela do salário-de-participação que ultrapassar a 10 (dez) UR. Desta forma, o participante poderá optar por percentuais de contribuição entre 1 e 100% da parcela do salário-de-participação que ultrapassar a 10 (dez) UR, sendo que a contribuição da empresa patrocinadora irá acompanhar até o percentual de 9% desta mesma parcela.
- b) **Realização de contribuição normal esporádica:** a contribuição normal esporádica não é obrigatória e pode ser realizada pelo participante através de desconto em folha da empresa patrocinadora (folha) ou através do pagamento de boleto bancário (avulsa). As contribuições normais ordinárias e esporádicas realizadas pelos Participantes serão creditadas na **Conta de Participante** e as contribuições realizadas pela empresa patrocinadora serão creditadas e acumuladas na **Conta de Patrocinador**.
- c) **Inscrição e manutenção de contribuições no Plano Prevaler:** O Plano Prevaler está na categoria de plano de previdência instituído setorial, que aumenta ainda mais o número de pessoas com acesso à Previdência Fechada e possui as vantagens dos investimentos no setor previdenciário. o Plano Prevaler possui características diferentes dos demais planos patrocinados da Valia trazendo maior flexibilidade e não está vinculado a vida funcional do empregado/participante. A contribuição mensal não é obrigatória, o valor mínimo da contribuição é de R\$ 50,00, que é chamada Contribuição Básica e este valor corresponde a 20% da Unidade de Referência (UR), que é

atualmente R\$ 250,00. Permite a inscrição de Participantes e Assistidos de outros planos previdenciários administrados pela Valia, bem como seus dependentes e familiares, inclusive parentes por afinidade, como cônjuge, sogros, genros, noras e cunhados não perdem a condição de participantes do plano em caso de um eventual rompimento do vínculo familiar. O Plano Prevaler também permite mais de uma inscrição Ativa por participante, o que não é possível nos outros planos previdenciários administrados pela Valia.

Com a definição da necessidade de aumento de investimentos, o problema a ser resolvido foi classificado como um problema de Clusterização, uma vez que a Valia precisa identificar seus grupos de participantes, separados por tipo de investimento. Sendo assim, o trabalho será dividido em três partes, com um modelo diferente para cada tipo de aumento de investimento.

3. Coleta de dados

Uma vez entendida a necessidade da área de negócio, foram identificados alguns atributos considerados como mais relevantes na base de dados da Valia, como entradas dos modelos de Clusterização a serem desenvolvidos nas etapas deste trabalho.

Para realização do estudo, foram considerados os participantes Ativos nos planos de benefícios patrocinados, que podem realizar aumento de investimentos.

A base de dados utilizada para o estudo possui 90.605 registros e 22 atributos listados a seguir e foi extraída do ERP SIS Amadeus, dos módulos Corporativo e Capitalização.

3.1 Atributos

ID – Identificador do registro do empregado/participante.

Cod Emprs – código da empresa patrocinadora registrado no sistema.

Cod Sexo Emprg – sexo do empregado/participante.

Cod Estcv Emprg – estado civil do empregado/participante.

Cod Tppcp – tipo de participação do empregado/participante no plano de benefícios.

Cod Sitpar - situação do empregado/participante no plano de benefícios.

Vlr Idade Emprg - idade do empregado/participante em anos.

Num Plbnf - código do plano de benefícios registrado no sistema.

Qde Tempo Vinc Fund – tempo de filiação do empregado/participante ao plano de benefícios em anos.

Cod Prfinv – código da opção de investimentos do empregado/participante no plano de benefícios.

Qde Alter Perfil Inv – número de alterações da opção de investimentos do empregado/participante registrada no sistema.

Qde Filhos – quantidade de filhos do empregado/participante.

Qde Dependentes - quantidade de dependentes do empregado/participante.

Max Vlr Salar Emprg – valor do salário nominal do empregado/participante.

Pct Ctb Pctipr – percentual de contribuição ordinária mensal do empregado/participante.

Qde Alter Pct Ctb Pctipr – número de alterações de percentual de contribuição ordinária mensal do empregado/participante.

Vlr Espor Atv Folha – valor da contribuição esporádica mensal do empregado/participante ativa na folha de pagamento da empresa patrocinadora.

Qde Espor Folha hist – número de contribuições esporádicas mensais do empregado/participante registradas no histórico de contribuições.

Sum Vlr Espor Avulsa Hist – soma dos valores das contribuições esporádicas avulsas do empregado/participante registradas no histórico de contribuições.

Qde Espor Avulsa Hist – número de contribuições esporádicas avulsas do empregado/participante registradas no histórico de contribuições.

Adesao Prevaler – informação de adesão ao plano Prevaler associada ao CPF do empregado/participante do plano patrocinado.

Qde Adesao Prevaler – número de adesões ao plano Prevaler associadas ao CPF do empregado/participante do plano patrocinado.

Os arquivos foram extraídos em arquivo texto (.txt) e convertidos para Excel (.xlsx), seus respectivos scripts de extração serão anexados no repositório do GitHub, conforme orientação da coordenação do Curso BI Master, não será disponibilizada a base de dados, somente os modelos utilizados e seus respectivos resultados.

4. Tratamento de dados

Nesta etapa foi tratada toda a base de dados extraída que será utilizada no estudo, totalizando 90.815 registros e 22 atributos. Os registros identificados na análise como *outliers*, eram correspondentes aos executivos das patrocinadoras e optou-se pela permanência dos registros na base de dados. Todos os registros foram anonimizados para garantir a segurança dos dados pessoais e das informações dos participantes. Foram identificadas algumas inconsistências, tais como:

- a) Datas inválidas e faltantes (data de nascimento, data de inscrição, data de admissão, data de desligamento);
- b) Documento CPF inválido ou faltante;
- c) Sexo do empregado/participante não informado;
- d) Nome do empregado/participante faltante;
- e) Salário Nominal do empregado/participante faltante;

Como a base de dados utilizada para o estudo possui uma grande quantidade de registros (90.815 registros) e a quantidade de inconsistências foi inexpressiva (210 registros), o que corresponde a 0,23%, optou-se pela exclusão destes registros. Todas as inconsistências identificadas foram encaminhadas para a área de negócios responsável, para análise e realização de saneamento de dados.

Para cada modelo de Clusterização foram criadas duas bases de dados, dividindo os 90.605 registros. Em uma das bases de dados foram utilizados os registros dos participantes que contém o tipo de investimento a ser avaliado e uma outra base de dados com os registros dos participantes que não contém o mesmo tipo de investimento. Como serão desenvolvidos 3 modelos, no total temos 6 bases de dados.

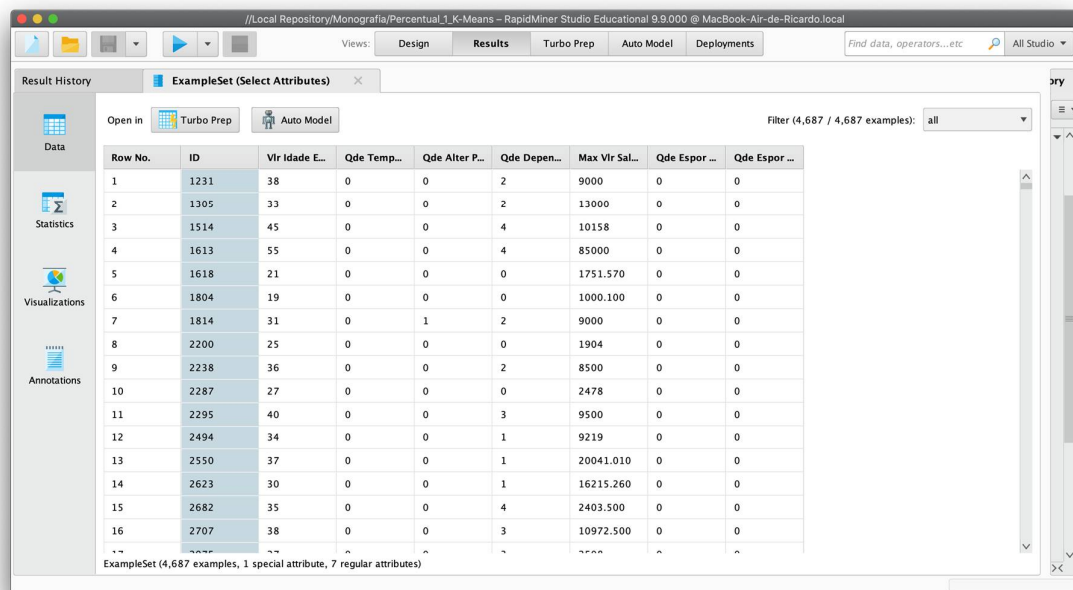
5. Clusterização 1: Aumento do percentual de contribuição normal ordinária mensal

Neste modelo de clusterização, será analisado o público de participantes Ativos nos planos de benefícios patrocinados administrados pela Valia. A base de dados utilizada neste modelo é formada pelos participantes que possuem percentual de contribuição normal ordinária mensal maior que 9%, isto é, os participantes com nível de contribuição acima do percentual acompanhado pela contribuição da sua Patrocinadora.

5.1 Análise Exploratória e Seleção Atributos

Foram escolhidos os atributos mais relevantes, considerando a interação dos participantes com a Valia e relacionados com o aumento de investimento, além dos atributos correlacionados com estes. Para a análise do aumento do percentual de contribuição normal ordinária mensal, os atributos referentes a este tipo de investimento foram retirados da base.

Desta forma, para esta análise foram considerados 4.687 registros, 01 atributo especial (ID) e 07 atributos regulares (figura 1).



Row No.	ID	Vlr Idade E...	Qde Temp...	Qde Alter P...	Qde Depen...	Max Vlr Sal...	Qde Espor ...	Qde Espor ...
1	1231	38	0	0	2	9000	0	0
2	1305	33	0	0	2	13000	0	0
3	1514	45	0	0	4	10158	0	0
4	1613	55	0	0	4	85000	0	0
5	1618	21	0	0	0	1751.570	0	0
6	1804	19	0	0	0	1000.100	0	0
7	1814	31	0	1	2	9000	0	0
8	2200	25	0	0	0	1904	0	0
9	2238	36	0	0	2	8500	0	0
10	2287	27	0	0	0	2478	0	0
11	2295	40	0	0	3	9500	0	0
12	2494	34	0	0	1	9219	0	0
13	2550	37	0	0	1	20041.010	0	0
14	2623	30	0	0	1	16215.260	0	0
15	2682	35	0	0	4	2403.500	0	0
16	2707	38	0	0	3	10972.500	0	0

Figura 1 – RapidMiner: Base Percentual Carregada

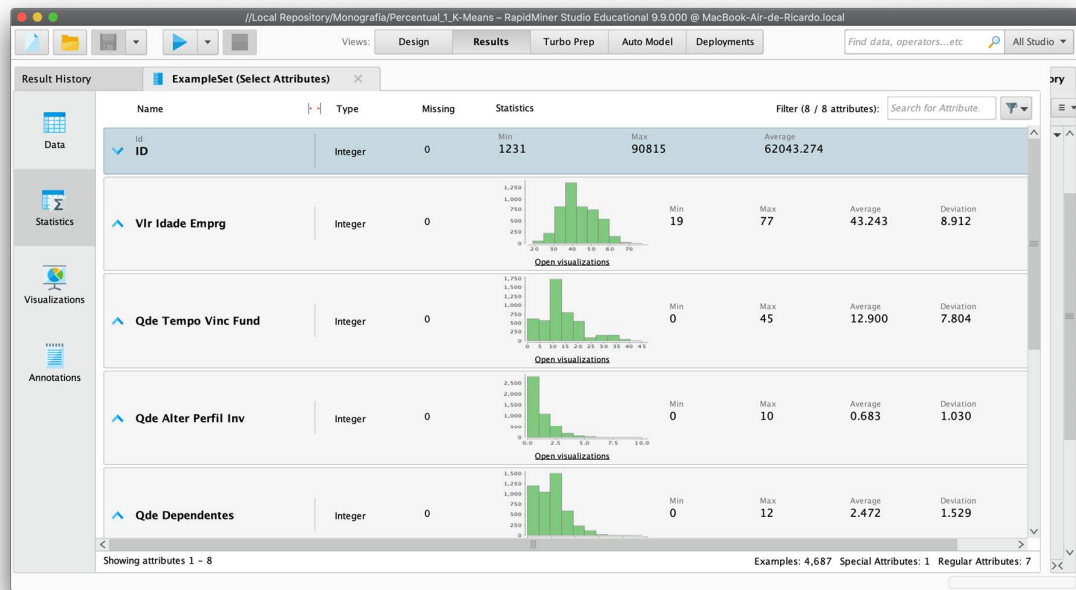


Figura 2 - RapidMiner: Base Percentual Estatísticas

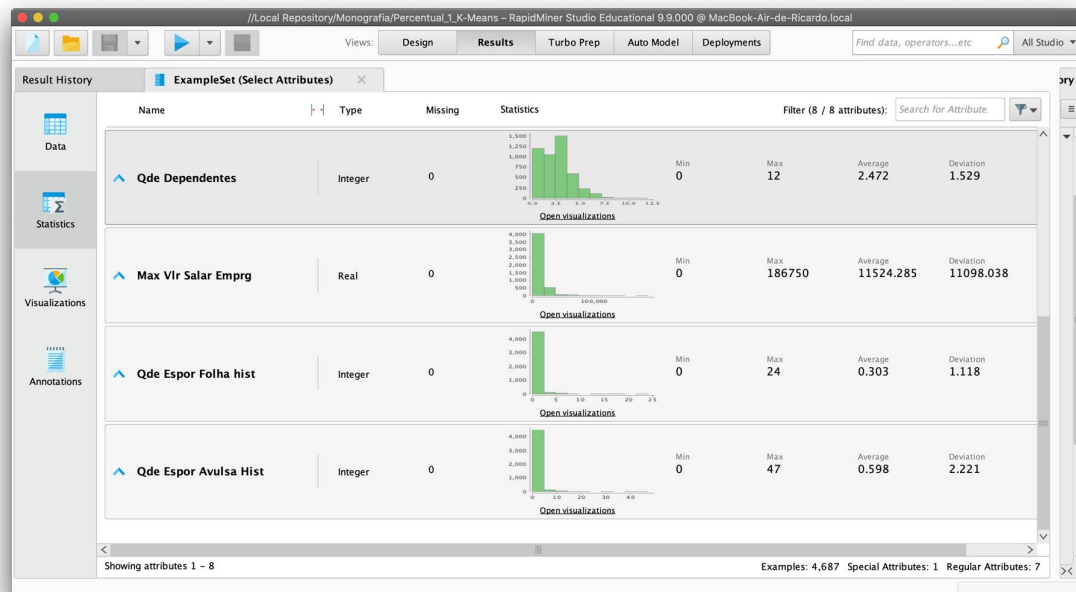


Figura 3 - RapidMiner: Base Percentual Estatísticas

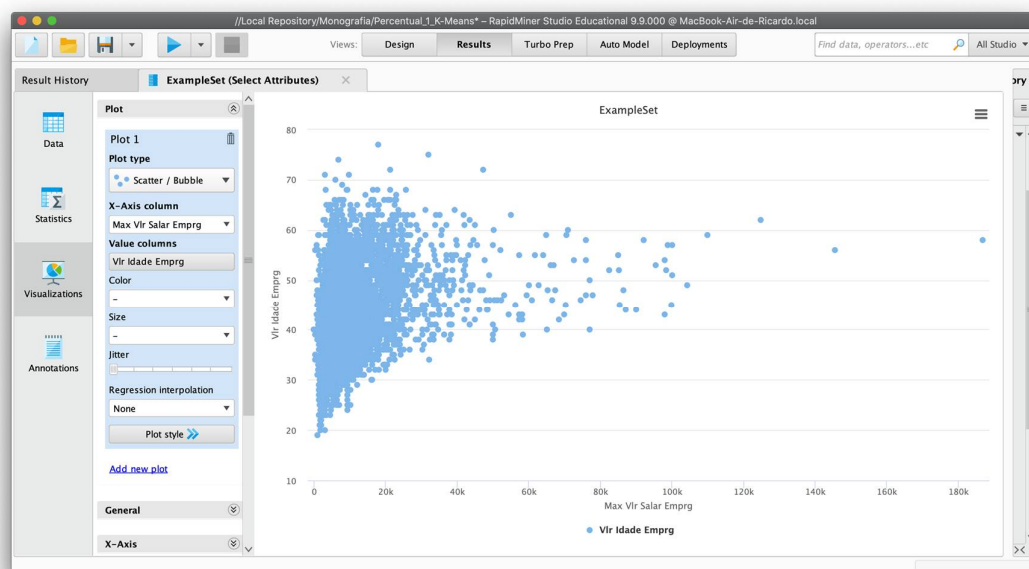


Figura 4 - RapidMiner: Base Percentual Gráfico de Dispersão

5.2 Modelagem de dados

Neste modelo foi utilizado o operador *Loop Parameters*, que juntamente com o operador de clusterização *K-Means* e o operador *Cluster Distance Performance*, para obter através de iterações de valores de “K”, o número ótimo de clusters pelo método conhecido como “Método do cotovelo” (figura 5).

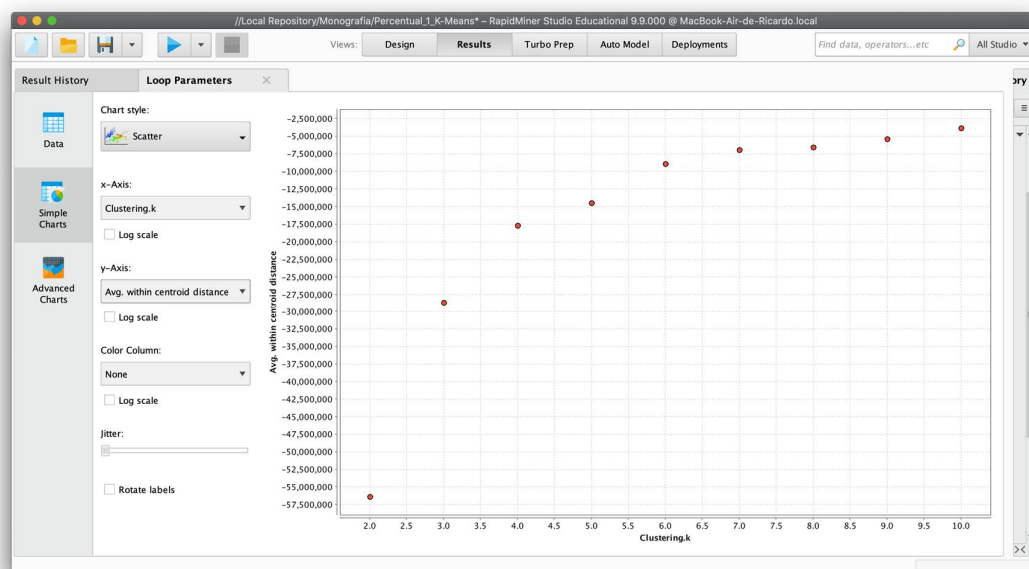


Figura 5 - RapidMiner: Base Percentual Gráfico de Dispersão “Método do Cotovelo”

Após a definição do número ótimo de clusters igual a 04, a base de dados foi apresentada aos mesmos operadores de Clusterização e a saída com o resultado foi salva em um arquivo Excel (.xlsx).

5.3 Resultados

Através da saída do modelo de clusterização, cada cluster foi analisado separadamente no Excel, onde foi realizada a análise de alguns dados estatísticos e a confirmação dos valores dos centróides.

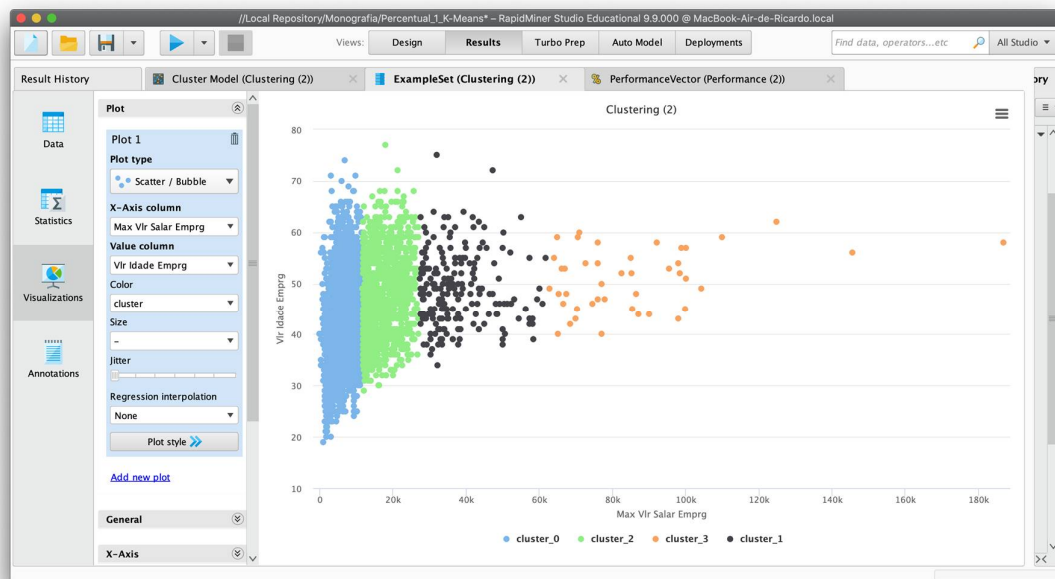


Figura 6 - RapidMiner: Base Percentual Gráfico de Dispersão Resultado Clusterização

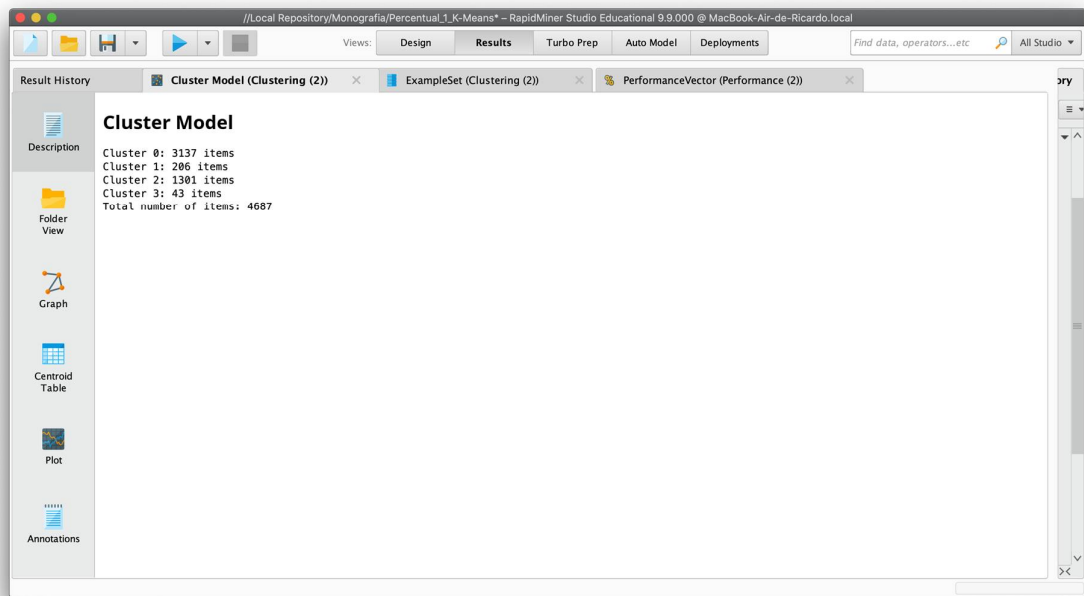


Figura 7 - RapidMiner: Base Percentual Resultado Clusterização

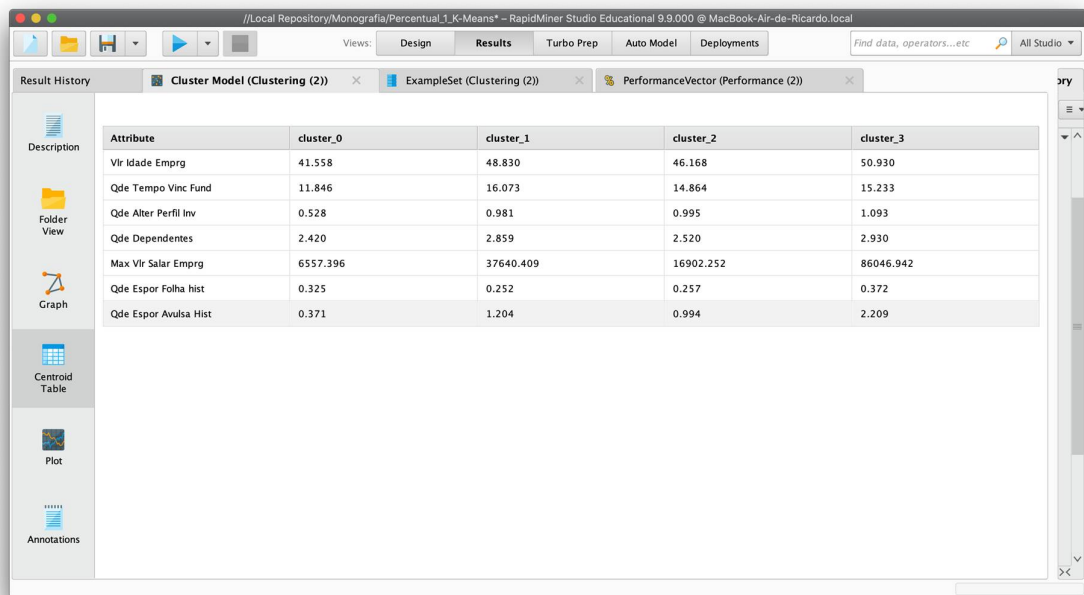


Figura 8 - RapidMiner: Base Percentual Tabela de Centroides

A maior concentração dos registros se deu no Cluster 0 com 66,9% da base analisada, este grupo apresenta valor médio do Salário nominal igual a R\$6.557,40, que corresponde a 79,5% dos participantes ativos nos planos administrados pela Valia (figura 8).

Saída K-Means Percentual										
cluster_0	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Qde Espor Folha hist	Qde Espor Avulsa Hist			
Minimo	19,00	,00	,00	,00	-	,00	,00			
Máximo	74,00	45,00	8,00	12,00	11.728,37	19,00	44,00			
Média	41,56	11,85	,53	2,42	6.557,40	,32	,37			
Mediana	40,00	12,00	,00	2,00	6.324,08	,00	,00			
Moda	37,00	13,00	,00	3,00	9.350,00	,00	,00			

Saída K-Means Percentual										
cluster_2	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Qde Espor Folha hist	Qde Espor Avulsa Hist			
Minimo	29,00	,00	,00	,00	11.752,10	,00	,00			
Máximo	77,00	40,00	10,00	10,00	27.219,70	24,00	47,00			
Média	46,17	14,86	1,00	2,52	16.902,25	,26	,99			
Mediana	45,00	13,00	1,00	3,00	15.928,53	,00	,00			
Moda	39,00	9,00	,00	3,00	13.295,95	,00	,00			

Saída K-Means Percentual										
cluster_1	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Qde Espor Folha hist	Qde Espor Avulsa Hist			
Minimo	34,00	,00	,00	,00	27.440,21	,00	,00			
Máximo	75,00	37,00	5,00	8,00	61.723,20	9,00	15,00			
Média	48,83	16,07	,98	2,86	37.640,41	,25	1,20			
Mediana	48,00	16,00	1,00	3,00	35.092,63	,00	,00			
Moda	43,00	19,00	,00	3,00	29.297,39	,00	,00			

Saída K-Means Percentual										
cluster_3	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Qde Espor Folha hist	Qde Espor Avulsa Hist			
Minimo	40,00	,00	,00	,00	62.864,64	,00	,00			
Máximo	62,00	35,00	5,00	5,00	186.750,00	8,00	13,00			
Média	50,93	15,23	1,09	2,93	86.046,94	,37	2,21			
Mediana	52,00	15,00	1,00	3,00	77.909,80	,00	,00			
Moda	58,00	15,00	,00	3,00	98.000,00	,00	,00			

Figura 9 – Excel: Base Percentual Análise do Resultado

Os valores limites encontrados na análise realizada pelo Excel (figura 9), foram aplicados na outra base de dados criada para a análise deste tipo de investimento, com os 85.918 registros de participantes que atualmente possuem percentual de contribuição normal ordinária mensal menor ou igual a 9%. Através da aplicação de filtros no Excel, com os parâmetros de saída da clusterização, foi definida uma lista de registros para cada um dos 04 clusters, que possuem maior propensão para o aumento deste tipo de investimento e farão parte de um “grupo de controle” para ações de campanha de marketing e posterior acompanhamento do percentual de conversão de sucesso.

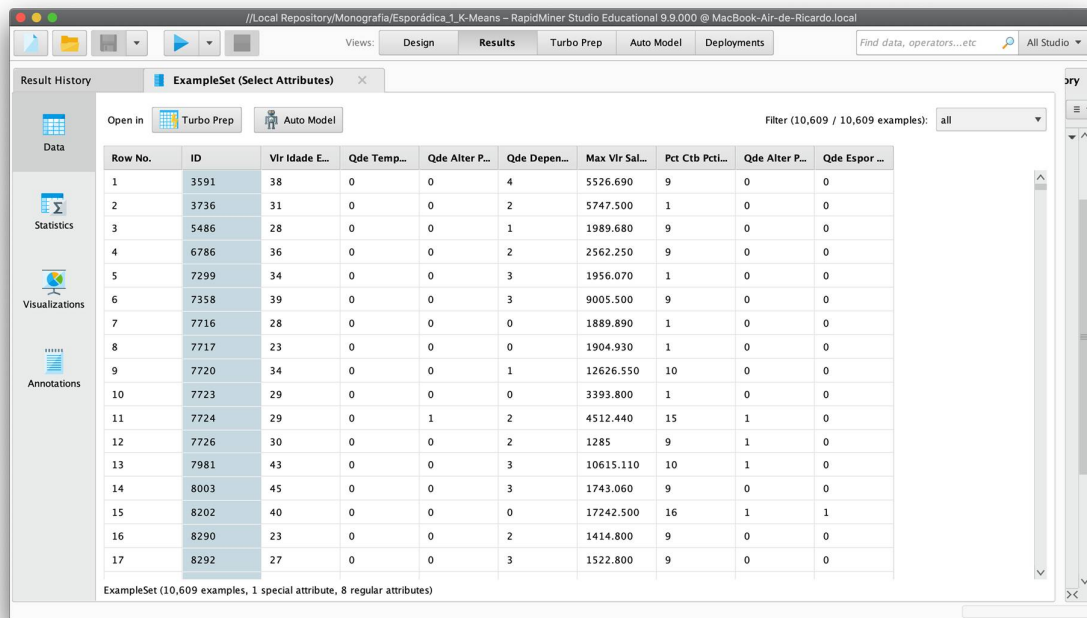
6. Clusterização 2: Realização de contribuição normal esporádica

A base de dados utilizada neste modelo é formada pelos participantes que possuem contribuição normal esporádica descontadas na folha de pagamento da Patrocinadora, isto é, os participantes com nível de contribuição acima do percentual acompanhado pela sua Patrocinadora.

6.1 Análise Exploratória e Seleção Atributos

Para a análise de realização de contribuição normal esporádica, os atributos referentes a este tipo de investimento foram retirados da base.

Desta forma, para esta análise foram considerados 10.609 registros, 01 atributo especial (ID) e 08 atributos regulares (figura 10).



Row No.	ID	Vir Idade E...	Qde Temp...	Qde Alter P...	Qde Depen...	Max Vir Sal...	Pct Ctb Pct...	Qde Alter P...	Qde Espor ...
1	3591	38	0	0	4	5526.690	9	0	0
2	3736	31	0	0	2	5747.500	1	0	0
3	5486	28	0	0	1	1989.680	9	0	0
4	6786	36	0	0	2	2562.250	9	0	0
5	7299	34	0	0	3	1956.070	1	0	0
6	7358	39	0	0	3	9005.500	9	0	0
7	7716	28	0	0	0	1889.890	1	0	0
8	7717	23	0	0	0	1904.930	1	0	0
9	7720	34	0	0	1	12626.550	10	0	0
10	7723	29	0	0	0	3393.800	1	0	0
11	7724	29	0	1	2	4512.440	15	1	0
12	7726	30	0	0	2	1285	9	1	0
13	7981	43	0	0	3	10615.110	10	1	0
14	8003	45	0	0	3	1743.060	9	0	0
15	8202	40	0	0	0	17242.500	16	1	1
16	8290	23	0	0	2	1414.800	9	0	0
17	8292	27	0	0	3	1522.800	9	0	0

Figura 10 - RapidMiner: Base Esporádica Carregada

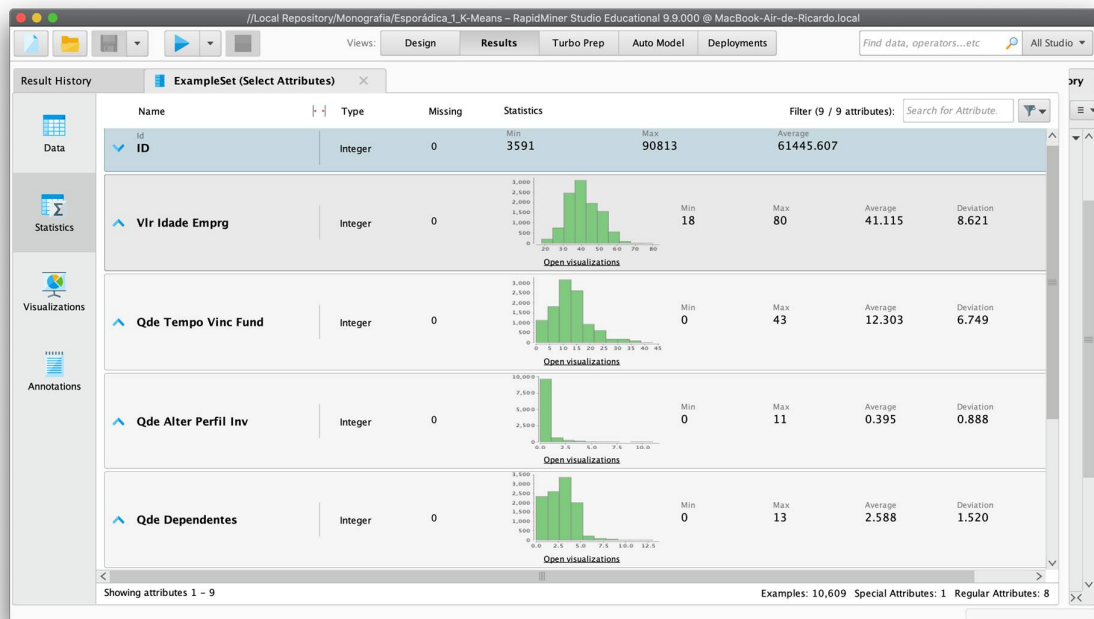


Figura 11 - RapidMiner: Base Esporádica Estatísticas

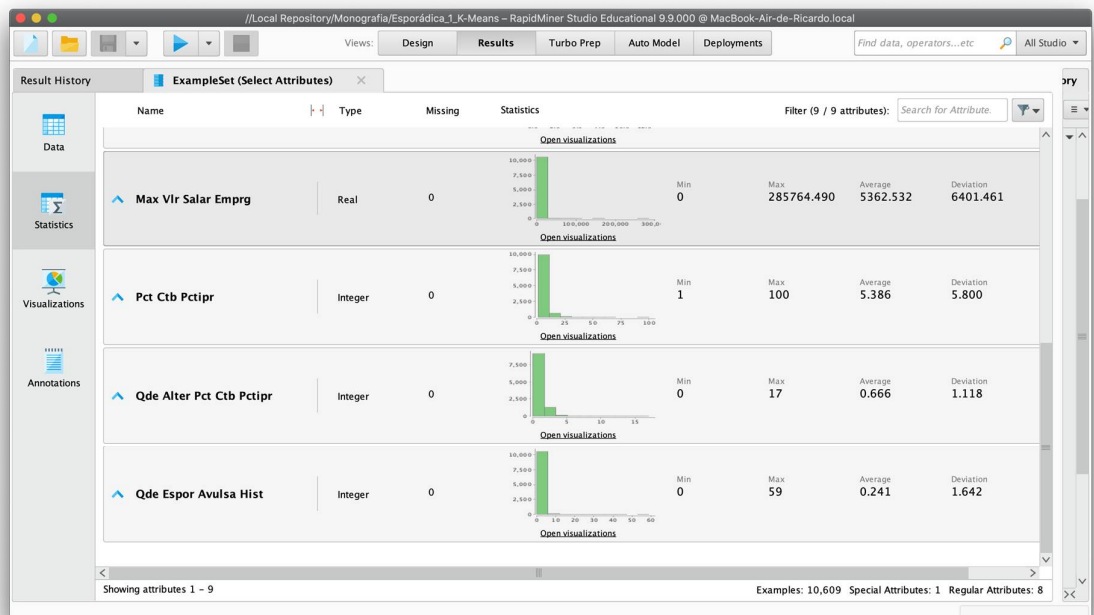


Figura 12 - RapidMiner: Base Esporádica Estatísticas

Como podemos observar na visualização do gráfico (figura 13), a maioria dos dados está concentrada na faixa salarial de até R\$25.000, onde existe se acumula o público que realiza contribuição normal esporádica descontada na folha de pagamento das patrocinadoras.

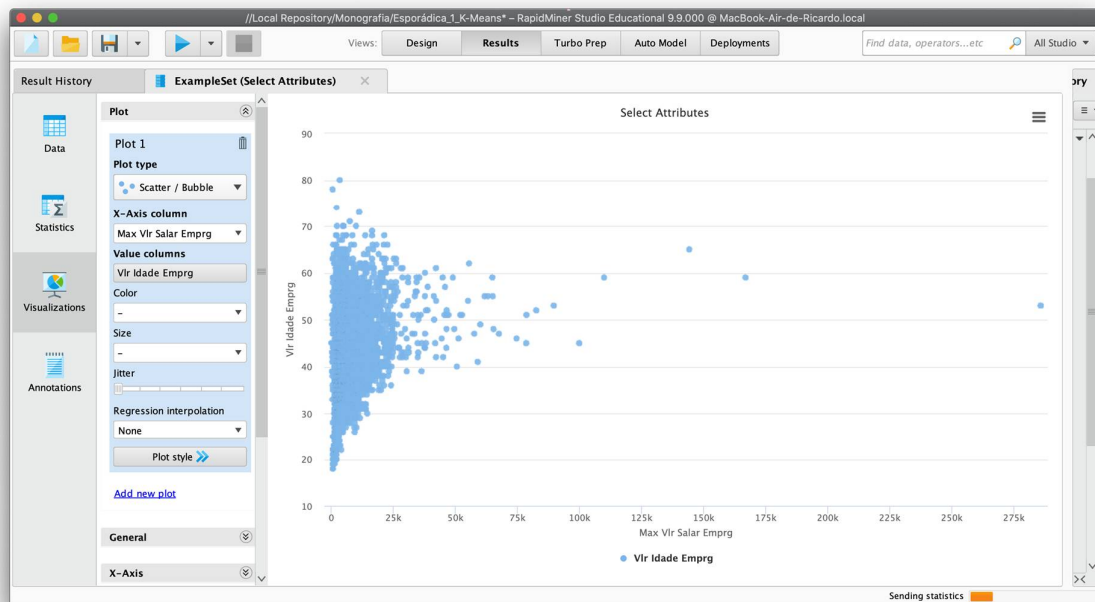


Figura 13 - RapidMiner: Base Esporádica Gráfico de Dispersão

6.2 Modelagem de dados

Neste modelo foi utilizada mesma metodologia do modelo anterior, com os operadores *Loop Parameters*, *K-Means* e *Cluster Distance Performance*, para obter através de iterações de valores de “K” (figura 14).

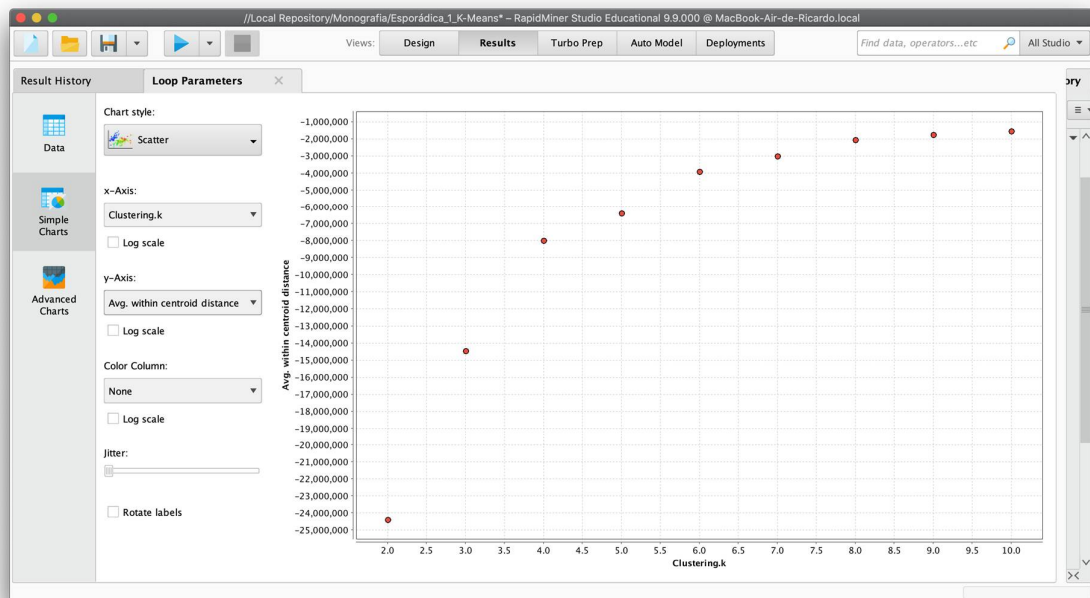


Figura 14 - Base Esporádica Gráfico de Dispersão “método do cotovelo”

Após a definição do número ótimo de clusters igual a 04, a base de dados foi apresentada aos mesmos operadores de Clusterização e a saída com o resultado foi salva em um arquivo Excel (.xlsx) para a continuação da análise.

6.3 Resultados

A maior concentração dos registros se deu no cluster 0 com 86,2% seguido pelo cluster 2 com 13,1%, ambos os clusters localizados na faixa salarial até R\$25.000, como pode ser observado no gráfico da figura 15. A maior concentração no cluster 0 é explicada pelo valor centróide do salário nominal em R\$3.770 que corresponde ao grupo de participantes com salários inferiores a 10 UR (Unidade de Referência), que por força regulamentar ficam limitados à contribuição normal ordinária mensal de 1% (um por cento), conforme identificado na descrição do problema de negócio. Desta forma a única opção de aumento de investimento dentro do mesmo plano de benefícios, é através da realização de contribuição normal esporádica. No entanto, os participantes com salários inferiores à 10 UR, atualmente não estão mais restritos a este tipo de aumento de investimento, podendo optar pela inscrição no plano Prevaler.

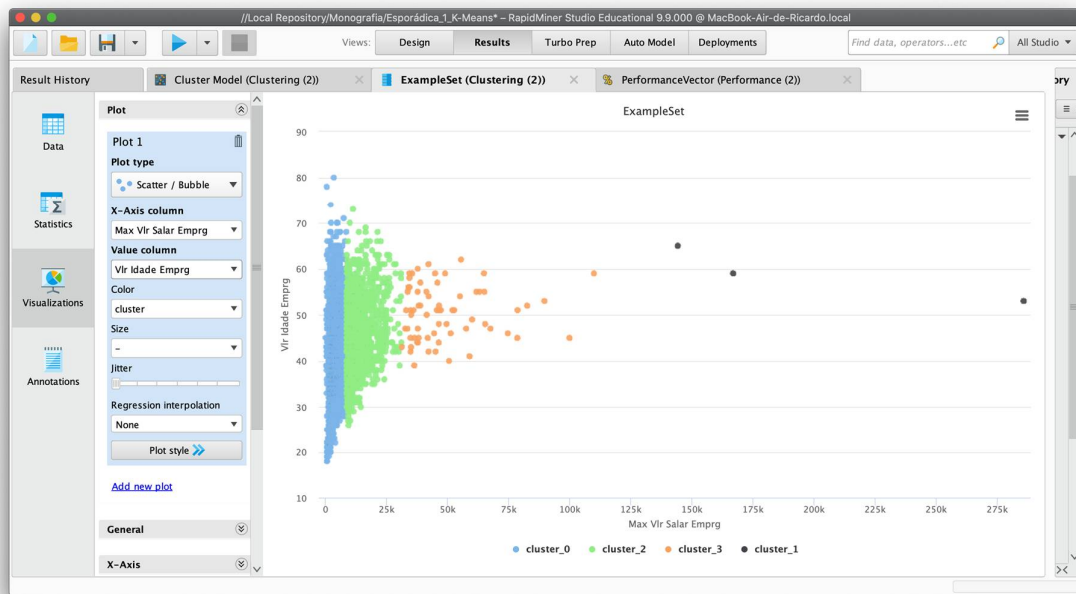


Figura 15 - Base Esporádica Gráfico de Dispersão Resultado Clusterização

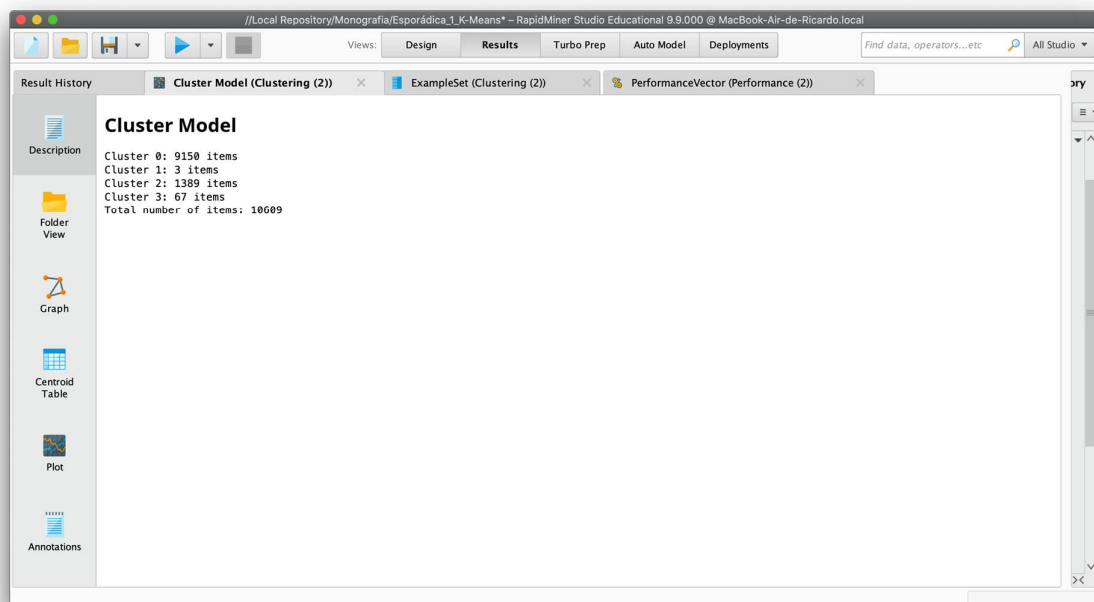


Figura 16 - RapidMiner: Base Esporádica Descrição Resultado Clusterização

Result History: Cluster Model (Clustering (2)) | ExampleSet (Clustering (2)) | PerformanceVector (Performance (2))

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Vlr Idade Emprg	40.415	59	45.234	50.433
Qde Tempo Vinc Fund	11.569	9.333	16.796	19.642
Qde Alter Perfil Inv	0.302	0	0.965	1.209
Qde Dependentes	2.568	3.667	2.701	2.970
Max Vlr Salar Emprg	3770.430	198925.690	13341.691	48705.695
Pct Ctb Pctipr	4.729	1	9.554	8.896
Qde Alter Pct Ctb Pctipr	0.521	1	1.533	2.478
Qde Espor Avulsa Hist	0.147	0	0.769	2.164

Figura 17 - RapidMiner: Base Esporádica Tabela de Centroides

Os valores limites encontrados na análise realizada pelo Excel (figura 18), foram aplicados na outra base de dados criada para a análise deste tipo de investimento, com os 80.206 registros de participantes que atualmente não possuem contribuição normal esporádica. Através da aplicação de filtros no Excel, com os parâmetros de saída do modelo 2, foi definida uma lista de registros para cada um dos 04 clusters, que possuem maior propensão para o aumento deste tipo de investimento e farão parte do “grupo de controle” para ações de campanha de marketing.

Saída K-Means Esporádica											
cluster_0	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctipr	Qde Alter Pct Ctb Pctipr	Qde Espor Avulsa Hist			
Mínimo	18,00	,00	,00	,00	-	1,00	,00	,00			
Máximo	80,00	43,00	10,00	13,00	8.553,35	100,00	17,00	59,00			
Média	40,42	11,57	,30	2,57	3.770,43	4,73	,52	,15			
Mediana	39,00	10,00	,00	3,00	3.397,10	1,00	,00	,00			
Moda	39,00	9,00	,00	3,00	2.760,48	1,00	,00	,00			

Saída K-Means Esporádica											
cluster_2	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctipr	Qde Alter Pct Ctb Pctipr	Qde Espor Avulsa Hist			
Mínimo	26,00	,00	,00	,00	8.566,25	1,00	,00	,00			
Máximo	73,00	41,00	11,00	13,00	31.022,97	40,00	11,00	56,00			
Média	45,23	16,80	,97	2,70	13.341,69	9,55	1,53	,77			
Mediana	44,00	16,00	,00	3,00	11.574,63	9,00	1,00	,00			
Moda	43,00	9,00	,00	3,00	9.350,00	9,00	1,00	,00			

Saída K-Means Esporádica											
cluster_3	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctipr	Qde Alter Pct Ctb Pctipr	Qde Espor Avulsa Hist			
Mínimo	39,00	2,00	,00	,00	31.240,81	1,00	,00	,00			
Máximo	62,00	35,00	3,00	8,00	109.880,23	25,00	11,00	15,00			
Média	50,43	19,64	1,21	2,97	48.705,70	8,90	2,48	2,16			
Mediana	51,00	18,00	1,00	3,00	44.431,74	9,00	2,00	1,00			
Moda	51,00	18,00	,00	3,00	46.427,17	9,00	2,00	,00			

Saída K-Means Esporádica											
cluster_1	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctipr	Qde Alter Pct Ctb Pctipr	Qde Espor Avulsa Hist			
Mínimo	53,00	3,00	,00	2,00	144.191,62	1,00	1,00	,00			
Máximo	65,00	17,00	,00	6,00	285.764,49	1,00	1,00	,00			
Média	59,00	9,33	,00	3,67	198.925,69	1,00	1,00	,00			
Mediana	59,00	8,00	,00	3,00	166.820,96	1,00	1,00	,00			
Moda	-	-	,00	-	-	1,00	1,00	,00			

Figura 18 - Excel: Base Esporádica Análise do Resultado

7. Clusterização 3: Inscrição e manutenção de contribuições no Plano Prevaler

A base de dados utilizada neste modelo é formada pelos participantes que possuem Inscrição no Plano Prevaler ou algum familiar/dependente. Os dados utilizados na análise são dos participantes dos planos patrocinados, uma vez que a Valia não possui dados suficientes de dependentes e familiares.

7.1 Análise Exploratória e Seleção Atributos

Para esta análise, os atributos referentes a este tipo de investimento foram retirados da base de dados. Desta forma, para esta análise foram considerados 1.293 registros, 01 atributo especial (ID) e 09 atributos regulares (figura 19).

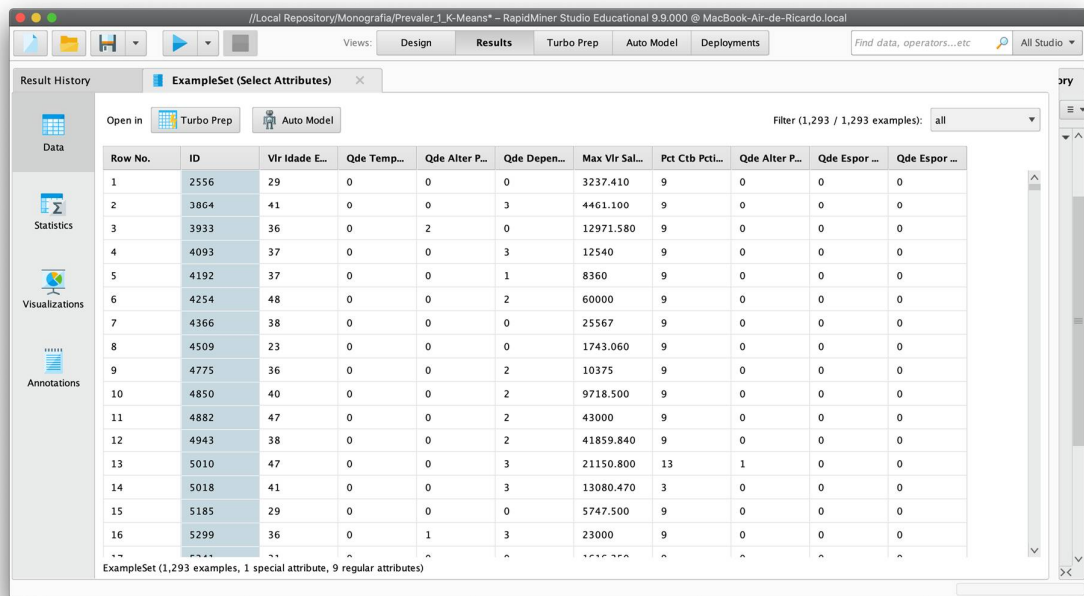


Figura 19 - RapidMiner: Base Prevaler Carregada

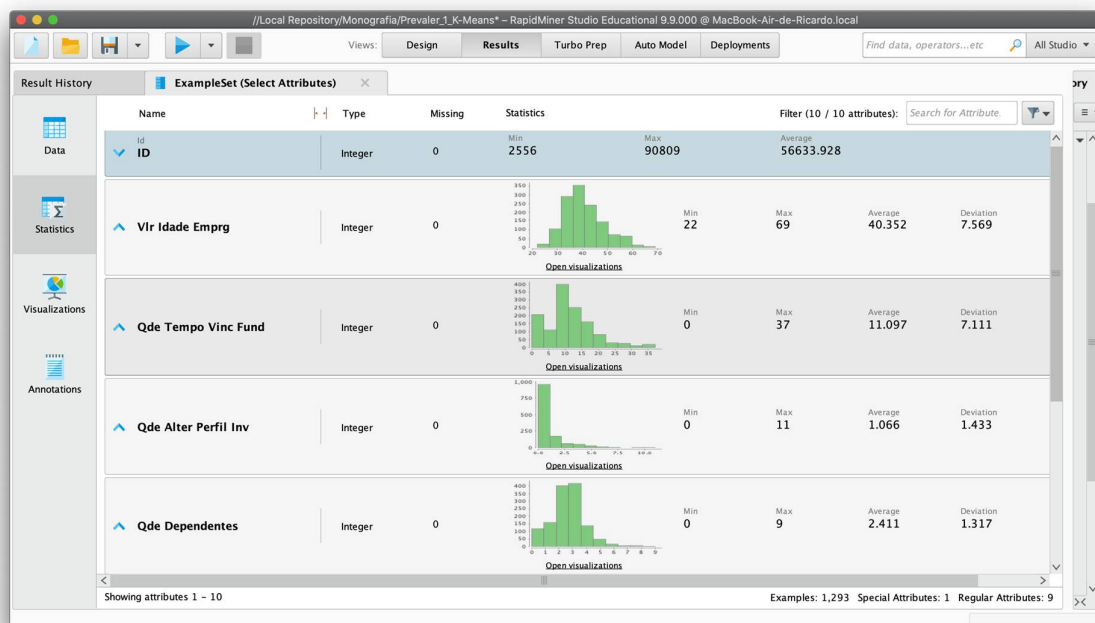


Figura 20 - RapidMiner: Base Prevaler Estatísticas

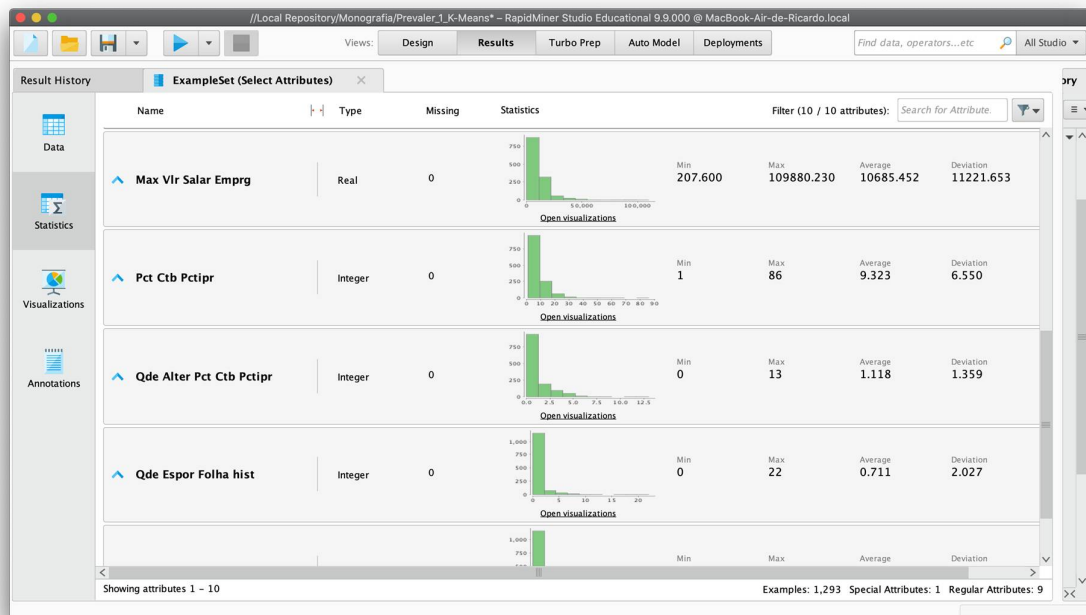


Figura 21 - RapidMiner: Base Prevaler Estatísticas

Como podemos observar na visualização do gráfico da figura 22, os dados estão mais distribuídos pelas faixas salariais, tendo em vista que o investimento no plano Prevaler não está intimamente ligado ao salário do participante.

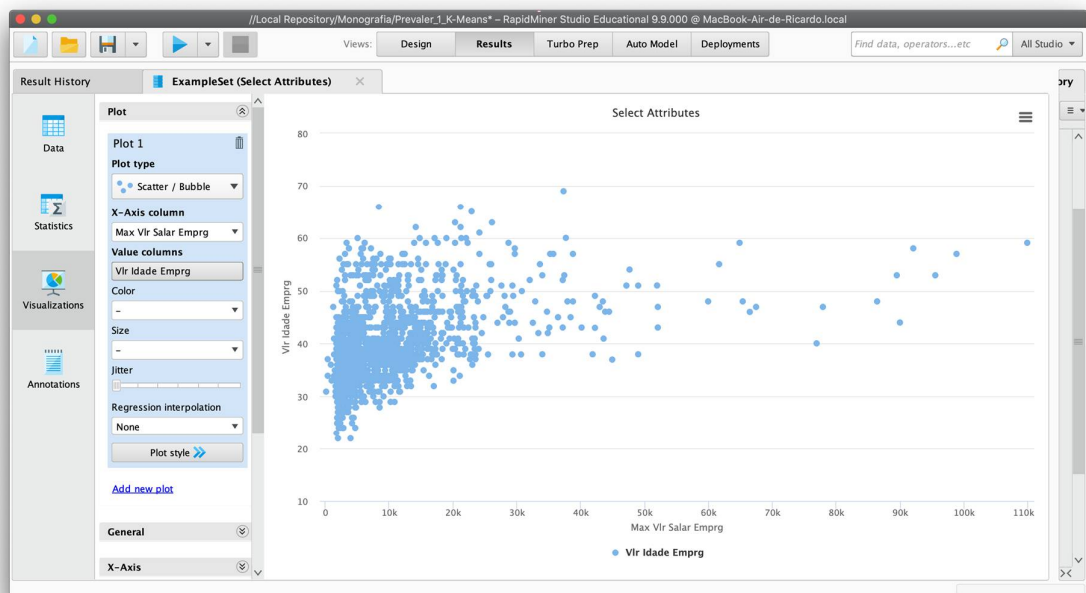


Figura 22 - RapidMiner: Base Prevaler Gráfico de Dispersão

7.2 Modelagem de dados

Neste modelo foi utilizada mesma metodologia do modelo anterior, com os operadores *Loop Parameters*, *K-Means* e *Cluster Distance Performance*, para obter através de iterações de valores de “K”.

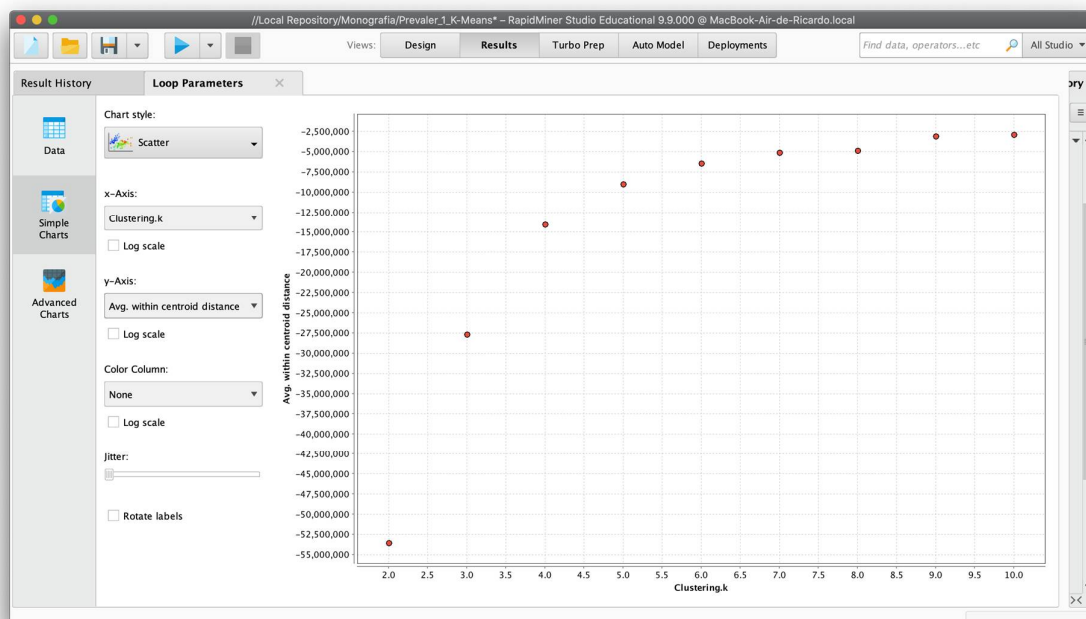


Figura 23 - RapidMiner: Base Prevaler Gráfico de Dispersão “método do cotovelo”

Com a maior distribuição dos dados pelas faixas salariais, tivemos o acréscimo de 01 cluster, totalizando em 05 clusters (figura 23), diferentemente dos modelos anteriores. A base de dados foi apresentada aos mesmos operadores de Clusterização e a saída com o resultado foi salva em um arquivo Excel (.xlsx) para a continuação da análise.

7.3 Resultados

A maior concentração dos registros se deu no cluster 4 com 52% seguido pelo cluster 0 com 34%, ambos os clusters localizados na faixa salarial até R\$20.000, como pode ser observado no gráfico da figura 24. A maior concentração no cluster 4 pode ser explicada pelo valor centróide do salário nominal em R\$4.332, que além de concentrar o maior número de participantes dos planos patrocinados, corresponde ao grupo de participantes com salários inferiores a 10 UR (Unidade de Referência), que encontram no plano Prevaler mais uma

oportunidade de investimento, com características mais flexíveis que as do plano patrocinado, conforme identificado na descrição do problema de negócio.

Já os participantes identificados no cluster 0, possuem valor centróide para o salário nominal em R\$12.032 e conforme também identificado nos demais clusters desta análise, possuem alta interação com a Valia em relação às demais alternativas de aumento de investimento, como o percentual de contribuição acima de 9% e a realização de contribuições esporádicas (figura 26).

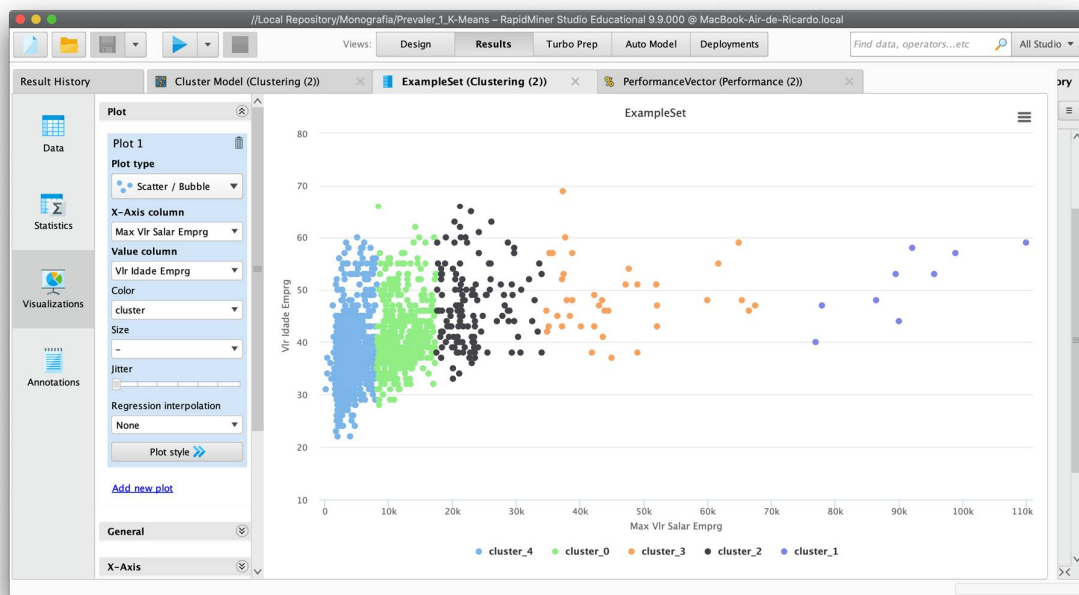


Figura 24 - RapidMiner: Base Prevaler Gráfico de Dispersão Resultado Clusterização

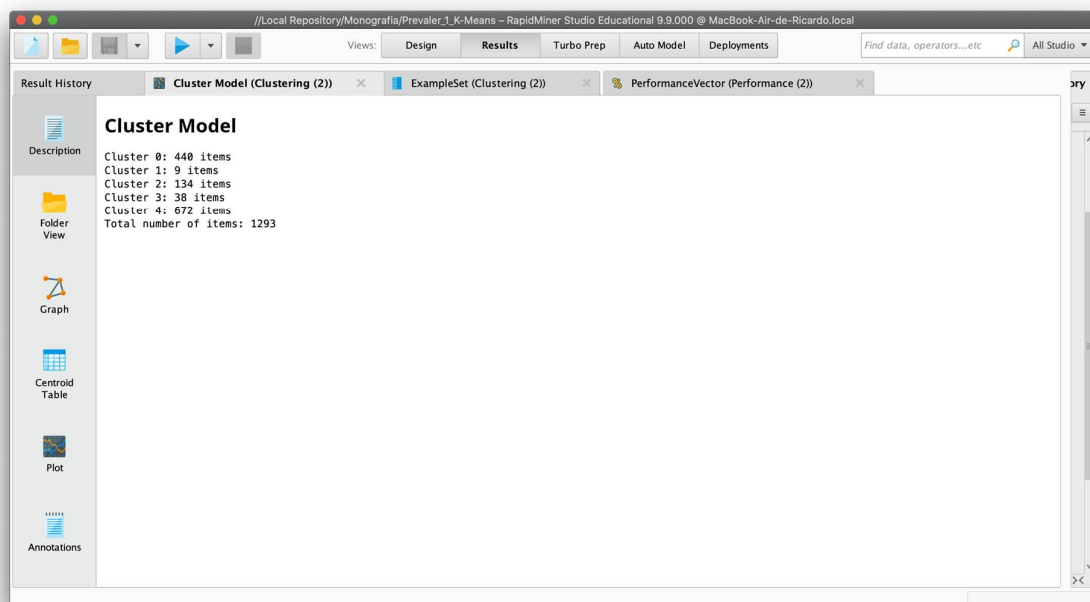


Figura 25 - RapidMiner: Base Prevaler Descrição Resultado Clusterização

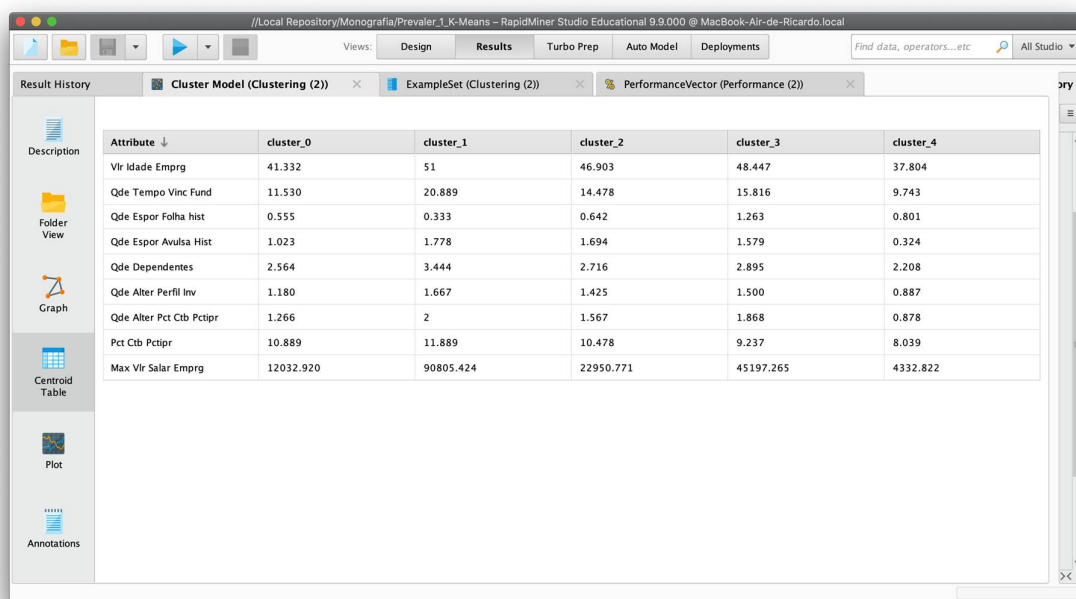


Figura 26 - RapidMiner: Base Prevaler Tabela de Centroides

Os valores limites encontrados na análise realizada pelo Excel, foram aplicados na outra base de dados com os 89.522 registros de participantes que atualmente não possuem inscrição no plano Prevaler. Através da aplicação de filtros no Excel, com os parâmetros de saída do modelo

3 (figura 27), foi definida uma lista de registros para cada um dos 05 clusters, que possuem maior propensão para efetivação de inscrição no plano e também farão parte do “grupo de controle” para ações de campanha de marketing.

Saída K-Means Prevaler												
cluster_4	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctpr	Qde Alter Pct Ctb Pctpr	Qde Espor Folha hist	Qde Espor Avulsas Hist			
Mínimo	22,00	,00	,00	,00	207,60	1,00	,00	,00	,00			
Máximo	60,00	34,00	10,00	8,00	8.161,68	86,00	13,00	22,00	9,00			
Média	37,80	9,74	,89	2,21	4.332,82	8,04	,88	,80	,32			
Mediana	37,00	9,00	,00	2,00	4.139,02	9,00	1,00	,00	,00			
Moda	37,00	9,00	,00	2,00	1.950,93	9,00	,00	,00	,00			
Saída K-Means Prevaler												
cluster_0	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctpr	Qde Alter Pct Ctb Pctpr	Qde Espor Folha hist	Qde Espor Avulsas Hist			
Mínimo	28,00	,00	,00	,00	8.201,30	1,00	,00	,00	,00			
Máximo	66,00	37,00	11,00	8,00	17.462,38	37,00	11,00	19,00	18,00			
Média	41,33	11,53	1,18	2,56	12.032,92	10,89	1,27	,55	1,02			
Mediana	40,00	10,00	1,00	3,00	11.584,72	9,00	1,00	,00	,00			
Moda	39,00	9,00	,00	3,00	9.350,00	9,00	,00	,00	,00			
Saída K-Means Prevaler												
cluster_2	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctpr	Qde Alter Pct Ctb Pctpr	Qde Espor Folha hist	Qde Espor Avulsas Hist			
Mínimo	33,00	,00	,00	,00	17.545,55	1,00	,00	,00	,00			
Máximo	66,00	37,00	7,00	9,00	34.052,81	30,00	8,00	18,00	29,00			
Média	46,90	14,48	1,43	2,72	22.950,77	10,48	1,57	,64	1,69			
Mediana	46,00	14,00	1,00	3,00	21.670,40	9,00	1,00	,00	,00			
Moda	38,00	9,00	,00	3,00	20.900,00	9,00	1,00	,00	,00			
Saída K-Means Prevaler												
cluster_3	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctpr	Qde Alter Pct Ctb Pctpr	Qde Espor Folha hist	Qde Espor Avulsas Hist			
Mínimo	37,00	,00	,00	,00	34.738,79	1,00	,00	,00	,00			
Máximo	69,00	36,00	5,00	6,00	67.485,60	20,00	5,00	11,00	12,00			
Média	48,45	15,82	1,50	2,89	45.197,26	9,24	1,87	1,26	1,58			
Mediana	47,50	16,00	1,00	3,00	42.638,48	9,00	2,00	,00	,00			
Moda	48,00	16,00	,00	2,00	-	9,00	,00	,00	,00			
Saída K-Means Prevaler												
cluster_1	Vir Idade Emprg	Qde Tempo Vinc Fund	Qde Alter Perfil Inv	Qde Dependentes	Max Vir Salar Emprg	Pct Ctb Pctpr	Qde Alter Pct Ctb Pctpr	Qde Espor Folha hist	Qde Espor Avulsas Hist			
Mínimo	40,00	9,00	,00	2,00	77.000,00	1,00	,00	,00	,00			
Máximo	59,00	35,00	5,00	5,00	109.880,23	17,00	6,00	2,00	7,00			
Média	51,00	20,89	1,67	3,44	90.805,42	11,89	2,00	,33	1,78			
Mediana	53,00	20,00	1,00	3,00	90.000,00	12,00	1,00	,00	1,00			
Moda	53,00	-	,00	3,00	-	12,00	1,00	,00	,00			

Figura 27 - Excel: Base Prevaler Análise do Resultado

8. Glossário

Assistido - participante ou seu beneficiário em gozo de benefício de prestação continuada.

Autopatrocinio - é o instituto que faculta ao participante manter o valor de sua contribuição e a do patrocinador, no caso de perda parcial ou total da remuneração recebida, para assegurar a percepção dos benefícios nos níveis correspondentes àquela remuneração ou em outros definidos em normas regulamentares.

Benefício Proporcional Diferido - é o instituto que, em razão da cessação do vínculo empregatício com o patrocinador ou associativo com o instituidor antes da aquisição do direito

ao benefício pleno programado, propicia a percepção de benefício a ser concedido quando cumpridos os requisitos de elegibilidade.

Conta de Participante - é a parcela do Saldo de Conta em que serão creditadas as contribuições individuais dos participantes mencionadas nos artigos 121 e 123 deste Regulamento, incluindo a respectiva rentabilidade líquida.

Conta de Patrocinador - é a parcela do Saldo de Conta em que serão creditadas as contribuições individuais do patrocinador mencionadas nos artigos 130 e 131 deste Regulamento, incluindo a respectiva rentabilidade líquida.

Herdeiro legal - para fins deste regulamento, considera-se herdeiro legal aquele como tal considerado pelo Código Civil, bem como aquele designado ou indicado em testamento.

Opção de investimento - é a estrutura de aplicação financeira do Saldo de Conta do participante, de acordo com o exercício ou não de sua opção, conforme normatização aprovada pelo Conselho Deliberativo da Valia.

Participante - a pessoa física que aderiu a este plano.

Portabilidade - é o instituto que faculta a transferência do direito acumulado pelo participante para outro plano.

Previdência Social - sistema governamental que tem como objetivo reconhecer e conceder benefícios previdenciários aos seus segurados e seus dependentes ou outro sistema de caráter oficial com objetivos similares.

Recursos Portados - constituídos pelos valores portados de outro plano de benefícios de entidade de previdência complementar ou de sociedade seguradora, segregados em subcontas por entidade aberta ou fechada de previdência complementar, conforme a origem.

Rentabilidade Líquida - corresponde à variação líquida proveniente da aplicação financeira das contribuições mencionadas neste artigo e dos rendimentos decorrentes da aplicação do patrimônio por elas já constituído, de acordo com a opção de investimento aplicável.

Resgate - é o instituto que faculta o resgate de contribuições vertidas ao plano, descontadas as parcelas do custeio administrativo, na forma regulamentada.

Saldo de Conta - é a soma dos valores das contas de participante e de patrocinador.

Saldo de Conta remanescente - é o valor apurado do Saldo de Conta do assistido, deduzidos os benefícios de renda pagos e acrescida a rentabilidade líquida.

Taxa de Administração - taxa incidente sobre o Saldo de Conta, Saldo de Conta remanescente e Recursos Portados previstos neste regulamento, para custeio das despesas administrativas do plano.

Taxa de Carregamento - taxa incidente sobre os valores das contribuições para custeio das despesas administrativas do Plano.

Transação remota - qualquer operação à distância envolvendo o uso de plataforma digital disponibilizada pela Valia, que requeira manifestação expressa dos seus participantes e assistidos perante a Fundação, incluindo, mas não se limitando, às seguintes situações: Adesão; Modificação das opções disponibilizadas ao participante ou assistido; Cancelamento.

UR (Unidade de Referência) - corresponde a 1/10 (um décimo) de R\$ 4.377,10 (quatro mil trezentos e setenta e sete reais e dez centavos), considerando o mês de novembro de 2020 como data base para futura incidência de reajuste sobre esse valor. A UR será reajustada na data base do acordo coletivo da Instituidora, considerando a variação do IPC-FGV acumulada no período compreendido entre o mês anterior ao do seu último reajuste e o penúltimo mês imediatamente anterior ao da data base do acordo coletivo da Instituidora. Justificadamente, mediante autorização do Conselho Deliberativo da VALIA e após a autorização do órgão governamental competente, poderá ser adotado outro critério de correção para a UR. Esse critério considerará a aplicação de índice econômico a ser implementado a partir do mês de janeiro do exercício seguinte ao de sua aprovação, precedido de ampla divulgação aos participantes e patrocinadores.

Caso a variação do IPC-FGV seja negativa, o valor da UR não será alterado.

Participante – Pessoa física que, nas condições deste Regulamento, seja admitida neste Plano.

Participante Autopatrocinado – Aquele que, estando na condição de Participante, optar pelo instituto do Autopatrocínio.

Participante Vinculado – Aquele que, estando na condição de Participante, optar pelo instituto do Benefício Proporcional Diferido.

Plano ou Plano de Benefícios – Conjunto de direitos e obrigações reunidos neste Regulamento.

9. Referencias Bibliográficas

BEZERRA, E.; GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações**. 2ª ed. Rio de Janeiro: Elsevier, 2015.

BRUCE, P; BRUCE, A. **Estatística Prática para Cientistas de Dados**. 50 Conceitos Essenciais. 1ª ed. Rio de Janeiro: Alta Books Editora, 2019

FAWCET, T; PROVOST, F. **Data Science para Negócios: O que Você Precisa Saber Sobre Mineração de Dados e Pensamento Analítico de Dados**. 1ª ed. Rio de Janeiro: Alta Books Editora, 2016.