NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Seguro de Saúde Individual

Variáveis Diferenciadoras do Risco na cobertura de Internamento

Marta Pereira Barceló

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

SEGURO DE SAÚDE INDIVIDUAL

VARIÁVEIS DIFERENCIADORAS DO RISCO NA COBERTURA DE INTERNAMENTO

por

Marta Pereira Barceló

Dissertação como requisito parcial para a obtenção do grau de Mestre em Estatística e Gestão de Informação, Especialização em Análise e Gestão de Risco

Orientador: Professora Doutora Gracinda Rita Diogo Guerreiro

Coorientador: Mestre Pedro Brigas Marcelino

novembro 2022

AGRADECIMENTOS

A realização desta dissertação não seria possível sem a ajuda e o apoio de um conjunto de pessoas a quem agradeço, em particular:

À Professora Gracinda Guerreiro, orientadora desta dissertação, pela oportunidade de orientar este trabalho e pela disponibilidade demonstrada, em específico por me ter facultado o módulo de tarifação, essencial para a realização deste trabalho.

Ao Pedro Marcelino, também orientador desta dissertação, pelo acompanhamento na minha integração profissional e constante apoio em todas as etapas desta dissertação.

À Multicare, em especial ao Gabinete de Atuariado e Controlo Técnico, por me terem integrado e fornecido os dados que sem os quais não teria sido possível realizar este trabalho. Em especial, à equipa de Suporte Atuarial Standard pela sua colaboração e acompanhamento no dia a dia.

Ao meu namorado, João Martins, por fazermos este caminho juntos, e estar sempre disposto a ouvir-me, suportando-me em todos os momentos, tanto nos bons como nos mais exigentes.

À minha família, avós e irmã, e em especial aos meus pais pelo que me proporcionaram e pela confiança demonstrada em todas as etapas da minha vida.

A todos que direta ou indiretamente contribuíram para alcançar este objetivo, o meu muito obrigada.

RESUMO

O financiamento dos cuidados de saúde em Portugal é assegurado pelo Serviço Nacional de Saúde (SNS); no entanto, para um acesso mais rápido, com mais opção de escolha e comodidade, os portugueses podem recorrer a um seguro de saúde. Para isso, terão que pagar um prémio que é definido de forma idêntica para pessoas com fatores de risco semelhantes. O objetivo deste trabalho é estudar que variáveis, dentro de um conjunto de características pessoais, do seguro escolhido, socioeconómicas e de saúde, têm impacto no risco, ou seja, na frequência de sinistralidade e no custo associado.

O objetivo desta dissertação consiste na identificação de variáveis diferenciadoras do risco da cobertura de Internamento, para os seguros de saúde individuais. O modelo de risco que identifica as variáveis significativas assenta nos Modelos Lineares Generalizados e será analisado o comportamento do risco das variáveis significativas através de duas técnicas de *Machine Learning*: Análise de *Clusters* e Árvores de Decisão. Os dados da carteira foram fornecidos por uma Seguradora a operar em Portugal.

PALAVRAS-CHAVE

Variáveis Diferenciadoras do Risco do Internamento; Seguro de Saúde; Modelos Lineares Generalizados; Análise de *Clusters*; Árvores de Decisão

ABSTRACT

The financing of health care in Portugal is provided by the National Health Service (SNS); however, for faster access, with more choice and convenience, the Portuguese can take out health insurance. For this, they will have to pay a premium that is defined identically for people with similar risk factors. The objective of this work is to study which variables, within a set of personal characteristics, of the chosen insurance, socio-economic and health, have an impact on risk, that is, on the frequency of accidents and the cost associated with this loss.

The objective of this dissertation is to identify variables differentiating the risk of hospitalization coverage for individual health insurance. The risk model that identifies the significant variables is based on the Generalized Linear Models and the risk behavior of the significant variables will be analyzed through two machine learning techniques: Cluster Analysis and Decision Trees. The portfolio data were provided by an Insurance Company operating in Portugal.

KEYWORDS

Risk Differentiating Variables of Hospitalization; Health insurance; Generalized Linear Models; Cluster Analysis; Decision Trees

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AIC *Akaike Information Criterion*

APS Associação Portuguesa de Seguradores

ASF Autoridade de Supervisão de Seguros e Fundos de Pensões

CART *Classification and Regression Tree*

INE Instituto Nacional de Estatística

MAG Modelos Aditivos Generalizados

ML *Machine Learning*

MLG Modelos Lineares Generalizados

SNS Serviço Nacional de Saúde

1. INTRODUÇÃO

1.1. ENQUADRAMENTO E IDENTIFICAÇÃO DO PROBLEMA

O papel de uma Seguradora passa por proteger os riscos seguráveis (podem ser de danos ou de pessoas), estabilizando e salvaguardando a situação financeira das famílias e das empresas, com a contrapartida de elas efetuarem um pagamento fixo antecipado, denominado de prémio. O segurado transfere o risco para a Seguradora, através das condições estabelecidas na apólice de seguro, e a Seguradora recebe um prémio que deverá ser suficiente para cobrir as indemnizações do segurado na eventualidade da ocorrência de danos imprevisíveis: os sinistros. O prémio que o tomador do seguro tem a seu cargo deverá ser calculado de forma a garantir que a Seguradora consegue cumprir com responsabilidades futuras, relativamente a montantes desconhecidos *a priori,* sem comprometer a situação de solvência da empresa. Esta tarefa, desempenhada por atuários, consiste em estimar um valor de prémio justo e adequado para cada cliente. O prémio que o segurado paga é composto pelo prémio de risco ou prémio puro (valor esperado do montante de sinistros que a Seguradora terá de suportar) mais outras parcelas como despesas administrativas, despesas com gastos externos e encargos; no entanto, para este estudo apenas se vai considerar o prémio de risco.

Sendo a Seguradora uma empresa em que o produto vendido é uma proteção financeira face a um risco – incerteza associada a um acontecimento futuro, seja quanto à sua realização, ao momento em que ocorre e aos danos dele decorrentes – é fundamental que a sua quantificação tenha por base informação suficientemente sólida. O risco é um evento aleatório, há sempre uma margem de erro associada, mas se se tiver mais conhecimento sobre os fatores que podem influenciar a ocorrência desses eventos, poderá fazer-se uma melhor estimação do risco *a priori*. Nos seguros onde estão abrangidas muitas pessoas, que é o caso do seguro de saúde, é usual encontrar-se um certo grau de heterogeneidade nos riscos da carteira e, assim, deve-se definir uma tarifa baseada em características representativas do risco que distinguem os prémios por categorias. Assim, com base em características homogéneas de risco de vários clientes, constrói-se uma tarifa equilibrada tanto para a Seguradora (o valor é suficiente face ao custo estimado de sinistros) como para os indivíduos (o preço é acessível face ao seu poder de compra). A tarifa é construída com base em fatores tarifários que são características do segurado relevantes para o risco e esses fatores têm níveis tarifários, ou seja, para cada fator a pessoa toma um dos seus níveis (por exemplo, se a idade é um fator tarifário, um escalão etário é um nível tarifário).

Os seguros de saúde são seguros de pessoas e fazem parte do Ramo Não-Vida, representando 18% deste mercado (APS, 2021); normalmente, são designados como Ramo Doença e incluem, entre outras, coberturas como Internamento, Ambulatório, Estomatologia, Próteses e Ortóteses. A adesão ao seguro de saúde é de carácter voluntário e a apólice pode ser individual, no caso se ser um agregado familiar ou de grupo, se se tiver um conjunto de agregados. Em saúde, sabe-se que, à medida que a idade aumenta, o risco tende a aumentar, mas existem outros fatores que também podem influenciar o risco. No caso do género, desde o final de 2012, deixou de poder ser considerado em Portugal, decorrente de uma Diretiva Europeia[1], por ser considerado uma forma de discriminação no domínio

[1]Orientações sobre a aplicação ao setor dos seguros da Diretiva 2004/113/CE do Conselho, à luz do acórdão do Tribunal de Justiça da União Europeia no Processo C-236/09 (Test-Achats)

de bens e serviços, indo contra o princípio de igualdade entre homens e mulheres consagrado na Constituição de República Portuguesa.

Em Portugal, o Sistema Nacional de Saúde é composto pelo Serviço Nacional de Saúde (SNS), pelos vários subsistemas de saúde públicos e privados, pelo setor segurador e pelo setor privado "puro", financiado por pagamentos diretos dos indivíduos (Silva, 2009). O SNS foi criado em 1979, estabelecendo o direito à proteção da saúde de todos os cidadãos, independentemente da sua condição económica e social. Apesar de o Estado assegurar o financiamento dos cuidados de saúde, os portugueses podem ainda recorrer a um seguro de saúde. Essa escolha pode dever-se ao elevado grau de indeterminação da quantidade e da duração dos cuidados de saúde que a pessoa necessita ao longo da sua vida (Abrantes, 1985). Contudo, existem outros fatores, relacionados com as características da oferta e da procura, que influenciam a decisão de compra. Filipa Baptista (2019) concluiu que o preço, a percentagem de comparticipação/reembolso, o sexo e a idade têm uma grande influência na aquisição um seguro de saúde em Portugal. Além disso, Alberto Guerra (2014) mostra que a rapidez de acesso aos cuidados de saúde e a diferenciação da oferta por parte dos serviços que o seguro apresenta são determinantes na aquisição do seguro. A procura pelo seguro de saúde tem tido uma tendência crescente, como pode ser observado na Figura 1.1.

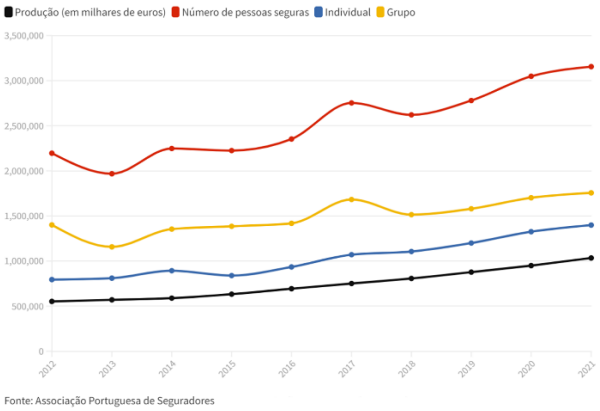


Figura 1.1 - Evolução do número de pessoas com seguro de saúde

Em 2021, havia um total de 3.155.252 pessoas seguras, representando 30,5% da população portuguesa. Por este incremento e pelas razões referidas, torna-se fundamental a Seguradora conhecer bem o risco do produto que vende, não só em termos da solvência da Seguradora, mas também para conseguir adequar melhor a sua oferta ao cliente, tornando-se mais competitiva. As pessoas não são todas iguais e, por isso, podem ter comportamentos diferentes perante o risco. Identificar as variáveis que levam a comportamento diferente do risco é essencial para a Seguradora, na medida em que irá conhecer melhor a sua carteira e fazer uma gestão da mesma, com um menor grau de incerteza.

1.2. Importância e Relevância do Estudo

A atividade Seguradora representa um papel fundamental na sociedade, a nível social garante a segurança de pessoas e bens, e a nível económico contribui para o seu desenvolvimento através do financiamento e da promoção de instrumentos que estimulam a poupança e/ou o investimento. Segundo a Associação Portuguesa de Seguradores (APS), no ano de 2021, a atividade Seguradora teve uma penetração de 6,2% na economia portuguesa, valor este que é obtido através do rácio entre o

volume de prémios da atividade e o Produto Interno Bruto (PIB). Adicionalmente nesse ano, o setor segurador estava no topo dos investidores institucionais, contribuindo com cerca de 24% do PIB. As instituições financeiras, nas quais está incluído o setor segurador, contribuem decisivamente para a expansão da economia de um país; contudo, é necessária uma constante análise e gestão de todos os riscos envolvidos. Sendo um seguro uma transferência do risco do tomador de seguro para a empresa Seguradora, esta tem de garantir que consegue corresponder a essas responsabilidades.

O estudo de características do segurado, que podem ser indicadoras do risco transferido, poderá contribuir para uma melhor perceção do risco e, consequentemente, uma melhor estimação do prémio. Assim, não só as Seguradoras terão mais confiança de que os valores que estão a cobrar são suficientes para cumprir com as suas obrigações, mas também os clientes poderão ter acesso a preços mais adequados ao seu perfil de risco.

Em estudos anteriores, foram apresentadas as razões que levam um indivíduo a comprar um seguro de saúde, inclusive qual o perfil do cliente que o compra. Contudo, podem ainda ser estudadas as características do indivíduo que têm impacto no risco que a Seguradora está a assumir. Nesta área, em (Bandeira, 2013) estudou-se a influência de algumas variáveis no custo e na frequência da cobertura de Ambulatório e concluiu-se que informações, como o número de coberturas do produto, o canal de vendas e a zona geográfica, permitem conhecer melhor o risco *a priori*.

Com este trabalho, pretende-se estudar as variáveis diferenciadoras da cobertura de Internamento e analisar o comportamento do risco, no caso dos seguros de saúde individuais. A realização deste estudo permitirá obter mais conhecimento do risco que, por si só, é um fenómeno aleatório, difícil de explicar.

1.3. OBJETIVOS DE ESTUDO

O objetivo desta dissertação passa por estudar se algumas variáveis são diferenciadoras do risco da cobertura de Internamento, nos seguros de saúde individuais. As variáveis utilizadas provêm de duas bases distintas: as variáveis internas, que se referem a características pessoais e do seguro adquirido; e as variáveis externas, de georreferenciação, fornecem informações socioeconómicas e de saúde. Será construído um modelo de risco, através dos Modelos Lineares Generalizados, que permite identificar as variáveis significativas. E, para analisar o comportamento do risco de cada uma dessas variáveis, serão utilizados métodos de *Machine Learning* como Análise de *Clusters* e Árvores de Decisão, com o objetivo de identificar a heterogeneidade dos riscos.

O trabalho está organizado em mais três capítulos, sendo que no capítulo 2 é feito um enquadramento sobre o Sistema Nacional de Saúde Português, o seguro de saúde, os fatores de risco em saúde, alguns métodos de quantificação do risco e a diferenciação do risco. No capítulo 3, é apresentada a metodologia utilizada, nomeadamente os Modelos Lineares Generalizados e alguns casos particulares destes, uma distribuição para tratar eventos extremos e métodos de agrupamento e de classificação, como Análise de *Clusters* e Árvores de Decisão. Uma vez descrito o problema e a forma de o resolver, são aplicados os métodos descritos a uma carteira de seguros de saúde individuais para a cobertura Internamento, no capítulo 4, e são apresentados os resultados obtidos.

2. REVISÃO DA LITERATURA

2.1. O SISTEMA DE SAÚDE EM PORTUGAL E OS SEGUROS DE SAÚDE

O Sistema Nacional de Saúde, em Portugal, é composto pelo Serviço Nacional de Saúde (SNS), pelos vários subsistemas de saúde públicos e privados, pelo setor segurador e pelo setor privado "puro", financiado por pagamentos diretos dos indivíduos (Silva, 2009). O SNS foi criado em 1979, estabelecendo o direito à proteção da saúde de todos os cidadãos, independentemente da sua condição económica e social.

Relativamente à procura por cuidados de saúde, entende-se que esta não depende apenas da vontade dos indivíduos ou dos seus recursos, é antes determinada pela perceção que o indivíduo tem da necessidade e da existência de um problema de saúde (Mulholland, et al., 2008). Pode-se referir que a procura por cuidados de saúde é influenciada: pela oferta; pelos custos implícitos no consumo de cuidados de saúde; por fatores culturais e demográficos; pela forma de financiamento dos cuidados de saúde (Rego, 2008). Em (Abreu, 2012) a autora sublinha que, segundo os economistas, o aumento desta procura irá depender também da capacidade económica (rendimento disponível) dos indivíduos e da sociedade porque um aumento no rendimento traduz-se num aumento da capacidade aquisitiva do indivíduo, levando a uma maior procura de cuidados de saúde (Rosko & R. Broyles, 1988). Estes autores destacam ainda que a formação académica do indivíduo tem impacto na procura de cuidados de saúde, isto porque se pode esperar que uma pessoa com maior formação académica compreenda melhor a importância que a obtenção de determinados cuidados tem para a sua saúde e ser mais eficiente na escolha que faz.

A saúde é uma das maiores preocupações dos cidadãos, seja no dispêndio financeiro ou na qualidade de serviços prestados, e isso traduz-se numa maior exigência por parte dos utentes face ao SNS (APS, 2009). Além disso, o SNS tem algumas dificuldades com que lidar, como o aumento do custo dos materiais, as novas tecnologias, a escassez de recursos, o envelhecimento populacional e as restrições financeiras (Abreu, 2012). Existem outros sistemas de comparticipação, complementares ou suplementares ao SNS, como é o caso dos Seguros de Saúde, que têm como objetivo apresentar alternativas aos cidadãos no financiamento e na qualidade de prestação dos cuidados de saúde.

Os seguros de saúde são produtos que asseguram o financiamento dos cuidados de saúde prestados aos seus beneficiários, com base em prémios ou quotizações que são suportados pelas próprias famílias (seguros individuais) ou pelas suas entidades patronais (seguros de grupo). O setor segurador, ao integrar o Sistema Nacional de Saúde, contribui para que alguns grupos da população passem a recorrer com menor frequência ao serviço público e, assim, contribui para a sustentabilidade do SNS. Neste trabalho, o universo de estudo será apenas o dos seguros de saúde individuais.

A procura por seguros de saúde individuais tem sido crescente e esta tendência pode dever-se ao elevado grau de indeterminação da quantidade e da duração dos cuidados de saúde que a pessoa necessita ao longo da sua vida (Abrantes, 1985). Em (Baptista, 2019), a autora concluiu que o preço, a percentagem de comparticipação/reembolso, o sexo e a idade têm uma grande influência na aquisição de um seguro de saúde em Portugal. Em (Guerra, 2014), evidenciou-se que a rapidez de acesso aos cuidados de saúde e a diferenciação da oferta relativamente aos serviços que o seguro apresenta são

determinantes na aquisição do seguro. E a Entidade Reguladora dos Seguros, em (ASF, 2021), refere que as pessoas procuram por uma maior liberdade de escolha e, simultaneamente, por uma melhoria global das condições de acesso. O facto de os indivíduos estarem cobertos por um seguro privado leva a uma melhoria no estado de saúde apercebido pelos indivíduos (Saúde, 2007). Em (Martinho, 2014), segundo a autora, o seguro de saúde é valorizado pela sociedade, uma vez que permite uma maior liberdade de escolha no acesso a cuidados de saúde, permite colmatar as falhas do SNS, oferece uma maior qualidade de serviço e possibilita um atendimento mais rápido. No que diz respeito ao perfil das pessoas que compram um seguro de saúde, em (Guiomar, 2010), conclui-se que a idade, o género, a existência de diabetes, a zona de residência, o nível de escolaridade, o rendimento e a perceção do estado de saúde têm impacto nesta decisão.

2.2. FATORES DE RISCO EM SAÚDE

A saúde é afetada por muitos fatores, sendo que os que estão associados a problemas de saúde, incapacidade, doença ou morte são conhecidos como fatores de risco. Um fator de risco é uma característica, condição ou comportamento que aumenta a probabilidade de ter uma doença ou lesão. Os fatores de risco podem ser classificados em:

- **Comportamentais** – Referem-se a uma ação tomada pelo indivíduo e, por isso, os fatores podem ser eliminados ou reduzidos através do estilo de vida ou escolhas comportamentais. Alguns exemplos de fatores de risco desta categoria são o tabagismo, o consumo de álcool em excesso, a inatividade física e as opções nutricionais.

- **Metabólicos** – São os fatores relativos à biologia do indivíduo e que podem ser influenciados por uma combinação de genética, estilo de vida e outros fatores. São exemplos o excesso de peso ou obesidade, a hipertensão arterial, o colesterol elevado e um nível de glicose elevado no sangue.

- **Ambientais/Ocupacionais** – Estes fatores abrangem uma vasta gama de tópicos como fatores sociais, económicos, culturais e políticos ou fatores físicos, químicos e biológicos. Alguns exemplos são o acesso à água potável, o saneamento e a poluição do ar.

Segundo as estimativas obtidas para Portugal, no âmbito do estudo *Global Burden of Diseases* (GBD) para 2019, os fatores de risco comportamentais são os que têm maior peso (25%) no número total de anos de vida saudável perdidos (DALYs), os fatores metabólicos tiveram um peso de 22% e os fatores ambientais/ocupacionais tiveram um peso de 7%, como pode ser visto em ANEXO 1. A Figura 2.1 apresenta com mais detalhe os fatores de risco com maior peso nos DAYLs, sendo que os cinco fatores com mais influência nos DAYLs são: a glicose plasmática em jejum elevada (10,6%), o tabagismo (10,4%), a pressão arterial elevada (8,8%), o índice de massa corporal elevado (7,9%) e os riscos dietéticos (7,3%).

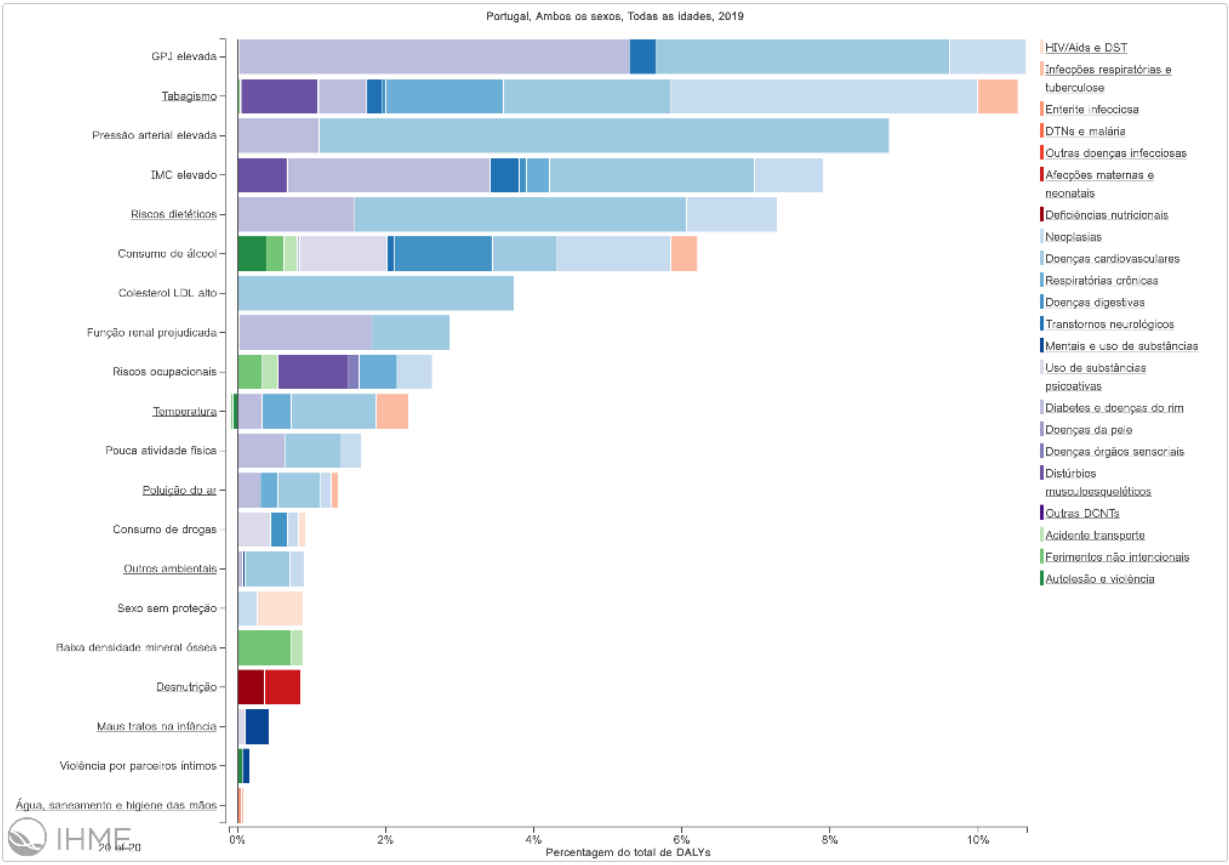


Figura 2.1 - Fatores de risco ordenados por peso no número total de DAYLs para Portugal, em 2019
Fonte: Elaborado com base nos dados para Portugal da GBD 2019

No âmbito dos Inquéritos Nacionais da Saúde, em (Vintém, 2008), concluiu-se que o género e a escolaridade têm impacto na auto-percepção do estado de saúde. Para o género, em geral, existe uma tendência para as mulheres subestimarem a apreciação positiva da sua saúde, enquanto os homens tendem a privilegiar essa mesma apreciação positiva. E para a escolaridade, as pessoas com níveis de escolaridade mais altos tendem a aderir melhor a medidas de prevenção da doença e promoção da saúde e corrigem com maior frequência os seus hábitos e estilos de vida menos saudáveis. No que diz respeito aos custos com despesas de saúde, em (Haynes & Dunnagan, 2002) concluiu-se que as pessoas com fatores de risco como pressão arterial, colesterol, IMC e fumadores tinham uma probabilidade aproximadamente 70% superior de reportar sinistros à Seguradora do que as pessoas que eram mais saudáveis.

Nos seguros de saúde, os fatores de risco são as características do risco que têm uma relação de causalidade com a sinistralidade, e que se levam em conta para o cálculo dos prémios. Para o seguro de saúde, é aceite que os preços individuais variem de acordo com a idade e outros indicadores de saúde que sejam justificáveis estatisticamente e sejam actuarialmente justos, e dependem de país para país (Duchêne & Boyer-Kassem, 2020).

Em (ASF, 2021), são descritos os tipos de seguros existentes em vários países da Europa. Em Portugal e Espanha, que têm um sistema de saúde idêntico, o risco de um seguro de saúde individual é mensurado pela idade. No caso do Reino Unido, para além da idade, o risco pode ser medido pelo facto de ser ou não ser fumador, ou a profissão. Na Alemanha, o risco do seguro de saúde

complementar ao seguro social que têm é dado pela idade e a história clínica. Na Bélgica, as Seguradoras definem os prémios de acordo com a idade do indivíduo e, por vezes, com o seu local de residência. Na Holanda, o seguro é social e cada segurado paga um prémio nominal, que não depende do risco individual, e um prémio indexado ao nível de rendimento, que é reembolsado pelo empregador e cujo risco é medido pela idade, nível socioeconómico e níveis de gastos com medicamentos e diagnósticos.

As escolhas dos indivíduos sobre os seguros de saúde são afetadas por uma variedade de características que não podem ser todas facilmente medidas, mas estas estão correlacionadas com outras características que são mais facilmente medidas como a idade, o género, a profissão e o rendimento. Para além destas, também está incluída a perceção do estado de saúde individual (e do marido/mulher e dos filhos), o conhecimento e preferências individuais na utilização e pagamento de cuidados de saúde, a sua capacidade de aceitar o risco e a sua capacidade de se envolver nas decisões sobre aderir, deixar ou continuar com um seguro de saúde (Shapiro, 1993).

2.3. MÉTODOS DE QUANTIFICAÇÃO DO RISCO

A tarifação de seguros consiste em estimar o risco, que é dado pela frequência de sinistralidade e pelo custo médio do sinistro. Para se obter este valor, existem várias abordagens que podem ser utilizadas e algumas delas são abordadas de seguida.

Distribuição Normal - Métodos baseados na distribuição normal são bastante utilizados para fazer inferências sobre a média da amostra ou regressões lineares. O problema destas abordagens é que são sensíveis a valores extremos e tendem a não ser eficientes em amostras pequenas ou médias se a distribuição não for realmente normal.

Métodos Não-Paramétricos - Duas abordagens comuns para avaliar diretamente os custos em ensaios aleatórios são o Teorema Limite Central e a metodologia *Bootstrap* (O'Hagan & Stevens , 2003). A primeira abordagem assenta na normalidade da média dos custos com base numa amostra de grande dimensão, independentemente da distribuição dos custos da população. O *Bootstrap* utiliza a distribuição empírica da média dos custos da amostra e com base na reamostragem estima a média das amostras de cada replicação feita. A utilização de vários *Bootstraps* e sugestão das abordagens mais recomendadas é apresentada por (Barber & Thompson, 2000).

Modelos Lineares Generalizados - Neste tipo de modelos, a média da variável resposta é dada de forma linear por variáveis explicativas da mesma. Têm sido estudadas extensões a esta abordagem relativamente à família de distribuições possíveis, à relação entre a média e a variância (Basu & Rathouz , 2005) e a estimadores que sejam mais robustos face a *outliers* (Cantoni & Ronchetti, 2001). O MLG mais usado em métodos de tarifação utiliza a função de ligação logarítmica e, apesar de mostrar algumas perdas de eficiência quando a variância do erro logarítmico é grande, quando a cauda é pesada, esta tem mostrado mais eficiência face a outras formas de distribuição (Manning & Mullahy, 2001).

Modelos multi-partes - Este tipo de modelos parece ser o mais flexível de incorporar os problemas dos dados, nomeadamente o excesso de zeros, a sobredispersão e as caudas pesadas, levando assim

a estimativas mais robustas. Estes modelos podem incorporar ainda a combinação de várias distribuições, o que poderá fazer sentido, pois as covariáveis podem não afetar as variáveis resposta da mesma maneira. Têm sido estudadas combinações de Poisson (Mullahy, 1997), de Binomial Negativa (Deb & Holmes, 2000) no que diz respeito à modelação da utilização e, para o custo, combinação de várias distribuições Gama (Deb & Holmes, 2000). As estimativas dos modelos com distribuições combinadas costumam ter melhores resultados que a utilização baseada numa só distribuição, mas a nível computacional fica mais exigente; nem sempre é fácil de identificar quando existem muitas componentes para modelar e algumas distribuições tendem a sobrepor-se. Um caso particular deste tipo de modelos é o modelo de duas partes em que só são permitidas duas componentes e uma delas é degenerativa, ou seja, a distribuição que a suporta consiste em apenas um valor; além disso, neste caso particular, as componentes são separadas e estimadas de forma independente, ao contrário dos outros modelos em que há combinação de distribuições. Esta abordagem é comum quando se pretende prever a taxa de utilização e se tem um grande número de não utilizadores.

Em (Mihaylova, Briggs, O'Hagan, & Thompson, 2011), ao apresentarem uma revisão dos métodos estatísticos para analisar a utilização e os custos de saúde, concluíram que os Modelos Lineares Generalizados são uma abordagem atrativa quando se tem covariáveis; permitem diferentes tipos de distribuições, é necessário é ter cuidado com a função de ligação escolhida. Além disso, quando se tem um grande número de zeros deve-se utilizar modelos de duas partes.

Nos dias de hoje, cada vez mais se fala em técnicas de *Machine Learning* (ML), termo este que surgiu quando, em 1959, o Engenheiro Arthur Samuel do MIT estava a construir uma máquina autónoma que funcionava com base na aprendizagem e melhoria. O autor da máquina descreveu o conceito como um campo de estudo que dá a computadores a capacidade de aprender, sem ter sido programado para fazê-lo (Samuel, 1959). Apesar do potencial que estas máquinas aparentavam ter, só mais tarde, com a evolução tecnológica e a criação de sistemas capazes de suportar bastante informação, é que estas técnicas ganharam relevo e começaram a ser utilizadas.

Esta máquina de aprendizagem auxilia no tratamento de um grande volume de dados, porque permite identificar relações e padrões que não são de fácil visualização pelo ser humano. Atualmente, com a crescente automatização de processos em muitos serviços, gera-se uma quantidade grande de dados que ficam armazenados, mas que nem sempre são estruturados e organizados consoante a sua relevância. Algumas das vantagens de ML são a redução de custos, aumento da eficiência e da produtividade e uma melhor gestão de riscos, no entanto, há algumas desvantagens que precisam de ser tidas em conta, como a sensibilidade a *outliers* e previsões erradas (Martin Leo, 2019). Por isso, o papel do ser humano é fundamental para fazer o julgamento necessário dos resultados obtidos e garantir a monitorização e a avaliação destas máquinas.

Em (Zhuang, 2013), foram utilizados dois métodos: um mais tradicional (Modelos Aditivos Generalizados (MAG)) e outro método de *Machine Learning* (Redes Neuronais). Nos MAG, que face aos MLG permitem que a relação entre o preditor linear e a variável resposta não seja linear, foram identificadas algumas limitações como a interação entre as variáveis explicativas, a colinearidade, o sobreajustamento e o pressuposto paramétrico. E as redes neuronais tiveram um excelente resultado

na previsão, mas também apresentaram a desvantagem do sobreajustamento. Mesmo assim, entre os dois métodos, as redes neuronais foram as que obtiveram melhores resultados.

2.4. DIFERENCIAÇÃO DO RISCO

Do regime jurídico do contrato de seguro, para o ramo Doença, importa salientar a proibição de práticas discriminatórias, previstos no artigo 15º. Nomeadamente, tal como previsto no artigo 13º da Constituição Portuguesa, a proibição de discriminação em razão de ascendência, sexo, raça, língua, território de origem, religião, convicções políticas ou ideológicas, instrução, situação económica, condição social ou orientação sexual. O princípio da igualdade é violado quando uma pessoa, em razão de deficiência ou de risco agravado de saúde, é tratada de maneira diferente de uma outra pessoa em situação comparável, nos termos da Lei nº46/2006, de 28 de agosto. A exclusão (da pessoa ou da doença) ou o agravamento do prémio deve ser objetivamente fundamentado, tendo por base dados estatísticos e atuariais rigorosos considerados relevantes nos termos dos princípios da técnica Seguradora e estão sujeitas à supervisão da Autoridade de Supervisão de Seguros e Fundos de Pensões (ASF). Assim, todos os fatores que a Seguradora utiliza para diferenciar o risco dos segurados devem ser tecnicamente justificáveis e a Seguradora deve prestar, ao proponente, informação sobre o rácio entre os fatores específicos e os fatores de risco de uma pessoa em situação comparável, mas não afetada por aquela deficiência ou risco agravado de saúde, nos termos dos n.os 3 a 6 do artigo 178º. No que diz respeito à vigência do contrato, não pode haver agravamento do prémio ou exclusão para doenças diagnosticadas (art.º 215.º da LCS).

Existem quatro princípios que um atuário deve seguir para tarifar um seguro: adequabilidade, razoabilidade, competitividade e equidade. O prémio deve ser adequado, na medida em que o prémio pago pelo segurado deve ser suficiente para cobrir as indemnizações e os custos da empresa. Deve ser razoável, ou seja, deve valer pela cobertura e pelos serviços que estão a ser disponibilizados. É importante ser competitivo, para atrair bons clientes (clientes com menos risco) e aumentar as vendas. Por último, deve respeitar a equidade, ou seja, pessoas com o mesmo risco devem pagar o mesmo prémio.

As características do segurado são determinantes na estimação do prémio a cobrar a cada apólice de seguro. Contudo, a diferenciação dos segurados em classes distintas de risco, que se traduzem em prémios pagos diferentes, tem os seus limites. Mais recentemente, com a evolução da tecnologia, nomeadamente a capacidade de recolher, armazenar e trabalhar uma grande quantidade de dados, algumas questões têm sido levantadas sobre a possibilidade de personalização da oferta. Em (McFall, 2019) define-se o mercado da personalização como a "combinação das novas tecnologias, técnicas de análise de correlação e metodologias que através do *Big Data* conseguem fazer recomendações e ofertas personalizadas.

No setor segurador existe um princípio base que é o da mutualização do risco; a tarifa que indica o prémio a pagar por pessoas com o mesmo risco deve assegurar as futuras indemnizações, mesmo que, em alguns casos, o prémio pago por um segurado não seja suficiente para cobrir as suas despesas. Na população considerada na tarifa, os segurados "compensam-se" uns aos outros e, por isso, no total a Seguradora consegue cumprir com as obrigações. O desenvolvimento da estatística durante o século XIX mostra evidências de que existe alguma regularidade dos eventos ao nível coletivo que não podem

ser explicadas ao nível individual (Foucault, 2009). Segundo (Knight, 1985), os seguros podem ser definidos como a transformação da incerteza de cada indivíduo num risco agregado mensurável. Isto é possível, se tivermos uma população suficientemente grande, aplicando a lei dos grandes números. Assim, o risco tem de ser agrupado, porque não pode ser calculado ao nível individual; só quando este é disperso pela população é que se consegue calcular. O trabalho da Seguradora consiste então em constituir essa população selecionando os riscos e dividindo-os.

A personalização do risco ao nível do indivíduo, segundo (Barry & Charpentier, 2020), tem os seus desafios em grande parte devido à resiliência conceitual do seguro descrita anteriormente; por isso, este autor concluiu que a gestão é e continuará a ser sobre a gestão coletiva de eventos incertos, que exige um conhecimento imperfeito dos indivíduos. Além disso, o seguro de saúde, em particular, tem uma regulação mais exigente, o que dificulta a introdução ou utilização destes novos mecanismos. Em (McFall, 2019), reforça-se a importância da solidariedade entre os grupos tarifários que equilibram a distribuição do risco. Com a personalização, o autor diz que não fica claro se qualquer eventual produto se qualificaria como seguro para a Seguradora.

Os fatores que devem ou podem ser levados em conta constituem um ponto de debate; alguns argumentam que nenhum parâmetro deve ser banido *a priori*; em (Walters, 1981) afirma-se assim que "não se deve legislar contra o uso do conhecimento numa sociedade livre". Outros consideram que os fatores de risco que não estão sob o controlo do segurado (como a idade ou o género) devem ser evitados (Landes, 2015). Outros gostariam ainda de introduzir fatores socioeconómicos que, dependendo do ramo do seguro, não refletem necessariamente riscos, a fim de promover a justiça social (De Witt & Van Eeghen, 1984). Em (Liukko, 2010), a classificação de risco é considerada um requisito indispensável para o funcionamento sustentável do seguro privado e isso é uma razão aceitável para discriminação, mesmo que noutras áreas da vida económica essa discriminação seja proibida.

Na Europa, o regulamento da proteção de dados tem o princípio de informações justas em que se defende que os dados devem ser minimizados e só podem ser usados para os fins para os quais foram recolhidos. O autor, em (McFall, 2019), afirma que, com a recolha de mais dados, a oferta dos seguros pode ser reorganizada, talvez até personalizada, mas tem de ser sempre regulamentada.

3. METODOLOGIA

Sendo o principal objetivo desta dissertação identificar as variáveis diferenciadoras do risco nos seguros de saúde individuais, é necessário começar por quantificar o risco. O risco pode ser mensurado pela frequência de sinistros multiplicado pelo seu custo médio, por isso estas serão as duas variáveis resposta que se pretendem avaliar. Tal como apresentado no capítulo 2, assume-se que estas são independentes e ajusta-se um modelo linear generalizado para cada uma.

No caso da distribuição do custo dos sinistros, existe a particularidade de esta apresentar, usualmente, uma cauda com valores extremos, ou seja, existe um pequeno conjunto da amostra com uma percentagem relevante do custo total com sinistros. Com este tipo de fenómeno, nem sempre é possível ajustar apenas uma distribuição aos custos totais. Uma abordagem frequente consiste em definir um valor $s > 0$ acima do qual se encontram os "grandes" sinistros, ou seja, definir dois tipos de sinistros $Y \leq s$ e $Y > s$. Assim, o valor esperado da variável resposta Y é dado por:

$$E[Y] = E[Y|Y \leq s]P[Y \leq s] + E[Y|Y > s]P[Y > s] \tag{3.1}$$

A determinação do nível ótimo de s é essencial para este tipo de modelação. Segundo o autor em (Santos, 2003), este valor não deve ser demasiado elevado, para não conduzir a uma grande variância; mas também não deve ser demasiado baixo, pois pode originar num maior enviesamento. Este valor deve ser efetuado com base na experiência do atuário, mas uma abordagem possível, apresentada em (Guerreiro, 2016), é analisar um quantil de elevada probabilidade para a distribuição do custo dos sinistros e/ou o nível de retenção do contrato de resseguro, salvaguardando a robustez estatística da modelação dos "grandes" sinistros.

Neste capítulo, serão apresentados os modelos lineares generalizados, em particular, o caso da regressão logística, da regressão gama e da regressão multinomial. Será ainda abordada a distribuição Pareto Generalizada, uma distribuição de extremos. Através dos modelos de regressão descritos anteriormente, será possível avaliar quais as variáveis explicativas da utilização do Internamento e/ou do custo médio associado, e mensurar os seus impactos. Além dos modelos de regressão, vão ser apresentados dois métodos de *Machine Learning*: Análise de *Clusters* e Árvores de Decisão, com o intuito de identificar os níveis de risco existentes nas variáveis explicativas e agrupá-los, caso sejam homogéneos.

3.1. MODELOS LINEARES GENERALIZADOS

Quando se pretende estudar a relação entre variáveis, ou mais particularmente, analisar a influência que uma ou mais variáveis explicativas têm sobre uma variável de interesse, que denominamos de variável resposta, um modelo de regressão é uma metodologia possível. O modelo de regressão linear normal foi apresentado inicialmente por Legendre e Gauss, no início do século XIX, e dominou a modelação estatística até meados do século XX. Contudo, este modelo tinha alguns requisitos que nem sempre eram fáceis de cumprir. Assim, Nelder e Wedderburn (1972) introduziram os Modelos Lineares Generalizados (MLG), com o objetivo de sintetizar alguns modelos que já tinham sido criados e que se baseavam em regressões lineares, mas que não cumpriam todos os requisitos do mesmo. Alguns exemplos desses modelos são, tal como refere (Lindsey, 1997), o modelo complementar log-log para

ensaios de diluição, os modelos probit e logit para proporções, os modelos log-lineares para dados de contagens e os modelos de regressão para análise de sobrevivência. Para além de apresentarem uma estrutura de regressão linear, estes modelos têm em comum o facto de a variável resposta seguir uma distribuição da Família Exponencial e haver uma estrutura de combinação linear entre regressores. Assim, os MLG passaram a englobar um grande número de modelos, que são facilmente utilizados devido ao rápido desenvolvimento computacional que se tem verificado nas últimas décadas.

3.1.1. Estrutura MLG

Os MLG estão estruturados em três componentes:

- **Componente aleatória** – esta componente é dada pela distribuição condicional da variável resposta Y_i, para o i-ésimo elemento de n observações independentes e identicamente distribuídas, dados os valores das variáveis explicativas no modelo.

$$E(Y_i|\boldsymbol{X_i}) = \mu_i = b'(\theta_i),\ i = 1, \dots, n \tag{3.2}$$

- **Componente sistemática** – esta componente representa uma função linear de regressores, o preditor linear

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ik},\ i = 1, \dots, n, k = 1, \dots, p \tag{3.3}$$

O vetor de regressores $\boldsymbol{X_i}^{\boldsymbol{T}} = (X_{i1}, \dots, X_{ip})$ são funções pré-especificadas das variáveis explicativas, onde $X_{ik}, k = 1, \dots, p$, representa a k-ésima covariável para o i-ésimo indivíduo, e $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ um vetor de parâmetro desconhecidos.

- **Função de Ligação** – esta componente é composta pela função de ligação $g(.)$, que transforma o valor esperado da variável resposta, $\mu_i = E(Y_i)$, no preditor linear:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ik},\ i = 1, \dots, n, k = 1, \dots, p \tag{3.4}$$

A função de ligação é uma função invertível, por isso também se pode escrever:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\big(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ik}\big),\ i = 1, \dots, n, k = 1, \dots, p \tag{3.5}$$

Esta função descreve a relação funcional entre a componente sistemática e o valor esperado da componente aleatória (a média da variável dependente).

De forma geral, diz-se que uma distribuição é da Família Exponencial se a sua função de probabilidade pode ser escrita na seguinte forma:

$$f(y|\theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{3.6}$$

y – variável resposta
θ – parâmetro canónico e está relacionado com a média da variável resposta

ϕ – parâmetro de escala ou de dispersão, é estritamente positivo e está relacionado com a variância da variável resposta
a, b e c – funções específicas que determinam unicamente a distribuição

As suas propriedades são:

$$\mu = E(Y) = b'(\theta) \tag{3.7}$$

$$\sigma^2 = Var(Y) = b''^{(\theta)} a(\phi) \tag{3.8}$$

3.1.2. Ajustamento dos Modelos

Estimação dos parâmetros

Num modelo linear generalizado o parâmetro de interesse é β, sendo estimado pelo método da máxima verosimilhança. O parâmetro de dispersão ϕ, quando existe, é estimado pelo método dos momentos. A função de verosimilhança do modelo, em função de β, é dado por (Turkman, 2000):

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\theta_i, \phi) = \prod_{i=1}^{n} exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\} = exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} + \sum_{i=1}^{n} c(y_i, \phi)\right\} \tag{3.9}$$

O logaritmo da função de verosimilhança é dado por:

$$\ln\big(L(\beta)\big) = l(\beta) = \sum_{i=1}^{n} \left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\} = \sum_{i=1}^{n} l_i(\beta) \tag{3.10}$$

l_i – contribuição de cada observação y_i para a verosimilhança

Os estimadores de máxima verosimilhança para β são obtidos a partir da solução do sistema de equações de verosimilhança. Essas equações são dadas por:

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta_k} = 0, k = 1, \dots, p \tag{3.11}$$

A equação é a derivada do logaritmo da verosimilhança em relação ao parâmetro β e pode-se chamar de *Score*. Estas equações de máxima verosimilhança não têm, em geral, uma solução analítica, sendo necessário recorrer a métodos numéricos para a sua resolução. Tendo em vista os MLG, (Wedderburn, 1972) construíram um algoritmo que permite resolver estas equações, o que em muito contribuiu para o sucesso destes modelos, por ser fácil de implementar a nível computacional e adaptável aos vários MLG. O algoritmo é denominado de Método Iterativo de Mínimos Quadrados Ponderados e baseia-se no método dos *Scores* de *Fisher*.

Método dos *Scores* de *Fisher*

Este método é uma generalização do método de *Newton-Raphson* e difere deste na medida em que substitui a segunda derivada, ou a matriz Hessiana, pelo seu valor esperado. A função *Score* é dada por:

$$S(\beta) = \frac{\partial l(\beta)}{\partial \beta} \tag{3.12}$$

A matriz de covariância da função *Score* é designada por matriz de Informação de *Fisher* e consiste no simétrico da matriz Hessiana:

$$I(\beta) = E\left(-\frac{\partial S(\beta)}{\partial \beta}\right) = E\left(-\frac{\partial^2 l_i(\beta)}{\partial \beta_k \partial \beta_k}\right) \tag{3.13}$$

O método iterativo é dado por:

$$\beta_{k+1} = \beta_k + I(\beta_k)^{-1} S(\beta_k) \tag{3.14}$$

com β_k – estimativa de β na k-ésima iteração

Como critério de paragem, é usual limitar o erro absoluto, ou seja, para um tal ε o método é interrompido e considera-se como solução x_k, este é dado por:

$$\|x_k - x_{k-1}\| \leq \varepsilon \tag{3.15}$$

Depois de terem sido obtidas as estimativas dos coeficientes da regressão, é necessário avaliar a qualidade de ajuste do modelo. Importa verificar a significância dos coeficientes estimados, ou seja, verificar se existe uma associação estatisticamente significativa entre as variáveis explicativas e a variável resposta. Para isso, será utilizado o teste de *Wald* e o teste de razão de verosimilhanças. Existem ainda métodos de seleção de variáveis que indicam quais as variáveis que têm mais importância para a variável resposta. Os métodos de seleção *Stepwise*, com base no critério de Informação de *Akaike*, calculam o quanto uma variável independente explica a variável resposta e a *Deviance* é apresentada como uma medida de avaliação de ajuste do modelo. Por fim, será introduzido o teste de *Tukey*, um teste de comparações múltiplas.

Teste de *Wald*

O teste de *Wald* permite inferir se um determinado parâmetro tem influência ou não no modelo, através do teste de hipóteses da nulidade do parâmetro:

$$H_0: \beta_k = 0 \; vs \; H_1: \beta_k \neq 0, k = 1, \dots, p \tag{3.16}$$

A estatística de teste, sob a validade de H_0 é dada por:

$$W_j = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \sim N(0,1) \tag{3.17}$$

$se(\hat{\beta}_j)$ – erro padrão da estimativa de máxima verosimilhança

Teste de razão de verosimilhanças

O teste de razão de verosimilhanças é utilizado quando se pretende comparar a qualidade de ajustamento de dois modelos encaixados, ou seja, modelos em que um tem o subconjunto de variáveis de outro modelo. Além disso, também permite verificar se o modelo é globalmente significativo, através do teste de significância dos coeficientes estimados simultaneamente. Dados dois modelos encaixados, M_p e M_q, com um número de variáveis p e q, respetivamente, tal que $p < q$, para comparar a qualidade de ajustamento de dois modelos aplica-se o teste de razão de verosimilhanças, sob a

hipótese de que as $q - p$ variáveis no modelo não apresentam acréscimo significativo na qualidade do modelo. O teste de hipóteses a testar é:

$$H_0: As\ q - p\ variáveis\ no\ modelo\ não\ são\ significativas$$
$$vs$$
$$H_1: As\ q - p\ variáveis\ no\ modelo\ são\ significativas \tag{3.18}$$

A estatística de teste, sob a validade de H_0 é dada por:

$$G = -2\left[\frac{\ln\left(L_{M_p}(\beta)\right)}{\ln\left(L_{M_q}(\beta)\right)}\right] \sim \chi^2_{q-p} \tag{3.19}$$

$\ln\left(L_{M_p}(\beta)\right)$ – função logaritmo da verosimilhança do modelo M_p com p variáveis

$\ln\left(L_{M_q}(\beta)\right)$ – função logaritmo da verosimilhança do modelo M_q com q variáveis

Métodos de seleção *Stepwise*

Quando se está perante inúmeras variáveis em que se pretende avaliar se são significativas para uma variável resposta, existe uma metodologia que escolhe as variáveis significativas a reter no modelo e que se denomina de ***Stepwise***. Existem duas abordagens, o ***Forward***, que consiste em começar com o modelo nulo (sem nenhuma variável) e ir adicionando as variáveis, uma a uma, caso contribuam para explicar a variável resposta; e o ***Backward***, em que o modelo começa com todas as variáveis e é retirada, uma a uma, a variável menos significativa. Em ambas as abordagens, é através do resultado do teste de Wald que se escolhe a variável a adicionar ou a retirar, e para avaliar a importância para explicar a variável resposta é feito o teste de razão de verosimilhanças.

Critério de Informação *Akaike* (AIC)

O critério de informação de *Akaike* foi desenvolvido por Hirotugu Akaike e proposto em 1974. Esta medida é uma estatística que tem por base o logaritmo da verosimilhança e penaliza o modelo com muitas variáveis. Com base neste critério, pretende-se examinar a complexidade do modelo e avaliar como se ajusta aos dados em estudo. A medida AIC é dada por:

$$AIC = -2[Log(L) - k] \tag{3.20}$$

k – número de parâmetros do modelo

L – valor da verosimilhança para o modelo estimado

O AIC permite comparar modelos, encaixados ou não encaixados, e escolher o melhor entre eles. Quanto menor for o valor do AIC menor será a informação perdida e, por isso, melhor será o ajustamento do modelo.

Deviance

A *deviance* é uma medida estatística que avalia a significância dos coeficientes estimados e tem por base o teste de razão de verosimilhanças. Considerem-se dois modelos, o primeiro com a variável

presente e o segundo sem essa variável. O teste da razão de verosimilhanças permite afirmar que, sob a hipótese de o modelo com a variável presente ser o verdadeiro modelo, a *deviance* é dada por:

$$D = -2Ln\left[\frac{L(modelo\ com\ a\ variável)}{L(modelo\ saturado)}\right] \sim \chi^2_{n-q} \tag{3.21}$$

O valor D representa o desvio do modelo ajustado em relação ao modelo saturado. Para avaliar a significância de uma variável explicativa no modelo, calcula-se a diferença entre o valor da *deviance* do modelo sem a variável e o valor da *deviance* do modelo com a variável. Quanto mais próximo o modelo ajustado, $\hat{\mu}$, estiver dos dados observados, y, menor será o valor de D, ou seja, um valor mais alto de *Deviance* indica um pior ajuste.

Teste de Tukey

Este teste foi desenvolvido por John Wilder Tukey e está apresentado em (Tukey, 1949). É um teste de comparações múltiplas que permite comparar a média dos níveis de uma variável categórica entre si. Caso o valor-p do teste seja inferior ao nível de significância pretendido, considera-se que existe uma diferença significativa na média de um dos níveis da variável, ou seja, os níveis são distintos entre si.

3.1.3. MLG na Tarifação de Seguros

No caso particular da tarifação de seguros, a fácil aplicação dos MLG não foi exceção e tem assumido, nas últimas décadas, um papel importante na construção de tarifas de seguros. Com base em informação histórica, o modelo identifica as variáveis que têm impacto no risco e, assim, estima-se o montante de indemnizações a pagar e o prémio necessário a cobrar. Devido ao resultado aleatório que advém do risco, a matemática e a estatística apresentam conceitos fundamentais como amostra representativa e a lei dos grandes números que permitem quantificar o risco. Sendo a mensuração do risco feita com base na **Frequência** e no **Custo Médio** do sinistro, estas serão as variáveis dependentes modeladas pelo GLM. Ajusta-se um modelo para cada uma delas porque as causas que explicam o seu comportamento nem sempre são as mesmas e porque as duas variáveis têm diferentes naturezas estatísticas, a frequência (**N**$(\boldsymbol{t})$ – número de sinistros ocorridos no intervalo de tempo é $[0, t[$) é uma variável discreta e o custo $\boldsymbol{X_i}$ – montante do i-ésimo sinistro, $i = 1, \dots, N$) é uma variável contínua e positiva. O montante agregado de perda é dado por:

$$S(t) = \sum_{i=0}^{N(t)} X_i \tag{3.22}$$

Quando as variáveis $\boldsymbol{X_i}$ são identicamente distribuídas, para $i = 1, \dots, N$, tem-se que o **Prémio de Risco** é dado por:

$$P(t) = E[N(t)] \times E(X) \tag{3.23}$$

3.1.4. Modelo de Regressão Logística e Logística Multinomial

No modelo de regressão logística, a partir de um dado conjunto de variáveis explicativas, é possível modelar uma variável dependente de natureza discreta com dois resultados possíveis: a presença de uma característica, ou a ausência da mesma (Agresti, 1996). Ou seja, trata-se de uma variável dicotómica podendo receber os valores 0 e 1, consoante se determinado fenómeno ou comportamento se verifica, ou pelo contrário, não se verifica. O modelo permite-nos estimar a probabilidade de um determinado evento ocorrer, face um conjunto de variáveis explicativas.

Formulação

Uma variável aleatória $Y \sim Binomial(n, p)$, onde Y é o número de sucessos entre n tentativas, dado uma probabilidade de sucesso p, a sua função massa de probabilidade é dada por:

$$f(\mathrm{y}|\mathrm{n},\mathrm{p}) = \binom{n}{y} p^y (1-p)^{n-y} \tag{3.24}$$

sendo $\mathrm{E}[Y] = np$ e $\mathrm{V}[Y] = np(1-p)$.

Esta distribuição pertence à Família Exponencial porque a sua função massa de probabilidade pode ser escrita da seguinte forma:

$$\begin{aligned} f(\mathrm{y}|\mathrm{n},\mathrm{p}) &= \exp\left[\log\binom{n}{y} + ylog(p) + (n-y)\log(1-p)\right] = \\ &= exp\left[\mathrm{y}\log\left(\frac{p}{1-p}\right) - (-n\log(1-p)) + \log\binom{n}{y}\right] \end{aligned} \tag{3.25}$$

sendo $\theta = \log\left(\frac{p}{1-p}\right), b(\theta) = -n\log(1-p)$ e $c(y,\phi) = \log\binom{n}{y}$.

Sendo a distribuição Bernoulli um caso particular da distribuição Binomial, $Y \sim Binomial(1,p) \Leftrightarrow Y \sim Bernoulli(p)$, também a distribuição Bernoulli faz parte da Família Exponencial.

Pelo primeiro termo da equação (3.25), pode ver-se que a função de ligação utilizada é o *logit,* que é o logaritmo da razão entre a probabilidade de sucesso e a probabilidade de insucesso, a sua equação é:

$$logit(p) = \log\left(\frac{p}{1-p}\right) \tag{3.26}$$

Consideremos X o conjunto de p covariáveis, $x_1, \ldots., x_p$, a probabilidade de o acontecimento de interesse ocorrer é dada por:

$$p = P(Y=1|X) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}, i = 1, \ldots, n \tag{3.27}$$

Para mais informações sobre este modelo, recomenda-se a leitura de (Hosmer, Lemeshow, & Rodney, 2013).

Quando se tem uma variável dependente com mais de duas categorias, deve-se optar por um modelo logístico multinomial se se pretende prever a probabilidade de ocorrência de cada categoria. Este modelo é uma extensão da regressão logística, por isso, o método de estimação das probabilidades para a variável resposta (condicionadas por covariáveis $X = (X_1, \ldots, X_k)$) e os pressupostos são similares ao modelo de regressão logística com a transformação *logit*. Considerando a variável dependente Y com $m+1$ categorias codificadas por $0, 1, \ldots, m$. Tal como na regressão logística binária, uma das categorias da variável dependente vai ser escolhida como categoria de referência e, na estimação dos *odds ratios*, cada uma das outras categorias é comparada com essa referência. Assim,

o modelo de regressão multinomial consiste num conjunto de $m+1$ modelos de regressão logística corrigidos, um para cada uma das $m+1$ categorias da variável dependente (Maroco J. , 2018).

Formulação

A distribuição multinomial generaliza a distribuição binomial permitindo que a variável resposta tenha mais do que um resultado possível dentro de categorias. A função massa de probabilidade para a distribuição multinomial é dada por:

$$f(\mathrm{y}|n,\mu) = \left(\frac{n!}{y_1!y_2!\dots y_m!}\right)\prod_{i=1}^{m} p_i^{y_i} = n!\prod_{i=1}^{m}\frac{p_i^{y_i}}{y_i!} \tag{3.28}$$

As *odds* (razão entre a probabilidade de sucesso e de insucesso) para cada categoria $j = 1, \dots, m$ em relação à categoria de referência $(Y = 0)$ são dadas por:

$$odds_j = \frac{P(Y=j|X)}{P(Y=0|X)} \tag{3.29}$$

A expressão geral para a probabilidade de se observar uma determinada categoria j, com $j = 0, 1, \dots, m$, onde todos os coeficientes de regressão da categoria tomada como referência $(j = 0)$ são nulos:

$$P(Y=j|X) = \frac{e^{\beta_{j0}+\sum_{i=1}^{k}\beta_{ji}X_i}}{1+\sum_{l=1}^{m}e^{\beta_{l0}+\sum_{i=1}^{k}\beta_{li}X_i}}\ , j = 0, 1, \dots, m. \tag{3.30}$$

Para mais informações sobre este modelo, recomenda-se a leitura de (Agresti, 2013).

3.1.5. Modelo de Regressão Gama

O custo dos sinistros é uma variável resposta contínua, estritamente positiva e apresenta assimetria à direita, como é usual no caso dos montantes das indemnizações a pagar. Assim, uma distribuição que geralmente se ajusta bem a este tipo de dados é a distribuição Gama. Segundo (Johnson P. E., 2014), esta distribuição é adequada quando se suspeita que a ligação entre a média e a variância é "fixa", ou seja, se se esperar um valor pequeno da variável resposta, deve se esperar também um pequeno valor na variância.

Admitindo que a variável aleatória $Y \sim Gama(\alpha, \beta)$, a função de densidade de probabilidade de distribuição $Gama(\alpha, \beta)$ é dada por:

$$f(\mathrm{x}) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}y^{\alpha-1}e^{-y/\beta}, y \geq 0, \alpha, \beta > 0 \tag{3.31}$$

sendo $\Gamma(\alpha) = \int_0^{+\infty} y^{\propto-1}e^{-y}dy$, o integral converge $\alpha > 0$ e define a função conhecida por Função Gama. Além disso, $\mathrm{E}[Y] = \frac{\alpha}{\beta}$ e $\mathrm{V}[Y] = \frac{\alpha}{\beta^2}$.

Para mais informações sobre este modelo, recomenda-se a leitura de (McCullagh & Nelder, 1989).

3.1.6. Distribuição Pareto Generalizada

Uma distribuição possível e adequada para modelar os "grandes" sinistros é a distribuição Pareto Generalizada (GDP). Esta distribuição foi introduzida por Pickands, em 1975, para modelar excedentes acima de um limiar.

Formulação

Uma variável aleatória Y segue uma distribuição GDP, com parâmetros de forma, localização e escala designados por $\varepsilon, \mu \in \mathbb{R}$ e $\sigma > 0$, respetivamente, se a sua função densidade de probabilidade é dada pela expressão:

$$f(y|\varepsilon,\mu,\sigma) = \begin{cases} \frac{1}{\sigma}(1+\varepsilon\frac{y-\mu}{\sigma})^{-\frac{1}{\varepsilon}-1}, & \varepsilon \neq 0 \\ \frac{1}{\sigma}(1+\varepsilon\frac{y-\mu}{\sigma})^{-\frac{1}{\varepsilon}-1}, & \varepsilon \neq 0 \end{cases} \tag{3.32}$$

Se $\varepsilon > 0$ então $y \geq \mu$, enquanto, se $\varepsilon < 0$ temos $\mu \geq y \geq \mu - \frac{\sigma}{\varepsilon}$.

A média e a variância são dadas pela expressão:

$$\mathrm{E}[Y] = \mu + \frac{\sigma}{1-\varepsilon}, \varepsilon < 1 \tag{3.33}$$

$$\mathrm{V}[Y] = \frac{\sigma^2}{(1-\varepsilon)^2(1-2\varepsilon)}, \varepsilon < 1/2 \tag{3.34}$$

A distribuição Pareto Generalizada é uma distribuição bastante versátil, dado que muitas distribuições de probabilidade são casos particulares desta, consoante o valor que os seus parâmetros tomam, por exemplo:

- Se $\varepsilon = 0$ e $\mu = 0$, reduz-se à distribuição Exponencial de valor médio σ.
- Se $\varepsilon = 0$ e $\mu = \frac{\sigma}{\varepsilon}$, coincide com a distribuição Pareto com parâmetro de forma $\frac{1}{\varepsilon}$ e de escala $\frac{\sigma}{\varepsilon}$.
- Se $\varepsilon = -1$ e $\mu = 0$, reduz-se à distribuição Uniforme em $(0, \sigma)$.

Para mais informações sobre este modelo, recomenda-se a leitura de (Pickands, 1975).

3.1.7. Vantagens e Desvantagens dos MLG

Os modelos lineares generalizados, para além das já apresentadas vantagens que trazem à tradicional regressão linear, estão bem estabelecidos em termos de literatura, aceitação regulatória e software disponível; pela experiência estão bem testados, tendo sido encontrados resultados significativos em dados de Seguradoras; são facilmente interpretáveis, percebendo o caminho percorrido pelo modelo desde o *input* até ao *output*. Por outro lado, existem ainda alguns pressupostos que limitam o modelo, como a função de ligação ou a função do erro; não conseguir incorporar variáveis explicativas correlacionadas; não detetar, diretamente, as interações existentes relevantes e a não linearidade; a sensibilidade a *outliers*; precisar de uma quantidade de dados significativa, sendo que quantas mais variáveis explicativas se queira incluir, maior tem de ser o tamanho da amostra.

Recentemente, têm surgido métodos de outro tipo que têm sido apontados como o futuro do *Big Data*. Esses são os modelos de *Machine Learning* (ML), um subconjunto da Inteligência Artificial, que, comparativamente com os MLG, podem resolver, na maioria, o mesmo problema de perspetivas diferentes; não é necessário cumprir com tantos pressupostos. Em termos de previsibilidade, acredita-se que, em geral, os modelos de ML têm capacidade superior aos MLG.

Neste trabalho utilizou-se tanto os MLG como modelos de ML, nomeadamente Análise de *Clusters* e Árvores de Decisão, cuja metodologia é apresentada de seguida.

3.2. ANÁLISE DE *CLUSTERS*

A análise de *clusters* é a arte de encontrar grupos nos dados (Leonard & Rousseeuw, 1990). Pode ser considerado um algoritmo de classificação ML, com base em aprendizagem não supervisionada, que consiste em identificar grupos homogéneos (*clusters*) de um conjunto de dados heterogéneos. Tendo uma base de dados dispostos numa matriz $n \times p$ (em que as n linhas correspondem às informações sobre os n objetos relativamente às p variáveis observadas (colunas)) é escolhida uma função de distância para relacionar os objetos, de tal modo que a variabilidade entre elementos do mesmo *cluster* seja mínima e a variabilidade entre *clusters* seja máxima. Para a análise de dados multivariados, esta técnica da estatística descritiva é utilizada com uma ferramenta exploratória e de redução da dimensionalidade que revela estruturas de classificação num conjunto de dados, sem informação prévia de possíveis agrupamentos. Na análise de *clusters,* o agrupamento de objetos/variáveis é feito a partir de medidas de semelhanças ou de dissemelhanças entre, inicialmente, dois objetos e, mais tarde, entre dois *clusters* (Maroco J. , 2003).

3.2.1. Medidas de Semelhança e de Dissemelhança

Uma medida de semelhança (ou de dissemelhança) é uma função, s (ou d), que a cada par de objetos faz corresponder um valor real (usualmente, tomam valores no intervalo $[0,1]$). A relação entre a medida de dissemelhança e a medida de semelhança é dada por $s = constante - d$, sendo a constante geralmente o valor máximo que d pode tomar. A partir da matriz de dados ($n \times p$) é possível construir a matriz de semelhanças (ou de dissemelhanças) ($n \times n$). A proximidade de i e j é tanto maior quanto menor é a dissemelhança entre eles.

Uma medida de dissemelhança d_{ij} entre um objeto i e um objeto j deverá satisfazer algumas propriedades:

- $d_{ij} \geq 0, \forall\, i,j = 1, \dots, n$;
- $d_{ii} = 0, \forall\, i,j = 1, \dots, n$; (Identidade)
- $d_{ij} = d_{ji}, \forall\, i,j = 1, \dots, n$; (Simetria)
- $d_{ij} \leq d_{ik} + d_{kj}, \forall\, i,j,k$ (Desigualdade Triangular)

No caso de as medidas de dissemelhança verificarem, além das três primeiras condições, a desigualdade triangular fala-se em distância.

Uma medida de semelhança s_{ij} entre um objeto i e um objeto j deverá satisfazer algumas propriedades:

- $0 \leq s_{ij} \leq 1, \forall\, i, j$
- $s_{ij} = s_{ji}\ \forall\, i, j = 1, \ldots, n$ (Simetria)
- $s_{ii} = 1\ \forall\, i = 1, \ldots, n$ (Identidade)

São várias as medidas que podem ser utilizadas como medidas de distância ou dissemelhança entre cada par de objetos, sendo que estas dependem da natureza das características que são observadas nos objetos. Assim para dois objetos i e j, para as variáveis quantitativas $v = 1, 2, \ldots, p$ podem ser utilizadas as medidas seguintes:

- **Distância Euclidiana**

$$d_{ij} = \left[\sum_{v=1}^{p} (X_{iv} - X_{jv})^2\right]^{1/2} \tag{3.35}$$

Apesar da distância euclidiana ser a mais utilizada, esta apresenta algumas desvantagens, como o seu comportamento para variáveis com variâncias muito distintas, muito correlacionadas, medidas em escalas diferentes e quando existem dados omissos. Para suplantar os defeitos da distância euclidiana, é apresentada a distância de *Manhattan* que deriva da distância Euclidiana.

- **Distância absoluta ou de *Manhattan* ou *City Block***

$$d_{ij} = \sum_{v=1}^{p} |X_{iv} - X_{jv}| \tag{3.36}$$

A distância de *Manhattan* mede a distância numa configuração retilínea. É menos afetada por *outliers* do que a distância euclidiana e mais fácil de interpretar.

3.2.2. Métodos de Classificação

Dentro dos métodos de análise de *clusters*, os mais utilizados são os Hierárquicos e os não Hierárquicos.

Nos **métodos hierárquicos** os *clusters* formam uma hierarquia em que, dados dois grupos, quaisquer que sejam, ou são disjuntos ou um deles está contido no outro. Dentro destes, podem ser do tipo aglomerativo (utilizados normalmente em amostras pequenas, n < 250), no caso de a construção da hierarquia se basear na fusão sucessiva de grupos, ou do tipo divisivo, no caso de a hierarquia se basear na divisão sucessiva de grupos. A representação gráfica das hierarquias formadas é, usualmente, feita através de um dendrograma, que permite visualizar as fusões ou divisões feitas em cada nível da análise de *clusters*. Este tipo de métodos tem o defeito de nunca poderem reparar o que foi feito em conjuntos prévios de agrupamento (Leonard & Rousseeuw, 1990). Para a construção da hierarquia é definida uma regra de ligação, com base nas semelhanças/dissemelhanças definidas anteriormente, as quais permitem selecionar a(s) classe(s) que se vão fundir ou dividir em cada passo.

- **Método da Ligação Simples ou do Vizinho mais próximo**

Sendo C_i e C_j duas classes, a dissemelhança entre elas será dada pelo valor da menor dissemelhança entre um elemento de C_i e um elemento de C_j , isto é:

$$D_{C_iC_j} = min\{d_{ij}: i \in C_i \wedge j \in C_j\} \quad (3.37)$$

(No caso de os dois elementos estarem à mesma distância é escolhido um deles de modo arbitrário)
Este método é mais eficaz para o tipo aglomerativo do que divisivo, pois, no caso das sucessivas divisões, o algoritmo é computacionalmente ineficiente.

- **Método da Ligação Completa ou do Vizinho mais afastado**

A dissemelhança entre duas classes C_i e C_j é definida como sendo a maior dissemelhança entre um elemento de C_i e um elemento de C_j :

$$D_{C_iC_j} = max\{d_{ij}: i \in C_i \wedge j \in C_j\} \quad (3.38)$$

- **Método da Ligação Média**

A dissemelhança entre duas classes C_i e C_j é dada pela média das dissemelhanças entre os elementos de todos os pares que se podem formar com um elemento de C_i e outro de C_j :

$$D_{C_iC_j} = \frac{\sum_{i=1}^{n_i}\sum_{j=1}^{n_j} d_{ij}}{n_i n_j} \quad (3.39)$$

n_i – número de objetos da Classe i
n_j – número de objetos da Classe j

- **Método do Centroide**

Neste método, a distância entre duas classes é dada pela distância entre os centros das classes (centroides). Os centroides das classes C_i e C_j são dados por:

$$\bar{x}_i = \frac{\sum_{i \in C_i} x_i}{n_i} \quad \bar{x}_j = \frac{\sum_{j \in C_j} x_j}{n_j} \quad (3.40)$$

x_i – é o vetor das n observações do objeto i
x_j – é o vetor das n observações do objeto j

Assim, a distância entre duas classes C_i e C_j é dada por:

$$D_{C_iC_j} = d(\bar{x}_i, \bar{x}_j) \quad (3.41)$$

- **Método da Ligação Mediana**

Neste método a distância entre duas classes, C_i e C_j , é dada pela mediana das distâncias entre todos os pares de elementos pertencentes a esses grupos, ou seja:

$$\bar{x} = \frac{\bar{x}_i + \bar{x}_j}{2} \quad (3.42)$$

- **Método de *Ward***

Neste método é utilizada uma medida baseada na média das dissemelhanças entre classes, sendo que essa é calculada como a soma dos quadrados das distâncias dos elementos das duas classes C_i e C_j, aos respetivos centroides sofrem quando se juntam estas duas classes, dada por:

$$\frac{n_i n_j d_{ij}^2}{n_i + n_j} \tag{3.43}$$

Quando se pretende decidir quais as classes a juntar, em cada ciclo do algoritmo, é calculado este incremento para todos os possíveis pares de classes; são escolhidas aquelas que correspondem a um menor incremento. Este incremento da soma dos quadrados significa perda de informação, por isso, pretende-se escolher o menor.

Os **métodos não hierárquicos** distinguem-se dos métodos hierárquicos por não constituírem hierarquias, permitindo, assim, reagrupar os objetos em *clusters* diferentes daqueles em que foram colocados inicialmente. Contudo, estes métodos necessitam que se defina o número de *clusters* que se pretende obter, que por vezes não é fácil definir a priori, por não se conhecer a estrutura dos dados. Para contornar esta questão, pode fazer-se primeiro um método hierárquico e utilizar o número de *clusters* obtido. Existem vários tipos destes métodos que variam consoante os princípios assumidos; os que se destacam e vão ser apresentados são os métodos de partição.

Os métodos de partição aplicam-se apenas a objetos e operam sobre a matriz de dados inicial, com base nela é construída uma partição, isto é, uma coleção de grupos distintos de objetos cuja reunião constitui o conjunto de objetos inicial. O método das K-*means* é o mais utilizado e mede a proximidade entre grupos através da distância euclidiana entre os centroides dos grupos. Os métodos de partição usam procedimentos que em geral seguem os passos seguintes (Johnson R. a., 2008):

1. Selecionar uma partição inicial dos n objectos em k *clusters*;
2. Calcular os centroides para cada um dos k *clusters*;
3. Agrupar os objetos aos clusters cujos centroides se encontram mais próximos, depois voltar ao passo (2) até não ocorrer variação significativa na distância mínima de cada objeto da base de dados a cada um dos centroides dos k clusters (ou até que o número máximo de iterações ou o critério de convergência, definido pelo analista, seja alcançado).

O método das K-*means* apresenta como vantagens a aplicação num conjunto de dados com grande dimensão e, usualmente, a rápida convergência. Contudo, ele tende a procurar clusters esféricos do mesmo tamanho, por isso é preciso fazer um esforço para minimizar a variância à volta do centroide do *cluster*.

Escolha do número de *clusters*

Existem várias técnicas para se determinar o número adequado de *clusters*. Esta escolha depende do tipo de dados e, por isso, é muitas vezes subjetiva. No caso dos métodos hierárquicos é possível definir o número de *clusters* através da visualização do dendrograma. Neste, procura-se grandes alterações na distância para as sucessivas fusões (barras de junção de classes relativamente compridas) e decide-se o ponto de corte do dendrograma. No caso dos métodos não hierárquicos, é necessário indicar à partida o número de *clusters*, assim pode-se utilizar um método hierárquico e escolher o número

conforme anteriormente apresentado, ou utilizar outros métodos de escolha baseados na variância explicada pela classificação. Para além do conhecimento e do tipo de experiência sobre os dados em questão, alguns dos métodos utilizados para escolher o número de clusters dos métodos não hierárquicos são:

- **Método de *Elbow*** – também conhecido como método do cotovelo, este método testa a variância dos dados em relação ao número de *clusters*. O número ótimo de *clusters* é aquele em que um aumento deste não traduz um aumento de diferenciabilidade, por isso não faz sentido aumentar o número de *clusters*. Para encontrar este número basta observar o gráfico e encontrar o valor de k (eixo x) no qual houve uma queda acentuada da variância (eixo y) e a partir do qual este valor estabiliza a variância. A denominação de método de cotovelo deve-se ao tipo de forma que o gráfico apresenta e o valor de k ótimo ser o cotovelo, se se imaginar que a linha do gráfico é um braço.

- **Método da Silhueta** – este método mede o quanto um ponto é semelhante ao *cluster* onde se encontra em comparação com os outros *clusters*. A silhueta é uma medida que permite quantificar a semelhança de objetos, varia entre -1 (não é semelhante) e 1 (é semelhante); se a maioria dos objetos tiver um valor alto de silhueta significa que a classificação de *clusters* é apropriada.

3.2.3. Avaliação da Classificação

Para avaliar as classificações obtidas por medidas de distância e métodos de ligação diferentes, pode-se construir uma nova matriz de dissemelhanças em que o elemento (i, j) é o valor da dissemelhança entre as classes que continham i e j imediatamente antes da sua fusão. Para validar a classificação deve-se comparar a matriz inicial com esta nova matriz.

Esta comparação pode ser feita para métodos hierárquicos, utilizando:

- **Coeficiente de Correlação Cofenética** – obtém-se calculando o valor do coeficiente de correlação usual entre os valores da matriz original e os valores da nova matriz.

$$r = \frac{\sum_i \sum_j (\delta_{ij} - \bar{\delta})(d_{ij} - \bar{d})}{n \times \sqrt{\sum_i \sum_j \frac{(d_{ij} - \bar{d})^2}{n-1}} \times \sqrt{\sum_i \sum_j \frac{(\delta_{ij} - \bar{\delta})^2}{n-1}}} \tag{3.44}$$

  Quanto mais próximo for o valor r de 1 melhor é a classificação.

Para avaliar as classificações obtidas por métodos de classificação distintos, serão utilizadas medidas internas que avaliam a classificação com base na coesão e separação dos dados armazenados em cada *cluster*.

- **Conectividade** – indica o grau de conectividade dos agrupamentos determinada pelos vizinhos mais próximos.

$$Conectividade = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i_{(j)}}} \quad (3.45)$$

$nn_{i_{(j)}}$ - observação do j-ésimo vizinho mais próximo de i

$x_{i,nn_{i_{(j)}}}$ - é 0 se i e $nn_{i_{(j)}}$ pertencem ao mesmo cluster e $\frac{1}{j}$ caso contrário

N – número de observações

L – parâmetro que determina o número de vizinhos que contribuem para a medida de conectividade.

Esta medida varia entre $[0, \infty[$ e o objetivo é maximizá-la.

- **Índice de *Dunn*** – este critério identifica *clusters* densos e bem separados. É definido como a razão entre a distância mínima entre *clusters* e a distância máxima entre *clusters.*

$$Dunn = \frac{min_{1 \leq i \leq j \leq n} d(i,j)}{max_{1 \leq k \leq m} d\prime(k)} \quad (3.46)$$

d – mede a distância inter-*cluster*

d′ – mede a distância intra-*cluster*

i, j, k – índice de *clusters*

O objetivo é maximizar esta medida pois o critério interno procura *clusters* com alta similaridade intra-*cluster* e baixa similaridade inter-*cluster*.

- **Coeficiente de Silhueta** – avalia a distância média intra-cluster (a) e a distância média mais próxima de outro cluster (b) para cada ponto da amostra.

$$Coeficiente\ Silhueta = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (3.47)$$

a_i – distância média entre i e todos os outros pontos pertencentes ao mesmo *cluster*

b_i – distância média entre i e todos os outros pontos pertencentes ao *cluster* mais próximo

Tal como descrito no número ótimo de *clusters* este valor varia entre -1 e 1 e o objetivo é maximizá-lo.

Em suma, a metodologia aplicada na Análise de *Clusters* consistirá na escolha do melhor método Hierárquico, com base na melhor distância entre a Euclidiana e a Manhattan e com base no melhor método de ligação (Simples, Completa, Média, Centroide, Mediana, *Ward*) através do valor do coeficiente de correlação cofenética. Além desse método, será utilizado o *K-means* com base no número ótimo de *clusters.* A escolha entre o método Hierárquico e o método não Hierárquico será feita com base nas medidas de avaliação internas (conectividade, índice de *Dunn* e o coeficiente de silhueta) e/ou com base na experiência e conhecimento sobre os dados.

3.2.4. Vantagens e Desvantagens da Análise de *Clusters*

A Análise de Clusters como método de agrupamento apresenta como vantagens:

- Os resultados obtidos são de fácil visualização e interpretação, que permitem distinguir bem os agrupamentos feitos;
- O agrupamento feito e que é desconhecido *a priori* permite descrever de forma mais eficiente e eficaz as características peculiares de cada um dos grupos identificados.

As desvantagens desta técnica que têm sido identificadas são:

- A redução da matriz de dados original para uma forma mais compacta leva a que algumas características dos objetos individuais percam variabilidade, pois estão a ser substituídas pelos valores médios de cada característica do conjunto de dados da amostra desagregada;
- Apesar de se ficar a conhecer melhor os dados, os critérios para definir o número ótimo de clusters são bastante subjetivos, o que leva a resultados distintos e à necessidade de haver um bom conhecimento por parte do investigador.

3.3. ÁRVORES DE DECISÃO

As árvores de decisão são um modelo preditivo de ML que pode ser utilizado como modelo de regressão ou de classificação. No caso de se querer prever um valor numérico, a árvore é denominada de **árvore de regressão**; no caso de se querer obter como resposta uma variável categórica, a árvore é denominada de **árvore de classificação**. Um modelo de árvores de decisão é constituído a partir de um conjunto de dados de treino, ou seja, é um método de aprendizagem supervisionada.

A árvore vai dividindo um conjunto de dados em grupos homogéneos, cada uma dessas divisões é representada pelo um nó na árvore. O nó raiz é o início da árvore e é o atributo que melhor divide o conjunto de dados, a ramificação da árvore continua com os nós de decisão que são formados quando é feita uma condição lógica. Após as ramificações que a árvore achar relevante fazer, obtém-se os nós terminais/folha que contêm uma classe (árvore de classificação) ou um valor que é a média das observações daquele ramo (árvore de regressão).

O objetivo deste método de *ML* é encontrar o(s) atributo(s) que geram a melhor divisão de dados homogéneos. Com base numa estrutura hierárquica, vão descrevendo sequências de acontecimentos no tempo e, por isso, os primeiros acontecimentos condicionam os seguintes. Os nós de decisão são constituídos pelas variáveis preditoras e, consoante o resultado da árvore, é possível identificar aquelas que possuem maior relação com a variável resposta.

A indução de árvores de decisão, segundo (Maimon, 2005), é definido como "dado um conjunto de treino T composto por um vetor de atributos $\boldsymbol{x} = x_{1,}x_2 , \dots , x_n$ e uma variável objetivo y com uma distribuição desconhecida, o objetivo é induzir um classificador ótimo com um mínimo de erro generalizado." O algoritmo que está atrás dos algoritmos de indução de árvores de decisão é o *Top-Down Induction of Decision Tree* (*TDIDT*). Através deste, são produzidas regras de decisão que vão dividindo as observações de acordo com os valores dos seus atributos preditivos. O algoritmo *TDIDT* é

um método recursivo que procura, num conjunto de atributos, os melhores que dividem o conjunto de dados em subconjuntos. Em cada nó, procura encontrar uma solução ótima local, acreditando que cada uma dessas escolhas leve à solução ótima global. Para o algoritmo escolher o atributo que melhor divide os dados, é utilizado um critério de divisão que se baseia em medidas como impureza, distância ou dependência. Neste trabalho, o algoritmo de indução de árvores de decisão escolhido foi um dos mais utilizados para obter árvores de regressão: o algoritmo CART.

3.3.1. Algoritmo CART

O termo *CART* (*Classification and Regression Tree*) foi introduzido pela primeira vez em 1984 por Breiman e Freidman no seu trabalho sobre árvores de decisão (Leo Breiman, 1984). A aplicação deste método pode ser dividida em quatro passos que são a construção da árvore, o fim da construção da árvore, a poda da árvore e a seleção da árvore ideal. A construção de uma árvore começa com o seu nó raiz, que inclui todos os valores. Para fazer a primeira divisão do nó é escolhida a melhor variável dentro de todas as possíveis variáveis de divisão e dentro de todos os possíveis valores de cada um destas variáveis com base no critério de divisão: Índice de *Gini*.

- ***Gini*** - Esta medida é baseada na pureza do nó. Este índice mede a heterogeneidade dos dados, por isso, a variável preditora com menor índice *Gini* indica maior ordem na distribuição dos dados em relação à variável resposta, e assim, é a escolhida para o nó raiz.

A medida *Gini* é dada por:

$$Gini(N) = 1 - \sum_{C=1}^{k} p(C|N)^2 \tag{3.48}$$

Também se calcula o ganho da medida de *Gini* através da equação:

$$GanhoEntropia(S) = Gini\left(N_{pai}\right) - \sum_{j=1}^{n} \frac{|N_j|}{|N_{pai}|} Gini(N_j) \tag{3.49}$$

Outra medida que pode ser usada é o erro da classificação, dado por:

$$ErroClassificacao(N) = 1 - max_C(p(C|N)) \tag{3.50}$$

O número divisões possíveis de uma variável nominal N com k níveis corresponde a $2^{k-1} - 1$. O critério da "melhor" divisão de um nó baseia-se na minimização da variância da variável resposta nos segmentos descendentes e é dado pela redução da impuridade em dividir o nó pai em dois nós filhos. Este processo de divisão, que se denomina de particionamento recursivo, é feito até que os nós sejam o mais puros possível. O critério de paragem pode ser a definição de um número mínimo até que o nó pare ou até à saturação, o que é preferível. Após o fim da construção da árvore, deve proceder-se à poda que consiste em encontrar uma árvore mais pequena com a máxima eficiência, em um ou mais conjuntos diferentes, face à utilizada para a construção. Isto porque, ao deixar a árvore crescer livremente, pode acontecer que os dados estejam demasiado ajustados ao conjunto de treino e conter nós de baixa significância estatística. A poda à árvore é feita com base numa medida chamada taxa de erro ajustada dada por:

$$EA(T) = E(T) + \propto ContadorFolhas(T) \tag{3.51}$$

$\propto$ - fator de ajuste incrementado gradualmente para criar novas subárvores

Esta medida incrementa cada taxa de classificação errada, de cada nó para o conjunto de treinamento, pela imposição de uma penalidade baseada no número de folhas da árvore. Começa-se pelos nós folha e calcula-se a soma das taxas de erros dos nós filhos resultantes de uma divisão. Essa taxa será a fração de elementos pertencentes ao nó que não pertencem à sua classe. Se a taxa do nó do pai for igual ou inferior à soma dos filhos, a árvores é podada, eliminando os nós filhos. Quando a taxa de erro ajustada para alguma subárvore for menor do que a taxa da árvore completa, então tem-se a primeira subárvore candidata e os ramos que não fazem parte desta são eliminados. O processo reinicia-se com esta subárvore até se obter uma subárvore que apresenta a menor taxa de erro, e essa será a árvore que melhor classifica novos dados por meio de um conjunto de validação. Existem também técnicas de pré-poda (aplicar a poda durante a indução da árvore); no entanto, apesar de serem mais rápidas, são menos eficazes que as de pós-poda descritas anteriormente. Isto porque corre-se o risco de interromper o crescimento da árvore ao selecionar uma árvore subótima. Quando o número de dados é suficiente, utiliza-se o conjunto de teste para testar o modelo que está a ser utilizado e avalia-se as classificações obtidas.

3.3.2. Vantagens e Desvantagens das Árvores de Decisão

De uma forma geral, as árvores de decisão apresentam as seguintes vantagens em relação a outros métodos de classificação (Lewis, 2000):

- Apresentam uma fácil e simples interpretação dos resultados; o resultado observado é facilmente explicado e comprovado por operações lógicas, enquanto isto não se verifica nos modelos ML de "caixa negra";
- Não requer um tratamento de dados intensivo posterior à implementação para garantir a qualidade dos dados;
- Opera tanto sobre dados numéricos como sobre dados categóricos;
- Apresenta um bom desempenho a analisar grandes volumes de dados.

São apontadas como desvantagens das árvores de decisão:

- A árvore resultante no final pode não ser a ótima, visto que a decisão ótima é tomada para o nodo em questão;
- É preciso ter em atenção o *overfitting* (sobreajustamento), pois dados de treino com pouca qualidade tendem a originar árvores demasiado ajustadas aos dados em questão e poderá estar a classificar incorretamente;
- Nos modelos que utilizam variáveis categóricas com diferentes níveis existe uma tendência para dar mais ênfase ao ganho das variáveis com mais nível de detalhe (Deng, 2011).

4. MODELAÇÃO E ANÁLISE DOS DADOS

4.1. DADOS E UNIVERSO DO ESTUDO

Para este trabalho, vão ser utilizados os dados de uma Seguradora a operar, em Portugal, o ramo Saúde. A amostra é composta por 294.388 pessoas e respeita alguns critérios:

- Cobertura em estudo: Internamento
- Período: 01.09.2017 a 31.08.2019
- Sem períodos de carência
- Exposição ao risco = 1

Sendo o Seguro de Saúde constituído por diversas coberturas, decidiu-se estudar apenas a cobertura de **Internamento,** no caso dos seguros individuais. Esta cobertura é a mais contratada e com custos mais dispendiosos, sendo assim a cobertura que tem mais risco. Além disso, é a que mais justifica a existência do seguro, tendo por base o princípio da mutualização, pois a probabilidade de ocorrência de um sinistro deste tipo é baixa (aproximadamente 3%), mas, quando acontece, apresenta uma grande variabilidade de custos.

A amostra utilizada corresponde a um total de **dois anos**, sendo que foram incluídas as pessoas com seguro de saúde entre os anos 2017 e 2019. Não se incluiu os dados dos anos mais recentes (2020 e 2021) por terem sido anos com a existência da pandemia COVID-19 e, por isso, a utilização do seguro de saúde poderia não refletir o comportamento normal do risco, principalmente nos meses de confinamento. Foi aplicado o critério de inclusão de pessoas **sem períodos de carência** (este período indica que, apesar de já terem o seguro de saúde, não o poderão utilizar nesse período; esta medida serve para proteger as Seguradoras da anti-seleção), porque, se uma pessoa estiver em período de carência, não poderá utilizar o seguro e isso não estará a refletir o seu comportamento de risco. Um outro critério foi a inclusão de pessoas com **exposição ao risco igual a um num ano**, ou seja, em cada ano consideraram-se apenas as pessoas que estiveram o tempo todo em risco. Devido à dimensão da carteira, foi possível remover as pessoas que não tiveram o período de risco total e, mesmo assim, ficar com uma amostra de dimensão adequada. A chave considerada para identificar univocamente cada cliente é dada por Apólice+Cliente+Ano, ou seja, no universo da amostra, se uma pessoa teve o seguro ativo durante os dois anos da amostra, é contabilizada por duas vezes, o risco do primeiro ano e o risco do segundo ano.

Os softwares utilizados foram o *SAS* para construção da amostra e o *R Project* para o desenvolvimento da Análise de *Clusters*, das Árvores de Decisão e dos Modelos Lineares Generalizados.

No que diz respeito às variáveis selecionadas para o estudo, optou-se por dois tipos de informação: **variáveis internas**, que são variáveis da pessoa e do seu seguro que estão disponíveis na base de dados da Seguradora, e **variáveis externas**, que não são disponibilizadas pelas pessoas, mas sim por bases de dados do País, nomeadamente, a Transparência (dados sobre o SNS), o Confidencial Imobiliário e o INE (dados sobre os censos). Na Figura 4.1, estão apresentadas as variáveis internas e externas que foram incluídas neste estudo. O objetivo destas variáveis externas é estudar se variáveis socioeconómicas e

variáveis relacionadas com o SNS têm impacto no risco que a Seguradora assume na cobertura de Internamento. No caso das variáveis socioeconómicas, idealmente pretendia-se estudar, por exemplo, o rendimento e a formação académica da pessoa, mas uma vez que a Seguradora não recolhe esses dados optou-se por utilizar estes dados de georreferenciação.

Variáveis Internas	Variáveis Externas		
Idade	Valor de oferta média/m2	Nº de indivíduos com o 1º ciclo do ensino básico completo	Poder de Compra
Género	Valor de transação média/m2	Nº de indivíduos com o 2º ciclo do ensino básico completo	% utilização global de consultas médias num ano
Parentesco	Valor de renda pedida média/m2	Nº de indivíduos com o 3º ciclo do ensino básico completo	% utilização global de consultas médias em três anos
Estado Civil	Valor de renda contratada média/m2	Nº de indivíduos com o ensino secundário completo	Nº de urgências
Zona	Nº de habitações com área até 50 m2	Nº de indivíduos com o ensino pós-secundário completo	Nº de doentes saídos de Internamento
Antiguidade	Nº de habitações com área entre 50 m2 e 100 m2	Nº de indivíduos com o ensino superior completo	Nº de intervenções cirúrgicas
Garantias do Produto	Nº de habitações com área entre 100 m2 e 200 m2	Nº de indivíduos sem atividade económica	Nº de consultas médicas
Capital	Nº de habitações com área superior a 200 m2	Nº de indivíduos empregados no setor primário	% de primeiras consultas realizadas em tempo adequado
Canal Comercial	Nº de habitações com 1 a 2 divisões	Nº de indivíduos empregados no setor secundário	% de mulheres com mamografia realizada nos últimos dois anos
Forma de Pagamento	Nº de habitações com 3 a 4 divisões	Nº de indivíduos empregados no setor terceário	
Tipo de Pagamento	Nº de indivíduos sem saber ler ou escrever	Nº de indivíduos pensionistas ou reformados	

Figura 4.1 - Variáveis internas e variáveis externas

É importante fazer uma distinção entre as variáveis internas e as variáveis externas. As variáveis internas estão ao nível de cada pessoa, mas as variáveis externas são variáveis de georreferenciação, ou seja, o valor que está associado à pessoa reflete as características da sua zona de residência, e estas podem não coincidir com as suas características. Assim, existe a possibilidade de haver algumas discrepâncias entre o comportamento da pessoa e o comportamento da sua zona, mas, neste momento, é a única maneira de estudar este tipo de informação, visto que no momento da subscrição de um seguro de saúde não é pedido este tipo de informações. Além desta restrição das variáveis externas, existem informações que podem estar desatualizadas no que respeita à pessoa, isto porque a informação proveniente do INE é relativamente aos censos de 2011. Apesar de já terem sido realizados censos em 2021, não foi possível utilizar já esses dados e, por isso, está a ser utilizada alguma informação de 2011 para anos da amostra relativos a 2017, 2018 e 2019.

Em termos de metodologia, optou-se por estudar primeiro quais as variáveis internas diferenciadoras do risco e, em segundo lugar, fez-se um novo modelo onde, para além das variáveis internas se introduziram mais as variáveis externas. Isto deveu-se ao facto de não se ter tido logo as variáveis externas disponíveis e, assim, para ganhar experiência com a metodologia construiu-se um modelo apenas com as variáveis internas. Além disso, a construção dos dois modelos permitiu fazer comparações como, por exemplo, avaliar se com a introdução de variáveis externas existem variáveis internas que deixam de ser significativas.

Por motivos de confidencialidade, alguns valores apresentados foram transformados e, por isso, não são os valores reais da amostra. Além disso, em algumas variáveis internas procedeu-se à renomeação das classes por letras, pelo mesmo motivo.

4.2. Tratamento dos Dados e Análise Descritiva das Variáveis

4.2.1. Variáveis Internas

No que diz respeito às variáveis internas, decidiu-se tratá-las como variáveis categóricas, por isso, a Idade foi transformada em escalões etários e a Antiguidade em escalões de antiguidade. Nas variáveis internas não foram encontrados dados omissos. De seguida, é feita uma breve análise descritiva das variáveis internas. No ANEXO 2 são apresentados uns gráficos com o comportamento de risco de cada uma das variáveis internas; as barras a verde indicam a quantidade de pessoas, a vermelho a média e

o desvio-padrão da taxa de utilização do Internamento e a azul a média e o desvio-padrão do custo médio do Internamento, por cada classe da variável.

Idade

A Idade é uma variável que está identificada como diferenciadora do risco em saúde e é utilizada atualmente nos modelos de tarifação. Neste estudo, pretende-se confirmar a sua significância e perceber de que maneira o risco se distribui ao longo das idades. Assim, foram considerados escalões etários de cinco em cinco anos. O escalão dos 41 anos aos 45 anos é o que tem mais pessoas (12,7% da amostra) e nas idades avançadas, para idades superiores aos 80 anos, temos poucas pessoas na amostra (694). Em termos de risco, percebe-se que existe uma tendência crescente tanto na utilização como no custo à medida que a idade aumenta, a partir dos 11 anos. Nas idades mais avançadas, não se pode tirar grandes conclusões devido à dimensão da amostra e, por isso, existe alguma variabilidade na utilização e no custo do Internamento.

Género

Na amostra considerada, a proporção homens-mulheres é dada por 44%-56%, é uma carteira equilibrada em termos de género, sendo que existem mais pessoas do género feminino. As pessoas que tinham o género indefinido foram excluídas da amostra. Em termos de comportamento do risco, não há diferenças na taxa de utilização do Internamento para homens e mulheres, mas, no caso do custo, os homens têm, em média, um custo superior ao das mulheres.

Estado Civil

O Estado Civil das pessoas pode ser solteiro, casado, separado ou viúvo. Na amostra, 61% são pessoas solteiras. Em termos de risco, este aparenta ser menor nos solteiros e maior nos viúvos.

Parentesco

O Parentesco indica a relação entre a pessoa segura e o tomador da apólice, e esta pode ser uma das seguintes categorias: Ascendente, Cônjuge, Titular, Pai ou Mãe, Filho (a), Neto (a), Irmão (ã), Outras relações. As primeiras classes aparentam ter um risco superior às últimas classes.

Antiguidade

A variável que indica há quanto tempo a pessoa tem o seguro de saúde na Companhia varia entre os 0 anos (as pessoas adquiram o seguro no primeiro ano da amostra) e os 30 anos. Contudo, a partir dos 18 anos, como existem poucas pessoas, iremos criar uma classe que abrange as pessoas com seguro há mais de 18 anos. Para as antiguidades iniciais, decidiu-se não as agrupar, pois nos primeiros anos pode haver uma utilização diferente do seguro de saúde: é um período de adaptação da pessoa ao seguro e poderá levar a um comportamento também diferente. Em termos de taxa de utilização e de custo, a tendência é ser crescente com a antiguidade.

Zona

Na variável Zona destaca-se a classe Grande Lisboa, sendo que 31% das pessoas da amostra vivem em Lisboa. Esta divisão das zonas do país está feita com base nas regiões de Portugal e o Estrangeiro, sendo que existem algumas exceções, pois têm também a ver com a distribuição da rede de prestadores da Seguradora. O Estrangeiro é a classe com menos pessoas (265) e terá de se juntar a outra classe com comportamentos de risco semelhantes para se poder tirar conclusões. Em termos do

risco, Lisboa é a zona com taxa de utilização mais alta, mas não é a que tem custo mais alto, sendo os Açores a zona que tem o custo mais alto. Contudo, é preciso ter particular atenção às classes que têm poucas pessoas, avaliando se não é necessário agrupá-las a outras classes que lhes sejam homogéneas em termos de risco.

Garantias do Produto

Esta variável contém as coberturas que o seguro tem, para além da cobertura do Internamento. Distinguiu-se o produto I + A (Internamento e Ambulatório) do I + Acons (Internamento e Ambulatório de consultas) porque, no primeiro, a pessoa tem um capital disponível para a cobertura de Ambulatório, enquanto, no segundo, apenas tem disponível um número limitado de consultas e, por isso, o risco envolvido entre os dois tipos de produto é diferente. A existência de mais coberturas pressupõe que estas tenham um capital disponível; assim, se apenas tiverem o acesso à rede de prestadores (o custo fica apenas a cargo do cliente, não há financiamento da Seguradora) nessa cobertura, não conta como a existência dessa cobertura. O produto I + A é o mais comum representado 31% da amostra. Em termos de risco, a taxa de utilização é superior para produtos mais completos (com mais coberturas) e o custo também aparenta ser assim, exceto na distinção entre os produtos I e I + Acons. O produto I + A + PO, apesar de ter três coberturas, não aparenta ter mais risco que o produto I + A, mas tem menos pessoas e pode ser da aleatoriedade da amostra.

Capital

Os capitais disponíveis para a cobertura de Internamento nesta amostra são 11 entre os 10.000 euros e os 500.000 euros. Contudo, existem uns que correspondem aos produtos mais comuns e, por isso, têm mais pessoas, como é o caso dos 25.000 euros, dos 50.000 euros e dos 100.000 euros. Os restantes capitais serão agrupados aos capitais adjacentes, que têm mais pessoas, para os resultados terem uma maior robustez estatística. Assim, serão analisados os capitais em quatro classes distintas: {10.000, 25.000}, {30.000, 35.000, 45.000, 50.000, 65.000, 75.000}, 100.000 e {250.000, 500.000}.

Canal Comercial

O Canal Comercial que indica onde feita a compra do seguro de saúde está dividido em oito classes distintas. Estas classes, por motivo de confidencialidade da Seguradora, estão classificadas por letras sem nenhuma ordem em específico e referem-se a canais de venda direta ou mediados. O Canal A é o mais comum, o Canal B e o Canal H têm apenas 892 e 140 pessoas, respetivamente, por isso serão agrupados a outros canais com comportamentos de risco idênticos. Em termos de taxa de utilização, o canal que apresenta maior valor é os C, e menor valor é o A. Quanto ao custo da utilização do Internamento, a F é a que, em média, tem maior custo, e o menor custo é dado no B e no H, mas cujo comportamento terá de ser mais bem analisado devido à dimensão da amostra.

Forma de Pagamento

A Forma de Pagamento é o tipo de fracionamento do prémio, por motivos de confidencialidade foram atribuídas letras às classes. A classe A é a mais comum (83% da amostra). Em termos de risco, o fracionamento A é o que apresenta menor taxa de utilização e menor custo, mas os outros tipos de fracionamento têm uma amostra reduzida, por isso, terão de ser melhor analisados.

Tipo de Pagamento

Quanto ao Tipo de Pagamento do prémio, este indica se o pagamento do prémio é de débito direto ou não, também foram atribuídas letras para codificar a variável. O mais comum na amostra (95%) é o A, este tipo de pagamento tem menor taxa de utilização e menor custo, mas a classe B contém pouca informação.

4.2.2. Variáveis Externas

Relativamente às variáveis externas, foi necessário fazer mais tratamentos nos dados, isto porque estes provêm de bases de dados distintas. O tratamento da informação das variáveis externas trouxe alguns desafios nomeadamente a verificação de moradas, a transformação de variáveis, os valores omissos e a análise de correlação. A informação das variáveis externas pode ser dividida em cinco categorias: Casa, Poder de Compra, Escolaridade, Atividade Económica e Saúde. De seguida, são apresentadas as etapas percorridas até se ter as variáveis externas prontas a utilizar.

Verificação de moradas

Para fazer a ligação da amostra com as informações das bases de dados externas, houve a limitação de se ter de utilizar as coordenadas geográficas atuais das pessoas da amostra, enquanto que a informação das pessoas da amostra corresponde aos anos de 2017 a 2019. Assim, foi necessário verificar se as coordenadas das pessoas atuais correspondiam às coordenadas dos anos da amostra, ou seja, se uma pessoa tiver mudado o local de residência entre 2017 e 2022, não poderá estar na amostra, isto porque as informações das variáveis externas dependem todas do local de residência, são variáveis de georreferenciação. Assim, das 294.388 pessoas da amostra inicial, 4,5% (13.263 pessoas) tinha uma morada atual diferente da amostra, por isso, a amostra passou a ter 281.125 pessoas.

Transformação das variáveis

Alguma da informação recolhida não estava pronta para ser logo incluída no modelo, tendo sido necessário fazer algumas transformações, por exemplo: tendo o número de indivíduos com uma característica e o total de indivíduos, interessa avaliar a proporção de indivíduos com aquela característica, numa determinada zona. Para cada uma das categorias expostas anteriormente são apresentadas as transformações feitas, no caso de ter sido necessário, o nível geográfico e o ano a que se referem.

<u>Casa</u>

- Valor Oferta Média/m2 (Freguesia, 2017 e 2018)
- Valor Transação Média/m2 (Freguesia, 2017 e 2018)
- Valor Renda Pedida Média/m2 (Freguesia, 2017 e 2018)
- Valor Renda Contratada Média/m2 (Freguesia, 2017 e 2018)
- Proporção de habitações com área até i m2 (Subsecção Estatística, 2011):

$$\frac{Nº\ de\ habitações\ com\ i\ m2}{Nº\ total\ de\ habitações}, i = até\ 50, 50\ a\ 100, 100\ a\ 200, > 200$$

- Proporção de habitações com i divisões (Subsecção Estatística, 2011):

$$\frac{Nº\ de\ habitações\ com\ i\ divisões}{Nº\ total\ de\ habitações}, i = 1\ a\ 2, 3\ a\ 4$$

Poder de Compra (Concelho, 2011)

- Indicador per Capita do Poder de Compra, tendo por referência o valor nacional (Portugal=100)

Escolaridade (Subsecção Estatística, 2011)

- Proporção de Analfabetos:

$$\frac{Nº\ de\ indivíduos\ sem\ saber\ ler\ ou\ escrever}{Nº\ total\ de\ indivíduos}$$

- Proporção de pessoas com ensino i completo:

$$\frac{Nº\ de\ indivíduos\ com\ o\ ensino\ i\ completo}{Nº\ total\ de\ indivíduos},$$
$$i = Básico, Secundário, Pós - Secundário, Superior$$

Nota: O Ensino Pós-Secundário é o Ensino Profissional.

Atividade Económica (Subsecção Estatística, 2011):

- Proporção de pessoas empregadas no setor i :

$$\frac{Nº\ de\ indivíduos\ empregados\ no\ setor\ i}{Nº\ total\ de\ indivíduos\ empregados}, i = primário, secundário, terciário$$

- Proporção de pessoas sem atividade económica:

$$\frac{Nº\ de\ indivíduos\ sem\ atividade\ económica}{Nº\ total\ de\ indivíduos\ com\ 15\ anos\ ou\ mais}$$

- Proporção de pessoas pensionistas ou reformados:

$$\frac{Nº\ de\ indivíduos\ pensionistas\ ou\ reformados}{Nº\ total\ de\ indivíduos\ com\ 15\ anos\ ou\ mais}$$

Saúde (Concelho, 2017 e 2018):

No que diz respeito aos dados de saúde, foi necessário extrair ainda o número de utentes inscritos em cuidados de saúde primários por cada ACES (Agrupamentos de Centros de Saúde). Esta informação foi obtida para Portugal Continental, não tendo sido possível encontrar os valores para os Açores, Madeira e Estrangeiro. Assim, para os Açores e Madeira considerou-se a zona por inteira e não dividida em ACES, e no caso do Estrangeiro não foram atribuídos valores.

- Taxa de utilização de consultas em 1 ano e em 3 anos
- Taxa de urgências:

$$\frac{Nº\ de\ urgências}{Nº\ \ total\ de\ utentes\ inscritos}$$

- Taxa de doentes saídos de Internamento:

$$\frac{Nº\ de\ doentes\ saídos\ de\ Internamento}{Nº\ \ total\ de\ utentes\ inscritos}$$

- Taxa de Intervenções Cirúrgicas:

$$\frac{Nº\ de\ Intervenções\ Cirúrgicas}{Nº\ \ total\ de\ utentes\ inscritos}$$

- Taxa de Consultas médicas:

$$\frac{Nº\ de\ Consultas\ Médicas}{Nº\ \ total\ de\ utentes\ inscritos}$$

- Taxa de primeiras consultas realizadas em tempo adequado
- Taxa de mulheres com mamografia realizada nos últimos 2 anos

Valores omissos

Devido à recolha das variáveis externas de bases de dados distintas e por valores de georreferenciação, existem alguns casos em que as pessoas da amostra não têm valor para a variável em questão; assim, decidiu-se retirar essas pessoas da amostra e verificar se, mesmo assim, se teria uma amostra com dimensão suficiente para construir o modelo. Das 281.125 pessoas presentes na amostra, já com a verificação das moradas, 56% tinham todos os valores das variáveis externas preenchidos e 124.048 pessoas foram retiradas da amostra por terem pelo menos um valor omisso numa variável externa. Assim, para utilizar as variáveis externas, será necessário considerar a amostra com apenas as pessoas com os valores todos preenchidos, o que corresponde a uma amostra de 157.077 pessoas. Como esta amostra ainda tem uma dimensão suficiente, decidiu-se apenas eliminar os valores omissos, em vez de os substituir por um valor, e também porque o valor original da variável já é uma aproximação à pessoa; se se fosse atribuir mais uma aproximação, a amostra poderia ficar ainda mais longe da realidade.

Análise de Correlação

As variáveis externas são todas numéricas, por isso, é possível obter a matriz de correlações. Sendo que estas variáveis são de georreferenciação, ou seja, são aproximações às pessoas da amostra, decidiu-se que para valores superiores a 0,75 se iria considerar que as variáveis eram correlacionadas. Este critério é menos exigente que o utilizado para as variáveis internas, devido ao tipo de informação que elas representam e ao que foi exposto anteriormente.

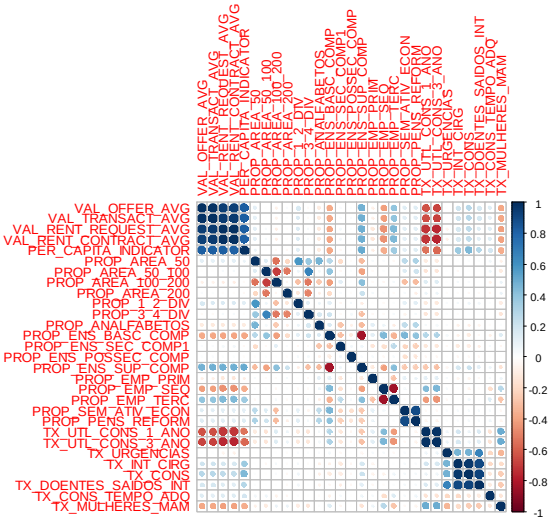


Figura 4.2 - Matriz de correlações das variáveis externas

Da matriz de correlações, verifica-se que as variáveis da Casa relativamente aos valores de oferta e de renda estão correlacionadas entre si (entre 0,94 e 0,99) e com o Poder de Compra (0,83). No que diz

respeito à Escolaridade, a Proporção de indivíduos com o ensino básico completo está correlacionada com a Proporção de indivíduos com o ensino superior completo (-0,84). Relativamente à Atividade Económica, a Proporção de indivíduos empregados no setor secundário e empregados no setor terciário estão correlacionados (-0,84) e a Proporção de indivíduos sem atividade económica está correlacionada com a Proporção de indivíduos pensionistas ou reformados (0,90). Nas variáveis da categoria Saúde, a Taxa de utilização de consultas em 3 anos está correlacionada com as variáveis, da categoria Casa, Valor médio de renda pedida (-0,76) e Valor médio da renda contratada por m2 (-0,76). Entre as variáveis de Saúde existem as seguintes correlações: Taxa de utilização de consultas em 1 ano e em 3 anos (0,99), Taxa de intervenções cirúrgicas e Taxa de consultas (0,97), Taxa de intervenções cirúrgicas e Taxa de doentes saídos de internamento (0,91) e Taxa de consultas e Taxa de doentes saídos de internamento (0,92). Como os modelos lineares generalizados não permitem que existam variáveis explicativas correlacionadas, vai-se ter que escolher das variáveis correlacionadas aquela que melhor explica o risco, isto será feito através do método de seleção de variáveis *Stepwise Forward*.

4.3. Modelação do Risco

O modelo de risco é dado pelo valor esperado de R, sabendo o impacto do vetor das variáveis explicativas $\boldsymbol{V}$. Este modelo é dado pela multiplicação do valor esperado da utilização do Internamento, N, dado o vetor de variáveis explicativas $\boldsymbol{V}$, pelo valor esperado do custo da utilização do Internamento, X, dado o vetor de variáveis explicativas V. O modelo de risco é então dado por:

$$\boldsymbol{E(R|V) = E(N|V) * E(X|V)} \tag{4.1}$$

$\boldsymbol{N}$ – variável aleatória que indica se a pessoa utilizou a cobertura de internamento num ano.
$\boldsymbol{X}$ – variável aleatória que indica o custo associado à utilização da cobertura de internamento de cada pessoa num ano.
$\boldsymbol{V}$ – vetor com o valor de cada variável explicativa que caracteriza a pessoa num ano.

Utilização

É usual utilizar como variável resposta a frequência de sinistralidade, contudo na cobertura de Internamento, em geral, existe uma taxa de utilização muito baixa. Isto porque um Internamento não é um ato médico que uma pessoa faça com regularidade e, quando ocorre mais do que uma vez ao ano, os diagnósticos médicos costumam estar relacionados, ou seja, não são independentes. Na amostra utilizada, a taxa de utilização média é de 4,3%. Tal como descrito na revisão literária, estamos perante um caso em que temos um ponto de massa para $N = 0$; assim não iremos estudar o número de vezes que uma pessoa foi internada, mas sim se uma pessoa foi ($N = 1$) ou não ($N = 0$) internada num ano. Faz então sentido modelar a utilização da cobertura de Internamento, e, para isso, será utilizado um modelo linear generalizado logístico.

Custo médio por Utilização

Pretende-se modelar o custo médio da utilização do Internamento, por isso o universo em estudo passa a ser apenas os que utilizaram a cobertura.

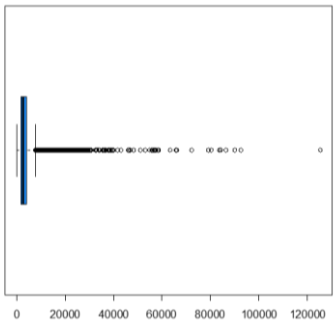


Figura 4.3 - *Boxplot* dos custos da utilização do Internamento

Tal como se verifica na Figura 4.3, a distribuição dos custos apresenta uma cauda bastante pesada. O custo médio do Internamento da amostra é 4.107 euros e os custos apresentam alguma variabilidade, sendo o desvio padrão de 5.535 euros.

Devido a estas características, não foi possível ajustar apenas uma distribuição ao custo e foi necessário dividi-lo entre três partes: os custos até 10.500 euros, os custos entre 10.500 e 27.550 euros e os custos superiores a 27.550 euros. Estes pontos de corte foram obtidos tendo como referência o quantil 95% de distribuição, e afinados até ser possível ajustar uma distribuição da Família Exponencial.

Para os custos do tipo I (até 10.500 euros), começou-se por tentar ajustar uma distribuição para os custos até 11.349 euros (quantil 95%); no entanto, os testes às distribuições rejeitavam todos a hipótese de esses custos seguirem uma distribuição da Família Exponencial. Por isso, testou-se com um limite superior mais baixo, e assim foi possível ajustar uma distribuição Gama (forma = 2,399126, escala = 1.175,868035) para os custos até 10.500 euros (quantil 94,24%).

Para os custos restantes, foi necessário definir novamente um limite superior tendo-se ajustado para os custos entre 10.500 euros e 27.550 euros uma distribuição Pareto Generalizada (forma = 0,18, escala = 2.303, localização = 10.500) e para os custos restantes uma distribuição Pareto Generalizada (forma = 0,24, escala = 9.198, localização = 27.550). Sendo que a maioria dos utilizadores (94%) teve custos até 10.500 euros, apenas 552 pessoas tiveram custos entre os 10.500 euros e 27.550 euros e 93 pessoas tiveram custos superiores a 27.550 euros. Apenas será possível construir um modelo linear generalizado (Modelo Gama) para os custos do tipo I e, para os restantes, por não se ter informação suficiente, vai-se considerar o valor médio das distribuições.

Para se obter o modelo do custo médio da utilização do Internamento é necessário considerar ainda a probabilidade de uma pessoa ter um dos tipos de custo; isto será feito através de um modelo linear generalizado multinominal logístico em que a variável resposta é categórica e toma um dos seguintes valores: a pessoa teve um custo até 10.500 euros (custo tipo I), a pessoa teve um custo entre os 10.500 e os 27.550 euros (custo tipo II), ou a pessoa teve um custo superior a 27.550 (custo tipo III).

Assim, temos que a modelação dos custos é dada pela divisão:

$$E(X|V) = E((X|x \leq 10.500)|V) \times P(X \leq 10.500|V) + E\big((X|x > 10.500\ \Lambda\ x \leq 27.550)\big|V\big) \times P\big((X|x > 10.500\ \Lambda\ x \leq 27.550)|V\big) + E((X|x > 27.550)|V) \times P(X > 27.550) \quad (4.2)$$

Logo o valor de risco é dado por:

$$E(Y|V) = E(N|V) \times \big[E((X|x \leq 10.500)|V) \times P(X \leq 10.500|V) + E\big((X|x > 10.500 \,\Lambda\, x \leq 27.550)\big|V\big) \times P\big((X|x > 10.500 \,\Lambda\, x \leq 27.550)|V\big) + E((X|x > 27.550)|V) \times P(x > 27.550)\big] \quad (4.3)$$

Para cada modelo, serão apresentados os resultados com as variáveis internas e com as variáveis internas mais as variáveis externas. Relativamente às variáveis candidatas a diferenciadoras do risco, em cada modelo, vão ser seguidos as etapas apresentadas na metodologia, nomeadamente:

1. Análise Estatística: Para incorporar as variáveis nos modelos é preciso garantir que existe uma amostra com dimensão suficiente para retirar conclusões estatísticas. As variáveis internas são todas categóricas; por isso, é preciso garantir que cada categoria tem expressão suficiente para fazer inferências. No caso das variáveis externas será ainda necessário escolher, de entre as variáveis correlacionadas, aquela que melhor explica a variável resposta.

2. Significância Estatística: Para ver se as variáveis candidatas a diferenciadoras do risco se revelam diferenciadoras da variável resposta, vai-se recorrer ao teste de razão de verosimilhanças e aos métodos de seleção *Forward*, *Backward* e *Stepwise Forward*. Será ainda feita uma análise de correlação para garantir que no modelo com as variáveis significativas não existem variáveis correlacionadas, visto que os MLGs têm essa restrição.

3. Identificação da Homogeneidade: No caso de existirem categorias de uma variável que têm comportamentos semelhantes face à variável resposta elas devem ser agrupadas. Para as variáveis internas não ordinais será feita uma análise de *clusters* e árvores de decisão e para as variáveis internas ordinais, será feito o teste de *Wald* e o teste de *Tukey,* para fazer comparações múltiplas e garantir que se obtém o modelo mais simplificado.

4.3.1. Modelação da Utilização

Modelo da Utilização com as variáveis internas

Análise Inicial

Como as variáveis internas são qualitativas, decidiu-se que cada classe teria de ter pelo menos 1.000 pessoas, que corresponde a 0,3% da amostra total, para ser estatisticamente representativa. Com este pressuposto, no caso das variáveis ordinais, as classes que não tinham pessoas suficientes foram agrupadas às adjacentes; para as variáveis não ordinais, agruparam-se à classe cujo comportamento da variável resposta em questão era o mais semelhante. Estas alterações verificam-se na variável Idade cujas idades superiores a 75 anos ficaram numa classe única até aos 95 anos; na Zona o Estrangeiro foi agrupado aos Açores por terem taxas de utilização semelhantes; no Canal Comercial, a classe Canal B foi agrupado ao Canal G; o Canal H foi agrupado ao Canal A.

Significância Estatística

Tendo as variáveis tratadas e analisadas, a etapa que se segue é ver quais são significativas para a utilização do Internamento. No caso do teste de razão de verosimilhanças, começou-se pela variável cujo valor-p é maior e cujas classes são todas não significativas. Os resultados podem ser observados na Tabela 4.1.

Variável	Valor-p
Tipo de Pagamento	0,483
Género	0,166
Forma de Pagamento	0,009
Parentesco	0,003
Estado Civil	0,001

Variável	Valor-p
Capital	1,11E-15
Canal Comercial	<2,2e-16
Zona	<2,2e-16
Garantias do Produto	<2,2e-16
Antiguidade	<2,2e-16
Idade	<2,2e-16

Tabela 4.1 - Significância das variáveis internas no modelo da utilização

As variáveis foram testadas pela ordem com que são apresentadas e, pelos valores-p, podemos ver que o Tipo de Pagamento e o Género não são significativas para a Utilização do Internamento, enquanto as restantes variáveis são significativas para um nível de significância de 5%. Pelos métodos de seleção *Backward*, *Forward* e *Stepwise Forward* os resultados foram semelhantes; contudo, neste caso o Género foi incluído na seleção, apesar de ter apenas contribuído para baixar o AIC em décimas. Apesar de, nos métodos de seleção de variáveis, o Género ter sido incluído para explicar a Utilização, como no teste de razão de verosimilhanças rejeitou-se a hipótese de esta variável ser significativa para a utilização com um valor-p de 0,166; então esta variável não vai ser incluída no modelo.

Das dez variáveis significativas para a utilização, fez-se uma análise de correlação para garantir que não havia variáveis correlacionadas no modelo. Os resultados encontram-se no ANEXO 3 e indicam que o Estado Civil e o Parentesco estão correlacionados com a Idade, sendo que, destas variáveis, a Idade é a variável que mais explica a utilização; e a variável Garantias do Produto está correlacionada com o Capital, sendo, neste caso, as Garantias do Produto que mais explicam a utilização. Assim, as variáveis Estado Civil, Parentesco e Capital foram retiradas do modelo e foi feito um novo teste de significância às varáveis restantes que indicou que todas continuavam relevantes, para um nível de significância de 5%.

Identificação da Homogeneidade

Sabendo quais as variáveis significativas para a Utilização, resta otimizar o modelo. Isto significa agrupar as classes cujo comportamento da variável resposta é semelhante, pois os modelos são sensíveis ao número de parâmetros estimados, e como cada classe tem um parâmetro, se se conseguir agrupar a outra classe reduz-se a complexidade do modelo e melhora-se o seu desempenho. No caso das variáveis não ordinais, para se conhecer melhor o comportamento das mesmas, fez-se uma análise de *clusters* e árvores de decisão para ver que classes poderiam ser agrupadas. Para decidir qual o agrupamento a utilizar, no caso de a análise de *clusters* dar resultados diferentes das árvores de decisão, utilizou-se as medidas *AIC* e *Deviance* para escolher o agrupamento a considerar no modelo final.

Zona

Para a Zona, a análise de *clusters* agrupou as 19 classes em quatro *clusters* e as árvores de decisão agruparam as zonas inicialmente em 6 nodos, que podem ser observados no ANEXO 5; contudo, como alguns deles tinham taxas de utilização muito semelhantes decidiu-se podar a árvore de modo a ter cinco nodos. Os resultados de ambas as técnicas são os da Tabela 4.2.

Zonas Agrupadas - *Clusters*	Taxa de utilização média
Beira Litoral Centro, Açores, Estrangeiro	2,0%
Madeira, Beira Litoral Sul, Beira Interior Sul, Beira Interior Norte, Alto Alentejo	2,9%
Trás-os-Montes, Ribatejo, Beira Litoral Norte, Baixo Alentejo, Alto Minho	3,6%
Oeste, Margem Sul, Marão, Grande Porto, Grande Lisboa, Baixo Minho, Algarve	4,6%

Zonas Agrupadas – Árvores de Decisão	Taxa de utilização média
Beira Litoral Centro	1,6%
Açores, Estrangeiro, Madeira, Beira Litoral Sul, Beira Interior Sul, Beira Interior Norte, Alto Alentejo	2,7%
Trás-os-Montes, Ribatejo, Baixo Alentejo, Alto Minho	3,6%
Beira Litoral Norte, Oeste, Margem Sul, Baixo Minho, Algarve	4,4%
Marão, Grande Porto, Grande Lisboa	4,9%

Tabela 4.2 - Agrupamento da taxa de utilização por *clusters* e árvores de decisão para a Zona

Os resultados são consistentes para ambas as técnicas, apenas existem algumas diferenças nomeadamente nos Açores e Estrangeiro que são agrupados nos *clusters* à classe que tem menor taxa de utilização (2,0%) e nas árvores de decisão estão no nodo com taxa de utilização 2,7%; a Beira Litoral Norte na análise de *clusters* está na classe cuja taxa de utilização é 3,6% e nas árvores de decisão é 4,4%. A árvore ao ter mais uma classe que a análise de *clusters* separa as zonas com mais taxa de utilização nomeadamente o Marão, o Grande Porto e a Grande Lisboa da Margem Sul, do Oeste, do Baixo Minho e do Algarve.

Modelo	Deviance	Deviance_Nulo – Deviance Modelo	AIC
Nulo	95142	0	95144
Zona	94537	605	94575
Zona AC	94578	564	94586
Zona AD	94547	595	94563

Tabela 4.3 - *Deviance* e AIC para os modelos da utilização para a Zona

Comparando os dois agrupamentos, obteve-se o valor mínimo de 564 para a distância entre a *Deviance* do modelo nulo e a *Deviance* do modelo com os agrupamentos da análise de *clusters* (AC). O menor AIC foi obtido para o agrupamento pelas árvores de decisão (*AD*). Como não existe um método que seja melhor nas duas medidas, decidiu-se utilizar o agrupamento com mais classes, o agrupamento dado pelas árvores de decisão, porque depois ainda poderá ser otimizado, se for realmente melhor, pelo teste de *Tukey*.

Garantias do Produto

Para as Garantias do Produto, obteve-se três *clusters* e uma árvore com quatro nodos que podem ser observados no ANEXO 5. Tal como na Zona, fez-se uma poda à árvore por haver taxas de utilização semelhantes e, assim, ficou-se com três nodos.

Garantias do Produto Agrupadas - *Clusters*	Taxa de utilização média
I	1,9%
I + Acons, I + A + PO	3,6%
I + A, I + A +E, I + A + E + PO	5,6%

Garantias do Produto Agrupadas - Árvores de Decisão	Taxa de utilização média
I	2,0%
I + Acons	3,6%
I + A, I + A + PO, I + A +E, I + A + E + PO	5,6%

Tabela 4.4 - Agrupamento da taxa de utilização por *clusters* e árvores de decisão para as Garantias do Produto

Enquanto os *clusters* incluíram no mesmo *cluster* os produtos I + A e os produtos I + A + PO, no caso das árvores estes últimos produtos ficaram num nodo diferente. Por fim, este método distinguiu os produtos com mais taxa de utilização que são os que têm capital de Ambulatório.

Modelo	Deviance	Deviance_Nulo – Deviance Modelo	AIC
Nulo	95142	0	95144
Garantias do Produto	93278	1864	93290
Garantias do Produto AC	93317	1826	93323
Garantias do Produto AD	93325	1817	93331

Tabela 4.5 - *Deviance* e AIC para os modelos da utilização para as Garantias do Produto

O agrupamento que se vai considerar para o modelo será o das árvores de decisão, pois apresentou melhor distância entre as *deviances*.

Canal Comercial

No caso do Canal Comercial, obteve-se quatro *clusters* e seis nodos, que, após a poda, ficaram quatro nodos, que podem ser observados no ANEXO 5. Ambos os métodos de agrupamentos deram as mesmas classes e as taxas médias da utilização estão dadas na Tabela 4.6.

Canal Comercial Agrupado - *Clusters* e Árvores de Decisão	Taxa de utilização média
A, H	2,9%
B, G	4,6%
D, E, F	5,4%
C	7,3%

Tabela 4.6 - Agrupamento da taxa de utilização por clusters e árvores de decisão para o Canal Comercial

Forma de Pagamento

Na Forma de Pagamento, tanto a análise de *clusters* como a árvore de decisão deram três classes em vez das quatro classes originais, que consistem no agrupamento da forma de pagamento D e B.

Forma de Pagamento Agrupado - *Clusters* e Árvores de Decisão	Taxa de utilização média
A	4,0%
B, D	5,6%
C	4,6%

Tabela 4.7 - Agrupamento da taxa de utilização por *clusters* e árvores de decisão para a Forma de Pagamento

Após os agrupamentos obtidos para as variáveis não ordinais, introduziram-se estas variáveis no modelo juntamente com as variáveis ordinais e testou-se se as classes já eram todas distintas entre si ou se ainda havia homogeneidade em algumas delas, em relação à variável resposta. Para isso, começou-se por testar as classes face à classe do segurado padrão pelo teste de *Wald* e, posteriormente, utilizou-se o teste de *Tukey* para testar a homogeneidade entre classes que não são do segurado padrão. Começou-se por testar as classes que têm maior valor-p. No caso da Idade, apenas se agruparam as idades entre os 21 anos e os 30 anos. Na Antiguidade, o tipo de utilização da classe dos 0 a 1 anos é idêntica à classe do segurado padrão (2 a 8 anos) e agruparam-se as antiguidades dos 12 aos 17 anos. Nas Garantias do Produto, os agrupamentos já foram todos feitos e as classes são todas distintas entre si. Quanto à Zona, a classe com segunda taxa de utilização mais alta

foi agrupada à classe com a taxa de utilização mais alta (a do segurado padrão), o que já tinha sido sugerido pela análise de *clusters* em ter apenas quatro classes. No caso do Canal Comercial, a taxa de utilização pode ser dada por três classes distintas, sendo que se agrupou a classe dos C à classe composta pela classe B e pela classe G; agruparam-se as classes D, E e F.

Modelo Final

Sabendo as variáveis significativas da utilização e quais as suas classes, garantindo que são heterogéneas, constrói-se o modelo e analisa-se o impacto de cada variável na utilização do Internamento. Para isto, é necessário ter em conta que o modelo tem como referência um certo segurado padrão. As características do segurado padrão são, em geral, as classes que têm maior dimensão, e as que foram utilizadas para o modelo da utilização do Internamento são as da Tabela 4.8.

Idade	Antiguidade	Zona	Garantias do Produto	Canal Comercial	Forma de Pagamento
[36, 45]	[0, 8]	Algarve, Baixo Minho, Beira Litoral Norte, Grande Lisboa, Grande Porto, Marão, Margem Sul, Oeste	I + A, I + A + E, I + A + PO, I + A + E + PO	D, E, F	A

Tabela 4.8 - Características do segurado padrão do modelo da utilização com as variáveis internas

Assim, quando se avalia os resultados que se seguem, é preciso ter sempre em conta que são os valores relativos a uma pessoa com as características do segurado padrão.

Idade

	E(N|V)
Segurado Padrão	6,0%
Idade [0,5]	5,9%
Idade [6,10]	3,3%
Idade [11,15]	2,5%
Idade [16,20]	4,5%
Idade [21,25]	4,5%
Idade [26,30]	4,5%
Idade [31,35]	5,4%
Idade [36,40]	6,0%
Idade [41,45]	6,0%
Idade [46,50]	8,7%

	E(N|V)
Segurado Padrão	6,0%
Idade [51,55]	10,1%
Idade [56,60]	11,1%
Idade [61,65]	15,2%
Idade [66,70]	21,4%
Idade [71,75]	21,4%
Idade [76,80]	29,1%
Idade [81,85]	29,1%
Idade [86,90]	29,1%
Idade [91,95]	29,1%

Tabela 4.9 - Taxa de utilização do modelo com as variáveis internas para a Idade

A idade é a variável que já sabíamos ser significativa para o risco em saúde, mas, dividindo por escalões etários, é possível observar que, face ao segurado padrão (36 aos 45 anos), os mais novos na fase inicial da vida, até aos cinco anos, utilizam mais a cobertura de Internamento do que dos seis aos 30 anos, sendo 2,5% o valor mínimo da taxa de utilização atingindo entre os 11 e os 15 anos. A partir dos 36 anos a taxa de utilização é crescente com a idade, sendo que para idades superiores a 75 anos a taxa de utilização é de 29,1%.

Antiguidade

	E(N|V)
Segurado Padrão	6,0%
Antiguidade 0	6,0%
Antiguidade 1	6,0%
Antiguidade 2	6,0%
Antiguidade [3, 5]	6,0%
Antiguidade [6, 8]	6,0%

	E(N|V)
Segurado Padrão	6,0%
Antiguidade [9, 11]	5,5%
Antiguidade [12, 14]	4,6%
Antiguidade [15, 17]	4,6%
Antiguidade [18, 30]	4,6%

Tabela 4.10 - Taxa de utilização do modelo com as variáveis internas para a Antiguidade

Relativamente à Antiguidade, o modelo distinguiu apenas três classes distintas: as pessoas que têm o seguro até há 8 anos, entre 9 e 11 anos ou as pessoas que têm o seguro há mais de 11 anos, sendo que a taxa de utilização do Internamento é decrescente ao longo da Antiguidade. Contudo, na análise descritiva desta variável, tinha-se visto que a taxa de utilização tinha uma tendência crescente à medida que a antiguidade aumentava. Na Tabela 4.11, vê-se a diferença entre os valores reais e os valores preditos pelo modelo e, ainda, a idade média que é a justificação para esta diferença. As três classes da Antiguidade têm idades médias muito distintas e crescentes com a antiguidade, se se sabe que a partir dos 30 anos a utilização aumenta com a idade, o facto de os valores reais estarem a crescer com a Antiguidade pode dever-se à idade e não à Antiguidade. Para se poder perceber comportamento da Antiguidade, teria de se ter classes com idade média semelhante para que não houvesse outros efeitos a influenciar a variável no modelo. O que o modelo faz é considerar o comportamento da antiguidade mantendo as outras variáveis constantes, neste caso a idade está fixa na classe do segurado padrão ([36, 45] anos), e vê quais das taxas de utilização por cada classe. O modelo mostra que mantendo a idade e as outras variáveis fixas no segurado padrão, a taxa de utilização não aumenta com a antiguidade, mas sim diminui. Assim, no modelo com as variáveis internas e as variáveis externas a antiguidade não será considerada, mesmo sendo indicada como significativa.

Antiguidade	Real	Previsto	Idade Média
[0, 8]	4,2%	6,0%	32
[9, 11]	4,1%	5,5%	35
[12, 30]	4,6%	4,6%	44

Tabela 4.11 - Comparação da taxa de utilização real e prevista pelo modelo com as variáveis internas para a Antiguidade

Zona

	E(N|V)
Segurado Padrão	6,1%
Beira Litoral Centro	3,7%
Açores, Alto Alentejo, Beira Interior Norte, Beira Interior Sul, Beira Litoral Sul, Estrangeiro, Madeira	5,1%
Alto Minho, Baixo Alentejo, Ribatejo, Trás-Algarve, Baixo Minho, Beira Litoral Norte, Grande Lisboa, Grande Porto, Marão, Margem Sul, Oeste	6,0% / 6,1%

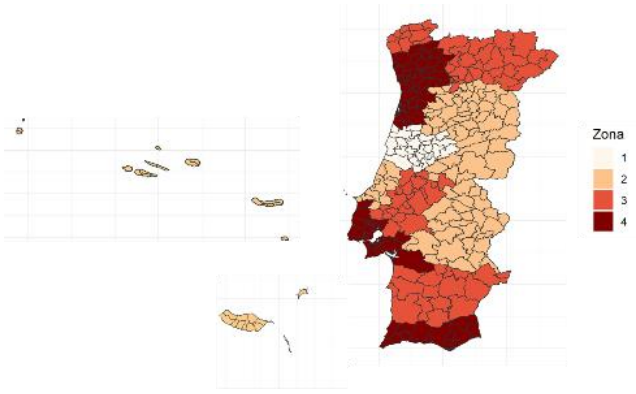


Figura 4.4 - Taxa de utilização do modelo com as variáveis internas para a Zona

No caso da Zona, relativamente aos agrupamentos feitos anteriormente, apenas se juntou a segunda zona com mais risco com a primeira; assim as zonas Algarve, Margem Sul, Oeste, Beira Litoral Norte, Baixo Minho, Marão, Grande Porto e Grande Lisboa têm todas uma taxa de utilização idêntica e é a classe com taxa de utilização mais alta face às outras zonas do país. Na Figura 4.4 , é possível observar a distribuição da taxa de utilização do Internamento por Portugal, sendo que as zonas estão ordenadas pelo risco, ou seja, a zona 1 é a zona de menor risco e a zona 4 é a de maior risco.

Garantias do Produto

	E(N|V)
Segurado Padrão	6,0%
Garantias Produto I	1,8%
Garantias Produto I + Acons	3,8%
Garantias Produto I + A	6,0%

	E(N|V)
Segurado Padrão	6,0%
Garantias Produto I + A + PO	6,0%
Garantias Produto I + A + E	6,0%
Garantias Produto I + A + E + PO	6,0%

Tabela 4.12 - Taxa de utilização do modelo com as variáveis internas para as Garantias do Produto

Quanto às garantias do produto, o modelo distingue os produtos em três tipos: os que só têm a cobertura de Internamento, os que têm Internamento mais um número de consultas limitado em Ambulatório e os que têm Internamento, Ambulatório e que podem ainda ter outra cobertura como Estomatologia ou Próteses e Ortóteses. Quanto maior a robustez do seguro, maior a taxa de utilização do Internamento.

Canal Comercial

	E(N|V)
Segurado Padrão	6,0%
Canal A	5,2%
Canal H	5,2%
Canal D	6,0%
Canal E	6,0%

	E(N|V)
Segurado Padrão	6,0%
Canal F	6,0%
Canal C	7,4%
Canal G	7,4%
Canal B	7,4%

Tabela 4.13 - Taxa de utilização do modelo com as variáveis internas para o Canal Comercial

No caso do Canal Comercial, o modelo indica que os seguros vendidos no Canal A e no Canal H são os que têm uma taxa de utilização do Internamento mais baixa. Depois vêm os produtos comprados no Canal D, no Canal E ou no Canal F. Por fim, os que apresentam uma maior taxa de utilização do Internamento são os seguros provenientes dos Canais B, C ou G.

Forma de Pagamento

	E(N|V)
Segurado Padrão	6,0%
Forma Pagamento A	6,0%
Forma Pagamento B	7,3%
Forma Pagamento C	7,3%
Forma Pagamento D	7,3%

Tabela 4.14 - Taxa de utilização do modelo com as variáveis internas para a Forma de Pagamento

Quanto à forma de pagamento, a taxa de utilização é idêntica para as Formas de pagamento B, C ou D e é superior para os que têm o fracionamento A.

Modelo da Utilização com as variáveis internas e as variáveis externas

Análise Inicial

As variáveis externas introduzidas no modelo estão selecionadas a verde na Tabela 4.15, sendo estas a que melhor explicam a utilização do Internamento, dentro das variáveis correlacionadas. Assim, das

31 variáveis externas, apenas 19 vão ser incluídas no modelo para verificar a sua significância. No total, vão ser introduzidas no modelo 28 variáveis, que podem ser observadas no ANEXO 7.

Casa	Escolaridade	Atividade Económica	Saúde	Casa e Poder Compra	Casa e Saúde
VAL_OFFER_AVG	PROP_ENS_BASC_COMP	PROP_EMP_SEQ	TX_UTL_CONS_1_ANO	VAL_OFFER_AVG	VAL_RENT_REQUEST_AVG
VAL_TRANSACT_AVG	PROP_ENS_SUP_COMP	PROP_EMP_TERC	TX_UTL_CONS_3_ANO	VAL_TRANSACT_AVG	VAL_RENT_CONTRACT_AVG
VAL_RENT_REQUEST_AVG		PROP_SEM_ATIV_ECON	TX_INT_CIRG	VAL_RENT_REQUEST_AVG	TX_UTL_CONS_3_ANO
VAL_RENT_CONTRACT_AVG		PROP_PENS_REFORM	TX_CONS	VAL_RENT_CONTRACT_AVG	
			TX_INT_CIRG	PER_CAPITA_INDICATOR	
			TX_DOENTES_SAIDOS_INT		
			TX_CONS		
			TX_DOENTES_SAIDOS_INT		

Tabela 4.15 - Variáveis externas introduzidas no modelo da utilização

Significância Estatística

Os resultados do teste de razão de verosimilhanças podem ser observados na Tabela 4.3.1.22.

Variável	Valor-p
Prop_ens_possec_comp	0,950
Prop_ens_sup_comp	0,944
Prop_3_4_div	0,781
Prop_pens_reform	0,734
Prop_emp_prim	0,675
Prop_emp_seq	0,515
Tx_cons_tempo_adq	0,550
Tipo de Pagamento	0,541

Variável	Valor-p
Taxa de urgências	0,354
Tx_int_cirg	0,181
Prop_analfabetos	0,152
Género	0,137
Val_rent_contract_avg	0,129
Tx_utl_cons_1_ano	0,222
Prop_area_50	0,090
Prop_area_200	0,194

Variável	Valor-p
Prop_area_100_200	0,252
Prop_area_50_100	0,545
Estado Civil	0,232
Prop_ens_sec_comp	0,088
Parentesco	0,042
Forma de Pagamento	0,029
Tx_mulheres_mam	0,018
Prop_1_2_div	2,35E-04

Variável	Valor-p
Capital	4,84E-08
Zona	6,36E-07
Canal Comercial	<2,2e-16
Antiguidade	<2,2e-16
Garantias do Produto	<2,2e-16
Idade	<2,2e-16

Tabela 4.16 - Significância das variáveis internas e das variáveis externas no modelo da utilização

As variáveis significativas, para um nível de significância de 5%, são as mesmas que se revelaram significativas no modelo só com as variáveis internas mais duas variáveis externas: a Taxa de mulheres com mamografia realizada nos últimos dois anos e a Proporção de habitações com 1 a 2 divisões. Pelo método de seleção *Stepwise Forward* os resultados são semelhantes; além das variáveis significativas também encontradas pelo teste de razão de verosimilhanças, foram identificadas, como diferenciadoras, a Proporção de indivíduos com o ensino secundário completo, a Proporção de habitações com área até 50 m2 e o Género. Estas variáveis baixavam o AIC, mas em valores muito pequenos; dependendo do nível de significância considerado, também poderiam ter sido escolhidas pelo teste de razão de verosimilhanças. Para este trabalho, o nível de significância utilizado é 5% e, por isso, as variáveis consideradas diferenciadoras da utilização são as identificadas pelo teste de razão de verosimilhança.

Das variáveis significativas, para além das variáveis internas já identificadas como correlacionadas, pode ser observado no ANEXO 3 que existe correlação entre a Zona e a Taxa de mulheres com mamografia realizada nos últimos dois anos (0,73) pelo que esta última variável será retirada do modelo por explicar menos a utilização que a Zona. Ao excluir estas variáveis do modelo e a Antiguidade, todas as outras variáveis continuaram significativas exceto a Forma de Pagamento.

Identificação da Homogeneidade

No modelo com as variáveis internas e as variáveis externas, optou-se por utilizar o teste de *Tukey* para testar se as classes eram todas heterogéneas. No caso das variáveis internas, os agrupamentos são semelhantes aos do modelo só com estas variáveis, as diferenças que existem podem dever-se à variabilidade da amostra. No caso das variáveis externas, decidiu-se transformar as variáveis significativas em variáveis categóricas. A Proporção de habitações com 1 a 2 divisões foi inicialmente em dividida em duas classes: <= 25% e > 25%, isto porque existem poucas pessoas com a segunda classe, e o teste de *Tukey* indicou que estas duas classes eram heterogéneas.

Modelo Final

O segurado padrão para o modelo da utilização das variáveis internas e das variáveis externas tem as características apresentadas na Tabela 4.17. Neste caso, vemos que as características não são exatamente as mesmas que no modelo só com as variáveis internas, por isso, a taxa de utilização de referência será diferente. Por exemplo, neste modelo tem-se uma taxa de utilização média mais baixa (6,0% vs 3,5%), isto porque estão a ser consideradas pessoas mais novas, de outras zonas, com produtos sem estomatologia, entre outros fatores.

Idade	Antiguidade	Zona	Garantias do Produto	Canal Comercial	Prop_1_2_div
[16, 35]	[0, 8]	Algarve, Grande Lisboa, Beira Interior Norte, Beira Litoral Sul, Marão, Margem Sul, Ribatejo, Baixo Alentejo, Alto Alentejo, Beira Interior Sul, Trás-os-Montes	I + A, I + A + PO	C, D, F	<= 25%

Tabela 4.17 - Características do segurado padrão do modelo da utilização com as variáveis internas

Idade

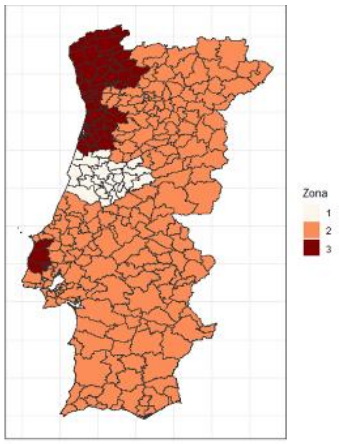
	E(N|V)
Segurado Padrão	3,5%
Idade [0,5]	4,8%
Idade [6,10]	2,3%
Idade [11,15]	2,3%
Idade [16,20]	3,5%
Idade [21,25]	3,5%
Idade [26,30]	3,5%
Idade [31,35]	3,5%
Idade [36,40]	4,7%
Idade [41,45]	4,7%
Idade [46,50]	6,0%

	E(N|V)
Segurado Padrão	3,5%
Idade [51,55]	7,5%
Idade [56,60]	7,5%
Idade [61,65]	10,9%
Idade [66,70]	15,3%
Idade [71,75]	15,3%
Idade [76,80]	20,0%
Idade [81,85]	20,0%
Idade [86,90]	20,0%
Idade [91,95]	20,0%

Tabela 4.18 - Taxa de utilização do modelo com as variáveis internas e as variáveis externas para a Idade

A idade continua a ser a variável mais significativa e observa-se o mesmo comportamento que no modelo só com as variáveis internas, ou seja, a taxa de utilização é mais baixa dos seis aos 15 anos e a partir daí cresce com as idades. Em termos de agrupamentos das categorias homogéneas, este modelo indicou que as pessoas entre os seis e os 15 anos têm o mesmo tipo de taxa de utilização, a classe dos [36, 40] tem taxa de utilização idêntica à classe [16,20] e o mesmo acontece para as classes dos [51, 55] e dos [56, 60]. O modelo só com as variáveis internas não tinha agrupado estas classes, sendo que elas tinham uma diferença, em média, de cerca 1% a 2%. Importa relembrar que os modelos têm amostras diferentes e algumas diferenças podem dever-se à variabilidade da amostra.

Zona



	E(N|V)
Segurado Padrão	3,5%
Beira Litoral Centro	1,8%
Algarve, Grande Lisboa, Beira Interior Norte, Beira Litoral Sul, Marão, Margem Sul, Ribatejo, Baixo Alentejo, Alto Alentejo, Beira Interior Sul, Trás-os-Montes	3,5%
Minho, Alto Minho	3,8%

Tabela 4.19 - Taxa de utilização do modelo com as variáveis internas e as variáveis externas para a Zona

No que diz respeito à Zona, o modelo com as variáveis internas e as variáveis externas distinguiu três classes de utilização distinta, enquanto que o modelo só com as variáveis internas tinha identificado quatro classes. Em ambos os modelos, a Beira Litoral Centro é a zona com taxa de utilização do Internamento mais baixa e esta é mais elevada no Grande Porto, na Beira Litoral Norte, no Oeste e no Baixo Minho. As zonas que pertencem à segunda classe, que têm uma taxa de utilização de 3,5%, também eram as zonas da classe com a segunda taxa de utilização mais baixa, com a exceção de Trás-os-Montes, que no modelo só com as variáveis internas estava na classe com a segunda taxa de utilização mais alta. As restantes zonas distribuem-se entre a segunda classe com taxa de utilização de 3,5% e a terceira classe com taxa de utilização de 3,8%. Nos dois modelos, não existe muita diferença entre as últimas classes, pelo que é natural que exista alguma diferença nos agrupamentos feitos.

Garantias do Produto

	E(N|V)
Segurado Padrão	3,5%
Garantias Produto I	1,5%
Garantias Produto I + Acons	2,2%
Garantias Produto I + A	3,5%

	E(N|V)
Segurado Padrão	3,5%
Garantias Produto I + A + PO	3,5%
Garantias Produto I + A + E	4,1%
Garantias Produto I + A + E + PO	4,1%

Tabela 4.20 - Taxa de utilização do modelo com as variáveis internas e as variáveis externas para as Garantias do Produto

Quanto às Garantias do Produto, o modelo com as variáveis internas e as variáveis externas distinguiu os produtos que têm estomatologia, dentro dos produtos com Internamento e Ambulatório, atribuindo-lhes a maior taxa de utilização, enquanto que, no modelo só com as variáveis internas, esses produtos têm taxas de utilização homogéneas. Em termos de comportamento da taxa de utilização do Internamento, em ambos os modelos, uma maior taxa de utilização está associada a produtos com mais coberturas.

Canal Comercial

	E(N|V)
Segurado Padrão	3,5%
Canal A	2,5%
Canal H	2,5%
Canal D	3,5%
Canal C	3,5%

	E(N|V)
Segurado Padrão	3,5%
Canal F	3,5%
Canal B	3,7%
Canal G	3,7%
Canal E	3,7%

Tabela 4.21 - Taxa de utilização do modelo com as variáveis internas e as variáveis externas para o Canal Comercial

No caso do Canal Comercial, o modelo indica que os seguros vendidos no Canal A ou no Canal H são os que têm, em média, uma taxa de utilização do Internamento mais baixa. No que diz respeito à taxa de utilização do Internamento mais alta, em ambos os modelos foram identificados o Canal B e O Canal G. Enquanto neste modelo o Canal E foi incluído na última classe, no modelo só com as variáveis internas esta estava na segunda classe e o Canal C estava na última classe. Tal como na variável Zona, o facto de estas duas classes apenas diferirem, em média, em 0,2 pontos percentuais pode resultar em algumas diferenças nos agrupamentos, devido à variabilidade das amostras.

Proporção de habitações com 1 a 2 divisões

	E(N|V)
Segurado Padrão	3,5%
Prop area 1_2div <= 25%	3,5%
Prop area 1_2div > 25%	5,6%

Tabela 4.22 - Taxa de utilização do modelo com as variáveis internas e as variáveis externas para a Proporção de habitações com 1 a 2 divisões

A proporção de habitações com 1 a 2 divisões é uma variável externa que se revelou significativa e indica que, quando esta é menor ou igual a 25%, a taxa de utilização do Internamento é menor do que quando esta é superior a 25%, ou seja, quando existe uma maior proporção de as casas terem 1 ou 2 divisões, isso leva a que a taxa de utilização do Internamento seja maior.

Em suma, podemos verificar que a utilização do Internamento tem algumas variáveis explicativas, sendo na maioria variáveis internas que explicam mais do que as variáveis externas. Das variáveis externas estudadas, apenas a Proporção de habitações com 1 a 2 divisões se revelou significativa. Das variáveis internas, apenas a Forma de Pagamento deixou de ser significativa quando se introduziram as variáveis externas. As variáveis mais diferenciadoras da utilização do Internamento são a Idade, as Garantias do Produto e o Canal Comercial (valor-p <2,2e-16).

4.3.2. Modelação do Custo

Para se construir o modelo do custo médio da utilização do Internamento é necessário considerar quatro partes distintas: os custos do Tipo I (até 10.500 euros), os custos do Tipo II (entre 10.500 euros e 27.550 euros), os custos do Tipo III (superiores a 27.550 euros) e a probabilidade de uma pessoa ter um dos tipos de custo.

Custos Tipo I

Modelo dos Custos Tipo I com as variáveis internas

Análise Inicial

Para os custos até 10.500 euros, que são os custos mais frequentes, existem 10.547 pessoas na amostra, assim considerou-se que cada classe das variáveis deveria ter no mínimo 500 pessoas para ser representativo. Com este pressuposto, foram feitos mais agrupamentos do que no modelo da utilização; no caso da Idade, foram agrupadas as classes dos 0 aos 15 anos, dos 16 aos 30 anos, os que têm mais de 70 anos e os restantes escalões mantiveram-se. Relativamente à Antiguidade, considerou-se uma classe única para 0 ou 1 ano e uma classe única a partir dos 15 anos, em vez dos 18 anos de antiguidade. Na Zona de residência foram criadas nove classes distintas sendo que, quanto maior o número da zona, maior o risco. Os resultados podem ser observados no ANEXO 4. Quanto às Garantias do produto, agruparam-se os produtos com Internamento e Ambulatório aos que têm essas coberturas mais a cobertura Próteses e Ortóteses. O Capital está agrupado nas mesmas classes que na utilização, e no Canal Comercial foi necessário agrupar a classe H à classe A, e a classe B à classe E.

Significância Estatística

Os resultados do teste de razão de verosimilhanças podem ser observados na Tabela 4.23.

Variável	Valor-p
Tipo de Pagamento	0,737
Estado Civil	0,643
Forma de Pagamento	0,601
Antiguidade	0,519
Capital	0,453
Garantias do Produto	0,328

Variável	Valor-p
Parentesco	0,079
Género	0,024
Canal Comercial	0,023
Zona	4,03E-15
Idade	<2,2E-20

Tabela 4.23 - Significância das variáveis internas e variáveis externas no modelo do custo do tipo I

As variáveis que se revelaram significativas para os custos até os 10.500 euros, pelo teste de razão de verosimilhança, foram a Idade, a Zona, o Canal Comercial e o Género para um nível de significância de 5%. Os métodos de seleção *Backward, Forward* e *Stepwise Forward* deram os mesmos resultados e a ordem das variáveis que mais contribui para explicar os custos do tipo I é dada pela Idade, a Zona, o Género e o Canal Comercial.

No custo do tipo I, as variáveis significativas não são correlacionadas entre si tal como pode ser visto no ANEXO 3.

Identificação da Homogeneidade

Idade

Na Idade, o teste de *Wald* indicou que, para as idades dos 16 aos 40, o custo era homogéneo e o teste de *Tukey* agrupou ainda as idades dos 41 aos 55 anos; a classe das idades mais avançadas, em vez de começar para quem tem mais de 75 anos, abrange os que têm mais de 55 anos.

Zona

No caso da Zona, a análise de *clusters* e a árvore de decisão deram os mesmos resultados. Das nove classes ficaram apenas quatro, sendo que as zonas de risco 1 a 4 e 6 a 8 foram agrupadas, como pode ser visto no ANEXO 6. Mais tarde, ao se introduzir estes novos agrupamentos no modelo, o teste de *Wald* ainda indicou que a zona 5 não diferia da zona 6 a 8 (segurado padrão).

Canal Comercial

No caso do canal comercial, das seis classes, a análise de *clusters* obteve apenas três classes heterogéneas e a árvore de decisão resultou em cinco nodos.

Canal Comercial Agrupado - *Clusters*	Custo médio de utilização
B, E	2.908 €
A, H	3.029 €
C, D, G	3.227 €
F	3.407 €

Canal Comercial Agrupado - Árvores de Decisão	Custo médio de utilização
B, E	2.908 €
A, H	3.029 €
C, D	3.197 €
G	3.269 €
F	3.407 €

Tabela 4.24 - Agrupamento do custo do tipo I por *clusters* e árvores de decisão para as Garantias do Produto

Em termos de distinção de custo do tipo I, as técnicas são coerentes, apenas colocam os Canais A, H e G em classes distintas; no caso das árvores, deixam-nas como classes únicas; no caso dos *clusters,* são agrupados à classe de risco adjacente inferior.

Modelo	Deviance	Deviance_Nulo – Deviance Modelo	AIC
Nulo	4675	0	185042
Canal Comercial	4642	33	184971
Canal Comercial AC	4642	33	184968
Canal Comercial AD	4642	33	184969

Tabela 4.25 - *Deviance* e AIC para os modelos do custo do tipo I para as Garantias do Produto

O agrupamento escolhido para entrar no modelo final foi o da análise de *clusters* porque apresentou menor valor de *AIC*. Após a introdução do novo agrupamento no modelo, o teste de *Wald* indicou que as classes A, B, C, D, E e G não diferiam entre si.

Modelo Final

Os resultados do modelo dos custos da utilização do Internamento até 10.500 euros são apresentados de seguida, tendo como referência as características do segurado padrão dadas na Tabela 4.26.

Idade	Género	Zona	Canal Comercial
[16, 40]	F	Alto Alentejo, Baixo Alentejo, Beira Interior Norte, Grande Lisboa, Madeira, Margem Sul, Ribatejo, Trás-os-Montes	A, B, C, D, E, G, H

Tabela 4.26 - Características do segurado padrão do modelo do custo do tipo I com as variáveis internas

Idade

	E(X<=10.500|V)
Segurado Padrão	2.994 €
Idade [0,5]	2.147 €
Idade [6,10]	2.147 €
Idade [11,15]	2.147 €
Idade [16,20]	2.994 €
Idade [21,25]	2.994 €
Idade [26,30]	2.994 €
Idade [31,35]	2.994 €
Idade [36,40]	2.994 €
Idade [41,45]	3.292 €
Idade [46,50]	3.292 €

	E(X<=10.500|V)
Segurado Padrão	2.994 €
Idade [51,55]	3.292 €
Idade [56,60]	3.689 €
Idade [61,65]	3.689 €
Idade [66,70]	3.689 €
Idade [71,75]	3.689 €
Idade [76,80]	3.689 €
Idade [81,85]	3.689 €
Idade [86,90]	3.689 €
Idade [91,95]	3.689 €

Tabela 4.27 - Custo médio do tipo I para a o modelo com variáveis internas por Idade

O modelo final indica que a idade é diferenciadora dos custos até 10.500 euros, sendo que as classes heterogéneas são menores do que no modelo da utilização; o comportamento dos custos é crescente ao longo das idades.

Género

	E(X<=10.500|V)
Segurado Padrão	2.994 €
Género F	2.994 €
Género M	3.076 €

Tabela 4.28 - Custo médio do tipo I para a o modelo com variáveis internas por Género

O género da pessoa revelou-se diferenciador para os custos até 10.500 euros, sendo que os homens têm, em média, um custo superior ao das mulheres.

Zona

	E(X<=10.500|V)
Segurado Padrão	2.994 €
Algarve, Alto Minho, Baixo Minho, Beira Litoral Centro, Beira Litoral Norte, Beira Litoral Sul, Grande Porto, Marão	2.725 €
Alto Alentejo, Baixo Alentejo, Beira Interior Norte, Grande Lisboa, Mdeira, Margem Sul, Ribatejo, Trás-os-Montes	2.994 €
Açores, Beira Interior Sul, Estrangeiro, Oeste	3.402 €

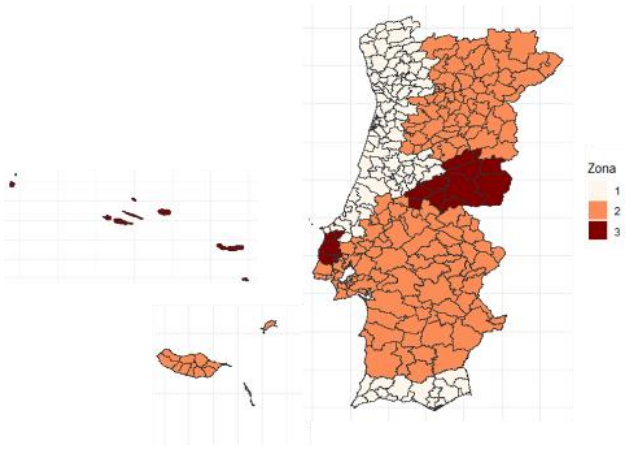


Figura 4.5 - Custo médio do tipo I para o modelo com variáveis internas por Zona

Quanto à Zona de residência, uma pessoa que viva numa das zonas Algarve, Alto Minho, Baixo Minho, Beira Litoral, Marão ou Grande Porto tem, em média, menos custos, dentro dos custos até 10.500. As pessoas que têm mais custos, dentro dos custos até 10.500, são as que vivem nos Açores, no Estrangeiro, no Oeste ou na Beira Interior Sul.

Canal Comercial

	E(X<=10.500|V)
Segurado Padrão	2.994 €
Canal A	2.994 €
Canal B	2.994 €
Canal C	2.994 €
Canal H	2.994 €

	E(X<=10.500|V)
Segurado Padrão	2.994 €
Canal D	2.994 €
Canal E	2.994 €
Canal G	2.994 €
Canal F	3.123 €

Tabela 4.29 - Custo médio do tipo I para o modelo com variáveis internas por Canal Comercial

Relativamente ao Canal Comercial, o modelo distingue as pessoas que compraram o seguro pela no Canal F dos restantes, sendo que os custos médios diferem, em média, em 129 euros.

Modelo dos Custos Tipo I com as variáveis internas e as variáveis externas

Análise Inicial

No modelo com as variáveis internas e as externas, existem 5.995 pessoas com custos até 10.500 euros. As variáveis externas introduzidas no modelo são as selecionadas a verde na Tabela 4.30. Assim, das 31 variáveis externas, apenas 19 vão ser incluídas no modelo para verificar a sua significância. No total, vão ser introduzidas no modelo 28 variáveis que podem ser observadas no ANEXO 7.

Casa	Escolaridade	Atividade Económica	Saúde	Casa e Poder Compra	Casa e Saúde
VAL_OFFER_AVG	PROP_ENS_BASC_COMP	PROP_EMP_SEQ	TX_UTL_CONS_1_ANO	VAL_OFFER_AVG	VAL_RENT_REQUEST_AVG
VAL_TRANSACT_AVG	PROP_ENS_SUP_COMP	PROP_EMP_TERC	TX_UTL_CONS_3_ANO	VAL_TRANSACT_AVG	VAL_RENT_CONTRACT_AVG
VAL_RENT_REQUEST_AVG		PROP_SEM_ATIV_ECON	TX_INT_CIRG	VAL_RENT_REQUEST_AVG	TX_UTL_CONS_3_ANO
VAL_RENT_CONTRACT_AVG		PROP_PENS_REFORM	TX_CONS	VAL_RENT_CONTRACT_AVG	
			TX_INT_CIRG	PER_CAPITA_INDICATOR	
			TX_DOENTES_SAIDOS_INT		
			TX_CONS		
			TX_DOENTES_SAIDOS_INT		

Tabela 4.30 - Variáveis externas introduzidas no modelo do custo do tipo I

Significância Estatística

As variáveis significativas dos custos até 10.500 euros são, para um nível de significância de 5%, são a Idade, a Zona, o Género e o Canal Comercial. As outras variáveis foram rejeitadas e os valores-p podem ser observados na Tabela 4.31. Pelo método de seleção *Stepwise Forward* os resultados são semelhantes; além das variáveis significativas também encontradas pelo teste de razão de verosimilhanças, foi identificada a variável Valor médio da renda contratada por m2, mas esta não reduziu o AIC, apenas o manteve. Para este trabalho, o nível de significância utilizado é 5% e, por isso, as variáveis consideradas diferenciadoras dos custos do tipo I do Internamento são as identificadas pelo teste de razão de verosimilhança.

Variável	Valor-p
Tipo de Pagamento	0,911
Capital	0,897
Prop_ens_sup_comp	0,876
Prop_3_4_div	0,849
Prop_ens_possec_comp	0,789
Prop_ens_sec_comp	0,769
Prop_pens_reform	0,751
Prop_analfabetos	0,698

Variável	Valor-p
Prop_1_2_div	0,746
Tx_utl_cons_1_ano	0,691
Tx_mulheres_mam	0,590
Estado Civil	0,576
Parentesco	0,572
Antiguidade	0,484
Tx_cons	0,480
Forma de Pagamento	0,388

Variável	Valor-p
Prop_ens_prim	0,363
Prop_area_50	0,308
Prop_area_100_200	0,578
Prop_emp_seq	0,326
Prop_area_50_100	0,194
Prop_area_200	0,278
Garantias do Produto	0,167
Tx_urgências	0,080

Variável	Valor-p
Tx_cons_temp_adq	0,120
Val_rent_contract_avg	0,154
Canal Comercial	0,012
Género	0,008
Zona	1,14E-06
Idade	<2,2e-16

Tabela 4.31 - Significância das variáveis internas e das variáveis externas no modelo do custo do tipo I

Neste caso, observa-se que a inclusão das variáveis externas não trouxe mudanças no modelo obtido anteriormente, as variáveis internas que antes se revelaram significativas continuaram a sê-lo e nenhuma variável externa veio contribuir para explicar os custos até 10.500 euros. O modelo obtido é idêntico ao modelo só com as variáveis internas, só que com uma amostra mais reduzida. As características do segurado padrão para o modelo da probabilidade do tipo de custos são as que estão na Tabela 4.32.

Idade	Género	Zona	Canal Comercial
[16, 40]	F	Alto Alentejo, Baixo Alentejo, Grande Lisboa, Margem Sul, Ribatejo, Trás-os-Montes, Baixo Minho, Alto Minho, Beira Interior Sul, Oeste	A, B, C, D, E, G

Tabela 4.32 - Características do segurado padrão do modelo do custo do tipo I com as variáveis internas e com as variáveis externas

Os resultados em termos de comportamento dos custos do tipo I para a Idade, para o Género e para o Canal Comercial foram semelhantes ao modelo só com as variáveis internas e podem ser observados no ANEXO 8. No caso da Zona, existem na mesma três classes de custo, mas os agrupamentos têm algumas diferenças que são apresentadas na Figura 4.6.

Zona

	E(X<=10.500|V)
Segurado Padrão	3.026 €
Beira Litoral Centro, Marão, Beira Interior Norte, Algarve	2.506 €
Beira Litoral Norte, Beira Litoral Sul, Grande Porto	2.757 €
Ribatejo, Alto Alentejo, Baixo Minho, Alto Minho, Margem Sul, Trás-os-Montes, Beira Interior Sul, Grande Lisboa, Baixo Alentejo, Oeste	3.026 €

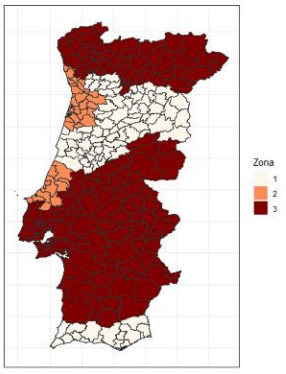


Figura 4.6 - Custo médio do tipo I para o modelo com as variáveis internas e as variáveis externas por Zona

As diferenças entre o modelo só com as variáveis internas e o modelo com as variáveis internas e as variáveis externas são maiores para o Alto Minho e o Baixo Minho, que, no primeiro modelo, estavam na Zona 1, e, no segundo modelo, mudaram para a Zona 3, ou seja, passaram da zona com menor risco para a zona com maior risco. O custo médio destas zonas na amostra do modelo só com as variáveis internas é 2.856 euros e, no modelo que inclui também as variáveis externas, 3.283 euros, ou seja, o agrupamento feito deve-se à aleatoriedade da amostra e, em cada um dos modelos, a classe foi bem atribuída. Em ambos os modelos, o número de observações é inferior a 1.000 pessoas; por isso, talvez seja necessário haver uma amostra com maior dimensão para retirar conclusões mais sólidas. No que diz respeito às diferenças entre outras zonas, a mudança foi mais pequena: a Beira Litoral Norte, a Beira Litoral Sul e o Grande Porto mudaram da Zona 1 para a Zona 2 e a Margem Sul, o Ribatejo e Trás-os-Montes mudaram da Zona 3 para a Zona 2. Importa salientar que as zonas que tinham sido consideradas de mais risco no modelo com as variáveis internas foram todas identificadas com maior risco no modelo com as variáveis internas e as variáveis externas.

Custos Tipo II e III

Para os custos do Tipo II e III apenas serão apresentados os resultados para o modelo com as variáveis internas e com as variáveis externas, porque há várias externas significativas e a metodologia para os dois modelos é idêntica, apenas alteram os valores estimados dos parâmetros das distribuições.

Na amostra existem 343 pessoas com custos do Tipo II e 61 pessoas com custos do Tipo III. Por não haver dimensão suficiente, considera-se o valor médio da distribuição que melhor se ajusta aos dados; em ambos os casos foi a distribuição Pareto Generalizada. No caso dos custos tipo II, a distribuição Pareto Generalizada com parâmetros de forma = 0,18, de escala = 2.278 e de localização = 10.500, o valor médio é dado por:

$$10.500 + \frac{2.278}{(1 - 0{,}18)} = 13.278\ €$$

E para os custos tipo III, a distribuição que se ajusta é a Pareto Generalizada com parâmetros de forma = 0,17, de escala = 7.437 e de localização = 27.550, e o valor médio é dado por:

$$27.550 + \frac{7.437}{(1 - 0{,}17)} = 36.510\ €$$

Probabilidade do Tipo de Custos

Modelo da Probabilidade do Tipo de Custos com as variáveis internas

Análise Inicial

Para este modelo utilizou-se como amostra todos os utilizadores e considerou-se que, para que as classes tivessem amostra suficiente, teria de se ter pelo menos 500 pessoas, o mesmo que foi assumido para o modelo do custo do tipo I. Com este pressuposto, na Idade foram criadas classes dos 0 aos 10 anos, dos 11 aos 20 anos, dos 21 aos 25, as pessoas com mais de 70 anos e os restantes escalões mantiveram-se. Na Antiguidade, agruparam-se dos 0 a 1 anos, dos 9 aos 14 anos e os com antiguidade igual ou superior a 15 anos. Na Zona de residência, foram criadas nove classes distintas, sendo que, quanto maior o número da zona, maior o risco, os resultados podem ser observados no ANEXO 4. Quanto às Garantias do produto e ao Canal Comercial, estão agrupados nas mesmas classes que no modelo dos custos do tipo I e o Capital está agrupado nas mesmas classes que no modelo da utilização.

Significância Estatística

Os resultados do teste de razão de verosimilhança podem ser observados na Tabela 4.33 e indicam que as variáveis que se revelaram significativas foram a Idade, a Zona e as Garantias do Produto. Pelos métodos de seleção *Backward*, *Forward* e *Stepwise* foram identificadas como significativas as mesmas mais o género, sendo que este fez baixar o AIC de 5.178,22 para 5.176,62. Contudo, o género não será considerado como significativo porque pelo teste de razão de verosimilhanças rejeita-se, para um nível de significância de 5%, a hipótese de o género ser uma variável significativa.

Variável	Valor-p
Forma de Pagamento	0,744
Tipo de Pagamento	0,725
Antiguidade	0,585
Parentesco	0,535
Canal Comercial	0,237
Capital	0,238

Variável	Valor-p
Estado Civil	0,147
Género	0,075
Garantias do Produto	0,028
Zona	2,85E-07
Idade	<2,2E-16

Tabela 4.33 - Significância das variáveis internas no modelo da probabilidade do tipo de custos

Da análise de correlação das variáveis significativas, conclui-se que todas estas variáveis vão continuar no modelo, os resultados estão no ANEXO 3.

Identificação da Homogeneidade

Para o modelo da probabilidade do tipo de custos não se fez nenhum agrupamento, porque a variável resposta pode tomar vários valores e, por isso, para comparar as classes, estas teriam de ser homogéneas entre si para todos os tipos de custo.

Modelo Final

Os resultados do modelo da probabilidade de ter um dos tipos de custo da utilização do Internamento são apresentados de seguida, tendo como referência as características do segurado padrão dadas na Tabela 4.34.

Idade	Zona	Garantias do Produto
[36, 40]	Açores, Baixo Alentejo, Beira Interior Norte, Grande Lisboa	I + A, I + A + PO

Tabela 4.34 - Características do segurado padrão do modelo da probabilidade do tipo de custos com as variáveis internas

Idade

	E(PX1|V)	E(PX2|V)	E(PX3|V)
Segurado Padrão	95,0%	4,3%	0,7%
Idade [0,5]	99,8%	0,0%	0,2%
Idade [6,10]	99,8%	0,0%	0,2%
Idade [11,15]	96,9%	2,7%	0,4%
Idade [16,20]	96,9%	2,7%	0,4%
Idade [21,25]	98,3%	1,6%	0,1%
Idade [26,30]	98,3%	1,6%	0,1%
Idade [31,35]	96,3%	3,2%	0,5%
Idade [36,40]	95,0%	4,3%	0,7%
Idade [41,45]	94,6%	5,1%	0,3%

	E(PX1|V)	E(PX2|V)	E(PX3|V)
Segurado Padrão	95,0%	4,3%	0,7%
Idade [46,50]	93,0%	6,3%	0,7%
Idade [51,55]	92,1%	6,8%	1,1%
Idade [56,60]	89,7%	8,4%	1,9%
Idade [61,65]	88,8%	9,9%	1,3%
Idade [66,70]	89,7%	8,7%	1,6%
Idade [71,75]	86,7%	11,3%	2,0%
Idade [76,80]	86,7%	11,3%	2,0%
Idade [81,85]	86,7%	11,3%	2,0%
Idade [86,90]	86,7%	11,3%	2,0%
Idade [91,95]	86,7%	11,3%	2,0%

Tabela 4.35 - Probabilidade do tipo de custos para o modelo com as variáveis internas por Idade

O modelo final indica que a Idade é diferenciadora da probabilidade do tipo de custos, sendo que, nas idades iniciais, entre os 0 e os 10 anos, existe uma maior probabilidade de ter um custo até 10.500

euros, enquanto na classe dos 11 aos 15 anos já existe mais probabilidade de ter um custo do tipo II. À medida que a idade aumenta, a probabilidade de ter um custo até 10.500 euros decresce e a probabilidade de ter custos mais elevados (custos do tipo II ou custos do tipo III) aumenta.

Zona

	E(PX1|V)	E(PX2|V)	E(PX3|V)
Segurado Padrão	95,0%	4,3%	0,7%
Trás-os-Montes, Baixo Minho, Madeira, Estrangeiro	97,9%	2,0%	0,1%
Beira Litoral Norte	97,8%	2,0%	0,2%
Alto Minho, Marão, Beira Interior Sul, Beira Litoral Sul	97,1%	2,6%	0,3%
Beira Litoral Centro, Algarve	96,8%	2,8%	0,4%
Grande Porto	96,6%	3,1%	0,3%
Oeste, Alto Alentejo	95,6%	4,4%	0,0%
Ribatejo	95,6%	3,6%	0,8%
Margem Sul	95,2%	4,3%	0,5%
Beira Interior Norte, Grande Lisboa, Baixo Alentejo, Açores	95,0%	4,3%	0,7%

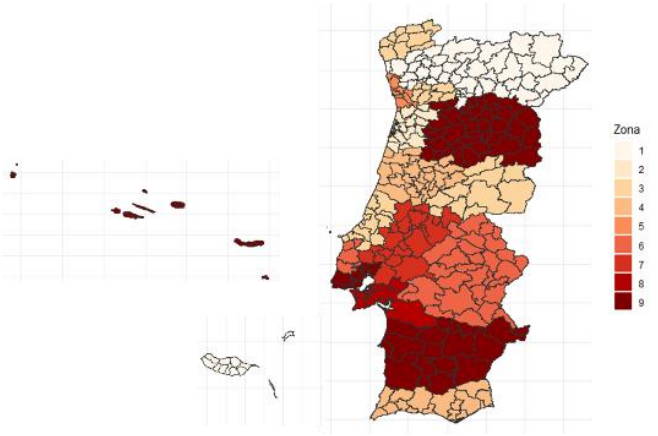


Figura 4.7 - Probabilidade do tipo de custos para o modelo com as variáveis internas por Zona

No caso da Zona, a probabilidade de ter um custo do tipo III é maior no Ribatejo e é menor no Alto Alentejo e no Oeste. Em termos da probabilidade de ter um custo do tipo II, as zonas com maior probabilidade são o Alto Alentejo e o Oeste, e as com menores probabilidades são o Baixo Minho, a Beira Litoral Norte, Trás-os-Montes, a Madeira e o Estrangeiro. No que diz respeito à probabilidade dos custos mais frequentes, até 10.500 euros, as zonas Baixo Minho, Trás-os-Montes, Madeira e Estrangeiro têm maior probabilidade e as zonas Beira Interior Norte, Grande Lisboa, Baixo Alentejo e Açores têm menos probabilidade.

Garantias do Produto

	E(PX1|V)	E(PX2|V)	E(PX3|V)
Segurado Padrão	95,00%	4,30%	0,70%
Garantias Produto I	95,9%	3,7%	0,4%
Garantias Produto I + Acons	94,96%	4,34%	0,70%
Garantias Produto I + A	95,00%	4,30%	0,70%
Garantias Produto I + A + PO	95,00%	4,30%	0,70%
Garantias Produto I + A + E	93,9%	4,9%	1,2%
Garantias Produto I + A + E + PO	93,2%	5,3%	1,5%

Tabela 4.36 - Probabilidade do tipo de custos para o modelo com as variáveis internas por Garantias do Produto

Relativamente às Garantias do Produto, um produto com mais coberturas tem menos probabilidade de ter custos do tipo I e maior probabilidade de ter custos do tipo II face a um produto mais simples.

Modelo da Probabilidade do Tipo de Custos com as variáveis internas e as variáveis externas

Análise Inicial

As variáveis externas introduzidas no modelo são as selecionadas a verde na Tabela 4.37, sendo, das variáveis correlacionadas, as que melhor explicam a probabilidade ter um dos tipos de custo. Assim,

das 31 variáveis externas, apenas 19 vão ser incluídas no modelo para verificar a sua significância. No total, vão ser introduzidas no modelo 28 variáveis que podem ser observadas no ANEXO 7.

Casa	Escolaridade	Atividade Económica	Saúde	Casa e Poder Compra	Casa e Saúde
VAL_OFFER_AVG	PROP_ENS_BASC_COMP	PROP_EMP_SEQ	TX_UTL_CONS_1_ANO	VAL_OFFER_AVG	VAL_RENT_REQUEST_AVG
VAL_TRANSACT_AVG	PROP_ENS_SUP_COMP	PROP_EMP_TERC	TX_UTL_CONS_3_ANO	VAL_TRANSACT_AVG	VAL_RENT_CONTRACT_AVG
VAL_RENT_REQUEST_AVG		PROP_SEM_ATIV_ECON	TX_INT_CIRG	VAL_RENT_REQUEST_AVG	TX_UTL_CONS_3_ANO
VAL_RENT_CONTRACT_AVG		PROP_PENS_REFORM	TX_CONS	VAL_RENT_CONTRACT_AVG	
			TX_INT_CIRG	PER_CAPITA_INDICATOR	
			TX_DOENTES_SAIDOS_INT		
			TX_CONS		
			TX_DOENTES_SAIDOS_INT		

Tabela 4.37 - Variáveis externas introduzidas no modelo da probabilidade do tipo de custos

Significância Estatística

Os resultados do teste de razão de verosimilhanças podem ser observados na Tabela 4.38.

Variável	Valor-p
Prop_area_200	0,999
Prop_ens_sec_comp	0,978
Tx_mulheres_mam	0,922
Prop_emp_prim	0,846
Tipo de Pagamento	0,820
Parentesco	0,802
Forma de Pagamento	0,761
Género	0,603

Variável	Valor-p
Prop_ens_sup_comp	0,604
Prop_area_100_200	0,355
Prop_area_50	0,397
Prop_1_2_div	0,626
Estado Civil	0,287
Prop_analfabetos	0,260
Prop_ens_possec_con	0,290
Tx_urgências	0,228

Variável	Valor-p
Tx_cons_tempo_adq	0,338
Canal Comercial	0,216
Prop_emp_seq	0,145
Prop_3_4_div	0,115
Prop_area_50_100	0,506
Tx_int_cirg	0,094
Capital	0,066
Garantias do Produto	0,196

Variável	Valor-p
Tx_utl_cons_1_ano	0,051
Zona	0,148
Prop_sem_ativ_econ	0,016
Val_rent_contract_avg	0,001
Idade	<2,2e-16

Tabela 4.38 - Significância das variáveis internas e das variáveis externas no modelo da probabilidade do tipo de custos

Por este método, as variáveis diferenciadoras da probabilidade de ter um dos tipos de custo são a Idade, o Valor de renda média contratada e a Proporção de indivíduos sem atividade económica, para um nível de significância de 5%, obtendo um modelo com valor AIC igual a 3.215. O modelo é dado por:

$$1: P_X \sim Idade + Val\_rent\_contract\_avg + Prop\_sem\_ativ\_econ$$

Pelo método *Backward*, as variáveis que contribuíram para um menor valor AIC (3.219) foram as identificadas pelo teste de razão de verosimilhanças, mais a Taxa de consultas realizadas em tempo adequado, a Taxa de urgências, a Proporção de habitações com área até 50 m2, a Proporção de habitações com área entre 50 a 100 m2, a Proporção de habitações com área entre 100 a 200 m2, a Proporção de habitações com área superior a 200 m2 e a Proporção de habitações com 3 a 4 divisões. Neste caso, pode escrever-se o modelo da seguinte forma:

$$2: P_X \sim Idade + Val\_rent\_contract\_avg + Prop\_sem\_ativ\_econ + Prop\_area\_50 + Prop\_area\_50\_100 + Prop\_area\_100\_200 + Prop\_area\_200 + Prop\_3\_4\_div + Tx\_urgências + Tx\_cons\_tempo\_adq$$

As variáveis selecionadas pelo método *Forward e Stepwise Forward* para o modelo com menor AIC (3.211) foram as identificadas pelo teste de razão de verosimilhanças, mais a Proporção de indivíduos com o ensino superior completo, a Taxa de consultas realizadas em tempo adequado e a Taxa de urgências. O modelo é dado por:

$$3: P_X \sim Idade + Val\_rent\_contract\_avg + Prop\_sem\_ativ\_econ + Tx\_urgências + Tx\_cons\_tempo\_adq + Prop\_ens\_sup\_comp$$

Pelos três resultados expostos acima, percebe-se que o método *Backward* é o que identifica mais variáveis explicativas, contudo as variáveis que foram selecionadas por este método e não foram pelos restantes podem ser de alguma forma correlacionadas e estar, por isso, a influenciar os valores. Apesar de se ter considerado o critério do nível de correlação a partir de 0,75, aqui poderá sentir-se a necessidade de ter usado um critério mais exigente; neste caso as correlações variam entre -0,04 e -0,69. Quanto às diferenças entre o método *Forward*/*Stepwise Forward* e o teste de razão de verosimilhanças, decidiu-se testar a significância das variáveis Tx_urgências, Tx_cons_tempo_adq e Prop_ens_sup_comp no modelo 3 pelo teste de razão de verosimilhanças. Os resultados deste teste, presentes na Tabela 4.39, indicam que, para um nível de significância de 5%, apenas a Idade, a Taxa de urgências e a Taxa de consultas em tempo adequado são significativas.

Variável	Idade	Val_rent_contract_avg	Prop_sem_ativ_econ	Tx_urgências	Tx_cons_tempo_adq	Prop_ens_sup_comp
Valor-p	<2.2e-16	0,0881	0,0515	0,0464	0,0227	0,1152

Tabela 4.39 - Significância das variáveis do modelo 3

Por outro lado, o modelo 1 apenas tem em comum como significativa a variável Idade. Assim, foi-se estudar a inclusão das variáveis Tx_urgências e Tx_cons_tempo adq no modelo 1, obteve-se que as duas variáveis não são significativas para 5% e os valores-p estão indicados na Tabela 4.40.

Variável	Tx_urgências	Tx_cons_tempo_adq	Prop_ens_sup_comp
Valor-p	0,1093	0,055	0,1128

Tabela 4.40 - Significância das variáveis do modelo 1

Apesar das variáveis Tx_urgências e Tx_cons_tempo_adq não terem sido significativas no modelo 1, elas foram no modelo 3; por isso, testou-se o seguinte modelo:

$$4\text{: } P_X \sim Idade + Val\_rent\_contract\_avg + Prop\_sem\_ativ\_econ + Tx\_urgências + Tx\_cons\_tempo\_adq$$

Variável	Idade	Val_rent_contract_avg	Prop_sem_ativ_econ	Tx_urgências	Tx_cons_tempo_adq
Valor-p	<2.2e-16	0,0015	0,0157	0,0455	0,0228

Tabela 4.41 - Significância das variáveis do modelo 4

Os resultados do modelo 4, apresentados na Tabela 4.41, mostram que todas as variáveis são significativas para 5%. A conclusão a que se chega destes testes é que as variáveis Tx_urgências e Tx_cons_tempo_adq são significativas no modelo 1 se ambas forem adicionadas, mas não são significativas se apenas uma delas entrar para o modelo, o que leva a crer que elas têm efeitos contrários (têm uma correlação de -0,2), porque, se tivessem o mesmo efeito, apenas uma delas já seria significativa.

Na modelação da probabilidade de ter um dos tipos de custo, vê-se que os métodos de seleção de variáveis deram resultados um pouco distintos, o que levanta algumas dúvidas sobre quais as variáveis que se devem considerar diferenciadoras. Os métodos de seleção *Backward*, *Forward* e *Forward Stepwise* baseiam-se em encontrar as variáveis que mais fazem baixar o *AIC,* e já se viu que, mesmo

que a variável cumpra esse requisito, nem sempre cumpre com o nível de significância de 5%. Por outro lado, as variáveis Tx_urgências e Tx_cons_tempo_adq não são significativas isoladamente, apenas o são na presença de outra variável. Assim sendo, decidiu-se que apenas se iria considerar que as variáveis diferenciadoras da probabilidade der ter um dos tipos de custo da utilização do Internamento, para um nível de significância de 5%, são a Idade, o Valor da renda média contratada por m2 e a Proporção de indivíduos sem atividade económica. Estas variáveis não estão correlacionadas tal como pode ser observado no ANEXO 3. As variáveis significativas externas foram transformadas em variáveis categóricas; o valor da renda contratada média por m2 foi dividida em quatro classes: [0,5],]5,8],]8,11] e >11; a proporção de indivíduos sem atividade económica foi dividida por três classes: <= 25%, 25% a 50% e > 50%.

Modelo Final

Os resultados do modelo da probabilidade de ter um dos tipos de custo da utilização do Internamento, com as variáveis internas e as variáveis externas, são apresentados de seguida, tendo como referência as características do segurado padrão dadas na Tabela 4.42.

Idade	Val_rent_contract_avg	Prop_sem_ativ_econ
[41, 45]	]5, 8]	25% a 50%

Tabela 4.42 - Características do segurado padrão do modelo da probabilidade do tipo de custos com as variáveis internas e as variáveis externas

Com a introdução das variáveis externas no modelo da probabilidade de ter um dos tipos de custo da utilização do Internamento, verifica-se que, das variáveis internas que se revelaram significativas no modelo só com as variáveis internas, apenas a Idade continua a ser significativa, tendo a Zona e as Garantias do Produto deixado de ser. Estas duas últimas variáveis internas foram substituídas por duas variáveis externas: o Valor médio da renda contratada por m2 e a Proporção de indivíduos sem atividade económica. O comportamento da Idade manteve-se, à medida que a idade aumenta diminui a probabilidade de ter um custo do tipo I e aumenta a probabilidade de ter um custo do tipo II ou do tipo III, e os resultados encontram-se no ANEXO 9. De seguida, são apresentados os resultados das variáveis externas que se revelaram significativas.

Valor médio da renda contratada por m2

	E(PX1|V)	E(PX2|V)	E(PX3|V)
Segurado Padrão	96,0%	3,8%	0,2%
Val_rent_contract_avg [0, 5]	96,5%	3,1%	0,4%
Val_rent_contract_avg]5, 8]	96,0%	3,8%	0,2%
Val_rent_contract_avg]8, 11]	94,7%	4,6%	0,7%
Val_rent_contract_avg >11	94,5%	4,8%	0,7%

Tabela 4.43 - Probabilidade do tipo de custos para o modelo com as variáveis internas e as variáveis externas por Valor médio da renda contratada por m2

O Valor médio da renda contratada por m2 é uma variável externa significativa para a probabilidade de ter um dos tipos de custo da utilização do Internamento, sendo que, para valores mais altos da variável (superiores a 8 euros), existe menos probabilidade de ter um custo do tipo I e mais probabilidade de ter um custo do tipo II ou do tipo III, face aos restantes valores de renda. Contudo, não é para valores mais baixos da renda ([0,5] euros) que a probabilidade de ter um custo do tipo I é maior e a probabilidade de ter um custo do tipo II ou III é menor. Na verdade, a classe dos]5,8] euros

é a que assume esses valores, mas a diferença entre estas duas classes não difere assim tanto: existe sim uma maior diferença entre estas duas classes e as últimas classes.

Proporção de indivíduos sem atividade económica

	E(PX1|V)	E(PX2|V)	E(PX3|V)
Segurado Padrão	96,0%	3,8%	0,2%
Prop_sem_ativ_econ <= 25%	94,8%	4,9%	0,3%
Prop_sem_ativ_econ 25% a 50%	96,0%	3,8%	0,2%
Prop_sem_ativ_econ > 50%	96,2%	3,6%	0,2%

Tabela 4.44 - Probabilidade do tipo de custos para o modelo com as variáveis internas e as variáveis externas por Proporção de indivíduos sem atividade económica

A proporção de indivíduos sem atividade económica é uma variável externa significativa para a probabilidade de ter um dos tipos de custo da utilização do Internamento, e pode ser analisada por dois níveis distintos: a proporção ser até 25% ou a proporção ser superior a 25%. Quando a proporção é até 25%, existe uma menor probabilidade de ter um custo do tipo I e maior probabilidade de ter custos do tipo II, ou seja, maior probabilidade de ter custo mais elevado. Quando numa zona existem algumas pessoas sem atividade económica, sendo a proporção superior a 25%, existe uma probabilidade maior de ter um custo do tipo I, ou seja, um custo mais baixo, e existe menor probabilidade de ter um custo do tipo II. Quanto à probabilidade de ter custos do tipo III, ou seja, custos muito elevados (superiores a 27.550 euros), ela não varia pela proporção de indivíduos sem atividade económica.

Os modelos construídos estão todos apresentados, de seguida será analisado o modelo do custo da utilização do Internamento que agrega as quatro componentes já explicadas anteriormente e, por último, o modelo de risco que resulta da junção da utilização do Internamento com o seu custo associado. O modelo da utilização e o modelo do custo da utilização do Internamento apresentado será com base nos modelos com as variáveis internas e as variáveis externas, isto porque, apesar do modelo do custo do tipo I ter identificado apenas variáveis internas, o modelo da utilização e o modelo da probabilidade do tipo de custos indicaram variáveis externas significativas.

Para o modelo do custo da utilização do Internamento, a Tabela 4.45 mostra as variáveis significativas e o impacto que estas têm no custo da utilização do Internamento. Das 29 variáveis introduzidas no modelo do custo da utilização do Internamento, apenas seis se revelaram significativas, sendo elas a Idade, o Género, a Zona, o Canal Comercial, o Valor médio da renda contratada por m2 e a Proporção de indivíduos sem atividade económica. A idade foi a única variável diferenciadora dos dois modelos: custos até 10.500 euros e probabilidade do tipo de custos. As restantes variáveis diferenciadoras como só são diferenciadoras numa componente do custo, nas outras componentes mantêm-se idênticas ao segurado padrão, assim o comportamento do custo final da utilização do Internamento terá o mesmo comportamento que nos modelos já apresentados. Relativamente à idade, quando se combina as quatro componentes do modelo, o custo continua a ser mais baixo dos seis aos 10 anos, sendo que depois tem tendência para ser crescente; contudo, na última classe, para idades superiores a 80 tem-se um custo inferior ao das classes dos 66 aos 80 anos, isto pode dever-se ao facto de que, nas idades mais avançadas, para além da doença em si, existem outros riscos na cirurgia que podem levar a que a decisão do Internamento não seja tomada, ou pelo menos não fazer uma cirurgia com tanto risco e

com custo elevado associado, por isso, o modelo indica que existe menor probabilidade de uma pessoa com mais de 80 anos ter um custo do tipo II ou do tipo III face a uma pessoa entre os 66 e os 80 anos.

	E(X1|V)	E(PX1|V)	E(X2|V)	E(PX2|V)	E(X3|V)	E(PX3|V)	E(X|V)
Segurado Padrão	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Idade [0,5]	2.142 €	99,9%	13.278 €	0,0%	36.510 €	0,1%	2.177 €
Idade [6,10]	2.142 €	100,0%	13.278 €	0,0%	36.510 €	0,0%	2.142 €
Idade [11,15]	2.142 €	95,9%	13.278 €	3,7%	36.510 €	0,4%	2.692 €
Idade [16,20]	3.026 €	98,4%	13.278 €	1,6%	36.510 €	0,0%	3.190 €
Idade [21,25]	3.026 €	99,7%	13.278 €	0,0%	36.510 €	0,3%	3.126 €
Idade [26,30]	3.026 €	97,9%	13.278 €	2,1%	36.510 €	0,0%	3.241 €
Idade [31,35]	3.026 €	96,6%	13.278 €	3,3%	36.510 €	0,1%	3.397 €
Idade [36,40]	3.026 €	96,0%	13.278 €	3,7%	36.510 €	0,3%	3.505 €
Idade [41,45]	3.313 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.758 €
Idade [46,50]	3.313 €	93,5%	13.278 €	6,0%	36.510 €	0,5%	4.077 €
Idade [51,55]	3.313 €	93,1%	13.278 €	6,4%	36.510 €	0,5%	4.116 €
Idade [56,60]	3.629 €	90,8%	13.278 €	8,2%	36.510 €	1,0%	4.749 €
Idade [61,65]	3.629 €	91,6%	13.278 €	7,4%	36.510 €	1,0%	4.672 €
Idade [66,70]	3.629 €	89,9%	13.278 €	9,4%	36.510 €	0,7%	4.767 €
Idade [71,75]	3.629 €	88,9%	13.278 €	10,1%	36.510 €	1,0%	4.933 €
Idade [76,80]	3.629 €	88,6%	13.278 €	10,1%	36.510 €	1,3%	5.031 €
Idade [81,85]	3.629 €	90,2%	13.278 €	9,4%	36.510 €	0,4%	4.668 €
Idade [86,90]	3.629 €	90,2%	13.278 €	9,4%	36.510 €	0,4%	4.668 €
Idade [91,95]	3.629 €	90,2%	13.278 €	9,4%	36.510 €	0,4%	4.668 €
Género F	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Género M	3.152 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.604 €
Zona ALGARVE	2.506 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	2.984 €
Zona BEIRA INTERIOR NORTE	2.506 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	2.984 €
Zona BEIRA LITORAL CENTRO	2.506 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	2.984 €
Zona MARÃO	2.506 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	2.984 €
Zona BEIRA LITORAL NORTE	2.757 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.224 €
Zona BEIRA LITORAL SUL	2.757 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.224 €
Zona GRANDE PORTO	2.757 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.224 €
Zona ALTO ALENTEJO	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona ALTO MINHO	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona BAIXO ALENTEJO	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona BAIXO MINHO	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona BEIRA INTERIOR SUL	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona GRANDE LISBOA	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona MARGEM SUL	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona OESTE	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona RIBATEJO	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Zona TRÁS-OS-MONTES	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Canal A	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Canal B	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Canal C	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Canal D	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Canal E	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Canal G	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Canal G	3.221 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.670 €
Canal F	3.221 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.670 €
Val_rent_contract_avg [0, 5]	3.026 €	96,5%	13.278 €	3,1%	36.510 €	0,4%	3.477 €
Val_rent_contract_avg]5, 8]	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Val_rent_contract_avg]8, 11]	3.026 €	94,7%	13.278 €	4,6%	36.510 €	0,7%	3.732 €
Val_rent_contract_avg >11	3.026 €	94,5%	13.278 €	4,8%	36.510 €	0,7%	3.752 €
Prop_sem_ativ_econ <= 25%	3.026 €	94,8%	13.278 €	4,9%	36.510 €	0,3%	3.628 €
Prop_sem_ativ_econ 25% a 50%	3.026 €	96,0%	13.278 €	3,8%	36.510 €	0,2%	3.482 €
Prop_sem_ativ_econ > 50%	3.026 €	96,2%	13.278 €	3,6%	36.510 €	0,2%	3.462 €

Tabela 4.45 - Modelo do custo da utilização do Internamento com as variáveis internas e as variáveis externas

Os resultados modelo do risco do Internamento construído estão apresentados na Tabela 4.46.

	E(N|V)	E(X|V)	E(R|V)
Segurado Padrão	3,5%	3.482 €	121 €
Idade [0,5]	4,8%	2.177 €	104 €
Idade [6,10]	2,3%	2.142 €	48 €
Idade [11,15]	2,3%	2.692 €	61 €
Idade [16,20]	3,5%	3.190 €	111 €
Idade [21,25]	3,5%	3.126 €	109 €
Idade [26,30]	3,5%	3.241 €	113 €
Idade [31,35]	3,5%	3.397 €	118 €
Idade [36,40]	4,7%	3.505 €	165 €
Idade [41,45]	4,7%	3.758 €	177 €
Idade [46,50]	6,1%	4.077 €	247 €
Idade [51,55]	7,5%	4.116 €	308 €
Idade [56,60]	7,5%	4.749 €	356 €
Idade [61,65]	10,9%	4.672 €	509 €
Idade [66,70]	15,3%	4.767 €	731 €
Idade [71,75]	15,3%	4.933 €	756 €
Idade [76,80]	20,0%	5.031 €	1.009 €
Idade > 81	20,0%	4.668 €	936 €
Género F	3,5%	3.482 €	121 €
Género M	3,5%	3.604 €	126 €
Zona BEIRA LITORAL CENTRO	1,7%	2.984 €	51 €
Zona ALGARVE	3,5%	2.984 €	104 €
Zona BEIRA INTERIOR NORTE	3,5%	2.984 €	104 €
Zona MARÃO	3,5%	2.984 €	104 €
Zona BEIRA LITORAL SUL	3,5%	3.224 €	112 €
Zona ALTO ALENTEJO	3,5%	3.482 €	121 €
Zona BAIXO ALENTEJO	3,5%	3.482 €	121 €
Zona BEIRA INTERIOR SUL	3,5%	3.482 €	121 €
Zona GRANDE LISBOA	3,5%	3.482 €	121 €
Zona MARGEM SUL	3,5%	3.482 €	121 €
Zona RIBATEJO	3,5%	3.482 €	121 €
Zona TRÁS-OS-MONTES	3,5%	3.482 €	121 €
Zona BEIRA LITORAL NORTE	3,9%	3.224 €	126 €
Zona GRANDE PORTO	3,9%	3.224 €	126 €
Zona ALTO MINHO	3,9%	3.482 €	136 €
Zona BAIXO MINHO	3,9%	3.482 €	136 €
Zona OESTE	3,9%	3.482 €	136 €

	E(N|V)	E(X|V)	E(R|V)
Segurado Padrão	3,5%	3.482 €	121 €
Garantias Produto I	1,6%	3.482 €	54 €
Garantias Produto I + Acons	2,2%	3.482 €	78 €
Garantias Produto I + A	3,5%	3.482 €	121 €
Garantias Produto I + A + PO	3,5%	3.482 €	121 €
Garantias Produto I + A + E	4,1%	3.482 €	141 €
Garantias Produto I + A + E + PO	4,1%	3.482 €	141 €
Canal A	2,6%	3.482 €	89 €
Canal H	2,6%	3.670 €	94 €
Canal C	3,5%	3.482 €	121 €
Canal D	3,5%	3.482 €	121 €
Canal F	3,5%	3.670 €	128 €
Canal B	3,8%	3.482 €	131 €
Canal E	3,8%	3.482 €	131 €
Canal G	3,8%	3.482 €	131 €
Prop_1_2_div <= 25%	3,5%	3.482 €	121 €
Prop_1_2_div > 25%	5,6%	3.482 €	196 €
Val_rent_contract_avg [0, 5]	3,5%	3.477 €	121 €
Val_rent_contract_avg]5, 8]	3,5%	3.482 €	121 €
Val_rent_contract_avg]8, 11]	3,5%	3.732 €	130 €
Val_rent_contract_avg >11	3,5%	3.752 €	131 €
Prop_sem_ativ_econ <= 25%	3,5%	3.628 €	126 €
Prop_sem_ativ_econ 25% a 50%	3,5%	3.482 €	121 €
Prop_sem_ativ_econ > 50%	3,5%	3.462 €	121 €

Tabela 4.46 - Modelo de risco do Internamento com as variáveis internas e as variáveis externas

O modelo do risco do Internamento é composto pela taxa de utilização e pelo custo (desagregado em quatro componentes), as variáveis diferenciadoras do risco podem ser apenas numa dessas componentes, mas, no fim, interessa avaliar o valor final do risco.

As variáveis que são diferenciadoras da utilização e do custo do Internamento são a Idade, a Zona e o Canal Comercial. As Garantias do Produto e a Proporção de habitações com 1 a 2 divisões nos últimos dois anos apenas são diferenciadoras na utilização do Internamento. O Género, o Valor de renda contratada média por m2 e a Proporção de indivíduos sem atividade económica são variáveis diferenciadoras apenas do custo da utilização do Internamento.

Em termos de comportamento do risco, apenas será descrito para as variáveis que são diferenciadoras da utilização e do custo do Internamento, porque, para as variáveis que diferenciam apenas uma das componentes, o comportamento do risco já foi apresentado.

Idade

No caso da Idade, uma pessoa nos primeiros anos de vida tem mais risco do que dos seis aos 15 anos, quer a utilização do Internamento é menor, quer o custo associado é menor. Entre os 16 e os 30 anos o risco não varia muito, porque a taxa de utilização é idêntica para estas idades. A partir dos 30 anos o risco começa a crescer conforme a idade, tanto na taxa de utilização como no custo médio, até atingir a classe dos 80 anos. Para idades superiores a 80, o risco desceu um pouco face à classe dos 76 aos 80 anos porque, apesar de terem taxas de utilização idênticas, o custo associado é menor.

Zona

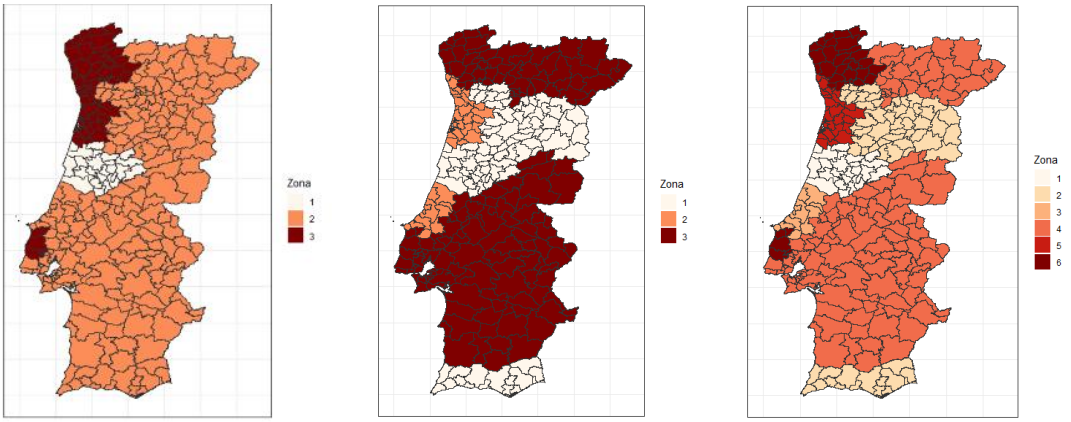


Figura 4.8 - Mapa com identificação das classes de risco do modelo da utilização, do custo e do risco do Internamento, respetivamente

No que diz respeito à Zona, como pode ser visto na Figura 4.8, as várias regiões não se distribuem da mesma forma pela utilização e custo do Internamento, apesar de estarem ambas divididas em três classes. Daqui surge a classificação do risco da variável por seis classes. As únicas zonas que têm igual classe de risco na utilização e no custo do Internamento são quatro de um total de dezassete zonas, a Beira Litoral Centro com menos risco, a Beira Litoral Sul com risco intermédio e o Alto Minho e o Oeste na classe com mais risco, estas zonas correspondes às zonas 1, 3 e 6 do modelo do risco, respetivamente. A zona 2 de risco é composta pelo Algarve, pela Beira Interior Norte e pelo Marão, que estão no nível mais baixo de custo, mas estão no nível intermédio de utilização. A zona 4 contém a maior parte das zonas, nomeadamente a área metropolitana de Lisboa, o Alentejo, o Ribatejo, a Beira Interior Sul e Trás-os-Montes, que têm nível intermédio de utilização e nível alto de custo do Internamento. Por fim, a zona 5 de risco é composta pela Beira Litoral Norte e pelo Grande Porto; estão no nível mais alto de utilização, mas estão no nível intermédio de custo.

Canal Comercial

Quanto ao Canal Comercial, o risco é menor no Canal A, tanto na utilização como no custo do Internamento face aos outros canais de venda. O Canal C e D aparentam ter o mesmo risco e os Canais Comerciais que aparentam ter mais risco são o B, o E e o G.

5. CONCLUSÃO

O seguro serve para proteger as pessoas do risco. Do lado das Seguradoras, surge a necessidade de conseguir quantificar o risco assumido, para conseguir cumprir com as suas obrigações com os segurados e garantir a sua sustentabilidade. Sendo o risco volátil, mesmo que se consiga obter uma estimativa do seu valor, existe sempre uma probabilidade de o comportamento não ser o esperado. Devido a esta incerteza, deve-se tentar conhecer ao máximo as causas (variáveis explicativas) do risco para se conseguir fazer uma melhor mensuração do mesmo.

Sendo o objetivo do estudo identificar as variáveis significativas do risco da cobertura de Internamento para seguros de saúde individuais, recolheram-se variáveis relacionadas com as características das pessoas, com as características do seguro que compraram e variáveis de georreferenciação relacionadas com as casas, o poder de compra, a escolaridade, a atividade económica e a saúde. Neste sentido, e porque o segundo tipo de variáveis é o de variáveis externas, que atualmente não são recolhidas pela Seguradora, decidiu-se construir primeiro um modelo com as variáveis internas, e depois outro modelo, onde se adicionaram as variáveis externas.

Os resultados mostraram que existem mais variáveis internas a explicar o risco do que variáveis externas, das que foram incluídas neste estudo. Além disso, as variáveis internas tendem a explicar mais o risco do que as variáveis externas. Importa salientar que as variáveis externas não estão ao mesmo nível que as variáveis internas, isto porque são variáveis de georreferenciação que sofreram ainda alguns tratamentos de dados antes de serem introduzidas nos modelos. Enquanto que as variáveis internas estão ao nível da pessoa, as variáveis externas referem-se à zona onde a pessoa mora, o que pode levar a enviesamentos, porque o comportamento da zona pode não representar o comportamento da pessoa que mora nessa zona.

No que diz respeito à mensuração do risco de Internamento, esta foi feita através de modelos lineares generalizados onde se construiu um modelo para a utilização, e outro para o custo da utilização do Internamento. As variáveis que são diferenciadoras da utilização do Internamento, para um nível de significância de 5%, são a Idade, as Garantias do Produto, o Canal Comercial, a Zona e a Proporção de habitações com 1 a 2 divisões, por ordem decrescente de importância. Para avaliar os custos da utilização do Internamento, foi necessário partir o modelo em quatro partes sendo elas os custos até 10.500 euros, os custos entre 10.500 euros e 27.550 euros, os custos superiores a 27.550 euros e a probabilidade de ter um destes tipos de custo. Para a primeira parte, os custos até 10.500 euros, construiu-se um modelo e as variáveis que se revelaram significativas foram apenas variáveis internas: a Idade, a Zona, o Género e o Canal Comercial, por ordem decrescente de importância. Para os custos entre 10.500 euros e 27.550 euros e para os custos superiores a 27.550, ajustaram-se distribuições Pareto Generalizadas, que são distribuições de valores extremos, e considerou-se o valor médio das mesmas. Na probabilidade de uma pessoa ter um dos tipos de custo, a Idade revelou-se significativa. Este foi o único modelo onde algumas variáveis externas substituíram as variáveis internas. No modelo sem as variáveis externas, as Garantias do Produto e a Zona eram diferenciadoras, mas, com a introdução das variáveis Valor médio da renda contratada por m2 e Proporção de indivíduos sem atividade económica, deixaram de o ser.

As principais conclusões sobre o comportamento do risco das variáveis significativas são que o risco aumenta com a idade, exceto nas idade entre os seis e os dez anos; os homens têm, em média, mais risco de internamento do que as mulheres; a zona com menor risco é a Beira Litoral Centro e as zonas com maior risco são o Minho e o Oeste; ter um produto com mais coberturas tem mais risco do que ter um produto mais simples; o canal de vendas com menos risco é o A e com mais risco são o B, E e G; se a proporção de habitações com 1 a 2 divisões de uma zona for superior a 25%, então o risco é maior do que se a proporção for inferior a 25%; se o valor da renda média contratada por m2 for superior a 8 euros, o risco é maior do que para valores mais baixos; para uma zona em que a proporção de indivíduos sem atividade económica é inferior a 25%, o risco é maior do que para uma proporção superior a 25%. Devido a problemas de correlação entre variáveis explicativas e de homogeneidade da amostra, não foi possível analisar o comportamento do risco das variáveis: Antiguidade, Capital, Taxa de Mulheres com mamografia realizada nos últimos 2 anos.

De um modo geral, foram encontradas mais variáveis diferenciadoras para a utilização do Internamento do que para o custo do Internamento. Importa referir que o Internamento é das coberturas com mais risco dentro do seguro de saúde, apresentando uma grande variabilidade de custos, que se torna difícil de explicar apenas através das variáveis internas e das variáveis externas utilizadas neste estudo. Além do mais, é importante não esquecer que o seguro de saúde em Portugal é complementar/suplementar ao SNS, por isso, nem todos os Internamentos que as pessoas façam estão contidos na amostra. Apesar de as pessoas procurarem ter um seguro de saúde para um serviço mais rápido e cómodo, existem cuidados médicos que apenas são tratados no prestador público. Assim, quando se avalia o risco do Internamento nos seguros de saúde deve-se ter presente que existem outros fatores como a qualidade de serviço e o tempo de resposta do SNS, entre outros, que nem sempre são fáceis de identificar ou mensurar, mas que podem ajudar a explicar a variabilidade do risco.

Não obstante a vontade de encontrar as variáveis diferenciadoras do risco, deve-se ter em especial atenção as restrições legais. O seguro de saúde lida com informação sensível e tem regras específicas na utilização de dados clínicos na diferenciação do risco. Neste estudo, o objetivo era conhecer melhor o risco e não tarifar com base nas variáveis que se revelaram significativas; mas se a Seguradora achar que alguma variável explicativa deve passar a ser considerada na tarifa, então deve assegurar-se que está a cumprir com o Regime Jurídico do Contrato de Seguro.

Em suma, conhecer o risco é essencial numa Seguradora. Apesar da sua aleatoriedade, existem fatores que o podem influenciar, e identificar as suas causas torna-se fundamental. Com o avanço das tecnologias, é possível a utilização de informação cada vez de forma mais fácil e rápida. Sendo o mercado segurador bastante competitivo, torna-se essencial utilizar todos os recursos disponíveis, para, por um lado, oferecer a proteção que as pessoas precisam por preço razoável, e, por outro, garantir que o que está a ser cobrado é adequado para o futuro da Seguradora.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Este trabalho, com as variáveis externas, pretendeu analisar informação das pessoas para além daquela que é atualmente disponibilizada. Contudo estas variáveis tinham algumas condicionantes que podem ter influenciado as conclusões, no que diz respeito a: enviesamento entre o comportamento da zona onde a pessoa mora e o seu comportamento real; diferença entre o período dos dados da amostra e a data das informações externas. Além disso, importa relembrar a importância da recolha dos dados e a qualidade dos mesmos. Nas variáveis internas a informação estava toda disponível sem valores omissos; no caso das variáveis externas, quase cerca de metade da amostra teve de ser excluída por ter informação em falta. Sendo a dimensão da amostra um fator tão importante para garantir conclusões robustas, é fundamental mitigar as perdas de informação por uma má recolha de dados.

Embora os métodos de seleção de variáveis utilizados tenham dado resultados semelhantes na maior parte dos modelos, pode ser interessante abordar outro tipo de métodos mais automatizados, percebendo se os resultados são consistentes e qual o ganho face a métodos mais tradicionais. Ainda no âmbito da metodologia, salienta-se a limitação de usar variáveis não correlacionadas; apesar de se ter feito algumas análises e limpezas dessas relações, não é possível garantir que as correlações tenham sido todas eliminadas, e isso pode condicionar as conclusões. Esta limitação advém de se ter utilizado Modelos Lineares Generalizados. Além desta restrição, estes modelos pressupõem que a relação entre a variável resposta e a variável explicativa é linear, o que pode não ser verdade. Para considerar relações não lineares, podem-se utilizar os modelos aditivos generalizados (MAGs); no caso de haver muitas variáveis correlacionadas, pode-se optar pelos modelos lineares mistos (MLMs) generalizados que consideram efeitos aleatórios no preditor linear ou as equações de estimação generalizadas (EEG), que permitem a correlação sem explicar a sua origem.

Um outro aspeto que deverá ser igualmente desenvolvido é a procura e recolha de mais variáveis. Na revisão da literatura, foram apresentados os fatores de risco que mais contribuíram para a morte dos portugueses: um quarto dessas causas foram fatores de risco comportamentais, ou seja, que podem ser mudados. Deste modo, seria interessante conseguir estudar este tipo de variáveis e perceber se são realmente diferenciadoras do risco em saúde. Com a evolução da tecnologia, a recolha de dados deste tipo está cada vez mais a ser feita, através da monotorização da saúde e hábitos de vida da pessoa. Com estes dados, a Seguradora conseguirá conhecer melhor o risco, mas também poderá ajudar a consciencializar melhor as pessoas do risco assumido, levando-as a adotarem estilos de vida mais saudáveis, o que por outro lado se traduz em menores custos para a Seguradora.

7. BIBLIOGRAFIA

Abrantes, A. (1985). QUEM PAGA OS CUIDADOS DE SAÚDE? *Acta Médica Portuguesa*, 165-169.

Abreu, M. (2012). *Taxas Moderadoras e a Racionalização da Procura de Cuidados de Saúde.* Lisboa: Trabalho Projeto de Mestrado, ENSP UNL.

Agresti, A. (1996). An Introduction to Categorical Data Analysis. *Open Journal of Statistic*.

Agresti, A. (2013). *Categorical Data Analysis.* John Wiley & SONS, Inc.

APS. (2009). OS SEGUROS DE SAÚDE PRIVADOS NO CONTEXTO DO SISTEMA DE SAÚDE PORTUGUÊS.

APS. (2021). Curso PDEAD de agente de seguros.

APS. (2021). *INVESTIMENTOS SETOR SEGURADOR.* Obtido de https://segurdata.apseguradores.pt/ords/f?p=100:0:17397358452738:APPLICATION_PROCESS%3DDOWNLOAD_FILE:::APP_FILE_ID,APP_FILE_ID_CHECK:100926,77

APS. (2021). *PRODUÇÃO DE SEGURO DIRETO.* Obtido de https://segurdata.apseguradores.pt/ords/f?p=100:0:17397358452738:APPLICATION_PROCESS%3DDOWNLOAD_FILE:::APP_FILE_ID,APP_FILE_ID_CHECK:100926,77

ASF. (2021). DESAFIOS PARA A SUPERVISÃO E REGULAÇÃO DOS SEGUROS DE SAÚDE EM PORTUGAL.

Assembleia da República. (2005). Constituição da República Portuguesa. (V. R. Constitucional, Ed.)

Bandeira, M. d. (2013). *SEGURO DE SAÚDE: CUSTOS DE AMBULATÓRIO - MODELIZAÇÃO LINEAR GENERALIZADA.* Obtido de Tese de Mestrado, FCUL UNL, Lisboa.

Baptista, F. (2019). *A realidade portuguesa dos seguros de saúde.* Obtido de Tese de Mestrado, FCUL UNL, Lisboa.

Barber, J., & Thompson, S. (2000). Analysis of cost data in randomized trials: an application of the non-parametricbootstrap. *Statistic in Medicine*, 3219-3236.

Barry, L., & Charpentier, A. (2020). Personalization as a promise: Can BigData change the practice of insurance?

Basu, A., & Rathouz , P. (2005). Estimating marginal and incremental effects on health outcomes using flexible link andvariance function models. *Biostatistic*, 93-109.

Cantoni, E., & Ronchetti, E. (2001). Robust inference for generalized linear models. *American Statistical Association*, 1022-1030.

De Witt, G., & Van Eeghen, J. (1984). Rate making and society's sense of fairness. *ASTIN Bulletin*, 151 - 164.

Deb, P., & Holmes, A. (2000). Estimates of use and costs of behavioural health care: a comparison of standard andfinite mixture models. *Health Economics*, 475-489.

Deng, H. a. (2011). Bias of importance measures for multi-valued attributes and solutions. *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)*, (pp. 293 - 300).

Duchêne, S., & Boyer-Kassem, T. (2020). On discrimination in health insurance.

Foucault, M. (2009). Security, Territory, Population. Basingstoke.

Guerra, A. (2014). *Factores Influenciadores da Aquisição de Seguro de Saúde Voluntário em Portugal: Validação Interna de um Questionário.* Obtido de CEAH.

Guerreiro, G. R. (2016). Manual de Construção de Tarifas com R – O Exemplo do Seguro Automóvel –.

Guiomar, J. (2010). OS SEGUROS DE SAÚDE VOLUNTÁRIOS – O perfil dos utilizadores e determinantes da procura.

Haynes, G., & Dunnagan, T. (2002). Comparing Changes in Health Risk Factors andMedical Costs Over Time. *American Journal of Health Promotion*.

Hosmer, D., Lemeshow, S., & Rodney, X. (2013). *Applied Logistic Regression* (Vol. 3rd Ed). Wiley.

Johnson, P. E. (2014). GLM with a Gamma-distributed Dependent Variable.

Johnson, R. a. (2008). Applied Multivariate Statistical Analysis. Pearson.

Knight, F. (1985). Risk, Uncertainty and Profit. Chicago.

Landes, X. (2015). How fair is actuarial fairness? *Journal of Business Ethics*, 519 - 533.

Leo Breiman, J. F. (1984). *Classification and Regression Trees.* Routledge Taylor & Francis Group.

Leonard, K., & Rousseeuw, P. J. (1990). Finding Groups in Data ,An Introduction to Cluster Analysis. *Wiley Inter-science.* Canadá.

Lewis, R. (2000). An introduction to classification and regression tree (CART)) analysis. *Annual Meeting of the Society for Academy Emergency Medicine.* San Francisco, California, USA.

Lindsey, J. (1997). Applying Generalized Linear Models. Heidelberg: Springer-Verlag.

Liukko, J. (2010). Genetic discrimination, insurance, and solidar-ity: An analysis of the argumentation for fair risk classifica-tion. . *New Genetics and Society*, 457 - 475.

Maimon, L. R. (2005). Top- Down Induction of Decision Trees Classifiers -A Survey. Em *IEE Transactions on Systems.*

Manning , W., & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Health Economics*.

Maroco, J. (2003). Análise Estatística com utilização do SPSS. Lisboa: Edições Sílabo.

Maroco, J. (2018). Análise Estatística Com o SPSS Statistics.

Martin Leo, S. S. (2019). Machine Learning in Banking Risk Management: A Literature Review. *Risk*.

Martinho, A. (2014). Evolução do seguro de saúde em Portugal nos últimos 15 anos. Tomar.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & all.

McFall, L. (2019). Supervised Learning Methodsfor Predicting Healthcare Costs: Systematic LiteratureReview and Empirical Evaluation. *Economy and Society*.

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 897-916.

Mulholland, K., E., L., S., Carneiro, I., Becher, H., & Lechman, D. (2008). Equity and child-survival strategies. *Bulletin of the World Health Organization, 86*, 399-407.

Mullahy. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Applied Econometrics*, 337-350.

O'Hagan , A., & Stevens , J. (2003). Assessing and comparing costs: how robust are the bootstrap and methods based onasymptotic normality? *Health Economics*, 33-49.

Pickands, J. (1975). Statistical Inference Using Extreme Order Statistic. *Annals of Statistic*, 119-131.

Rego, G. (2008). *Gestão Empresarial dos Serviços Públicos.* Vida Económica.

Rosko, M., & R. Broyles. (1988). The Demand for Medical Services. Em M. Rosko, & R. Broyles, *The Economics of Health Care: A Reference Handbook* (pp. 44-77). New York: Greenwood Press.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal*, Vol. 3, No. 3.

Santos, J. M. (2003). Aplicação da Teoria de Valores Extremos à Actividade Seguradora.

Saúde, C. p. (2007). *A Sustentabilidade Financeira do Serviço Nacional de Saúde.*

Shapiro, M. J. (1993). *Employment and Health Benefits: A Connection at Risk.* Committee on Employment-Based Health Benefits, Institute of Medicine .

Silva, S. (2009). *Os Seguros de Saúde Privados no Contexto do Sistema de Saúde Português.*

Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, 99-114.

Turkman, M. a. (2000). Modelos Lineares Generalizados - da teoria à prática. Lisboa: Edições SPE.

Vintém, J. (2008). Inquéritos Nacionais de Saúde:auto-percepção do estado de saúde: uma análise em torno da questão de géneroe da escolaridade.

Walters, M. (1981). Risk classification standards. *Proceedings of the Casualty Actuarial Society*, 68, 1–23.

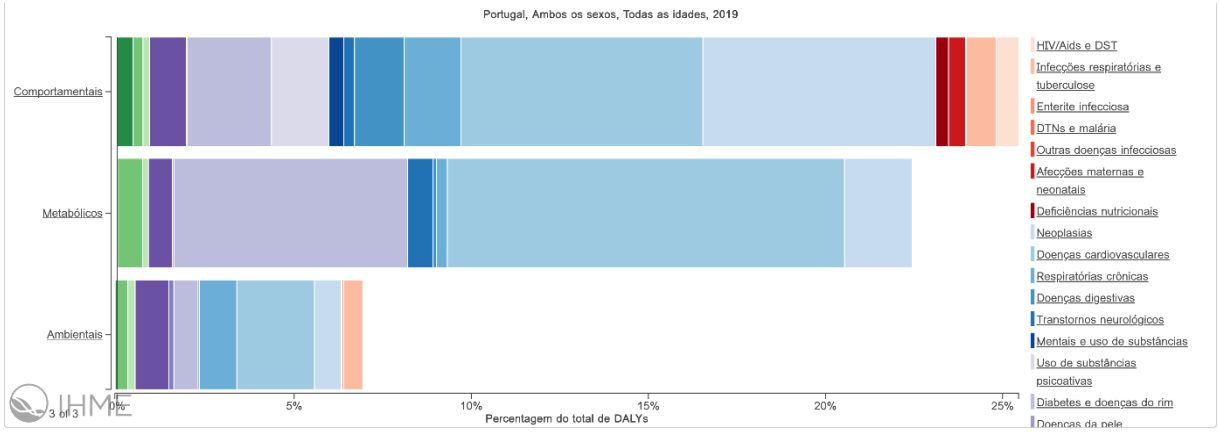
Wedderburn, J. N. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 370-384.

Zhuang, G. (2013). Neural Network Model of Pricing Health Care Insurance.

8. ANEXOS

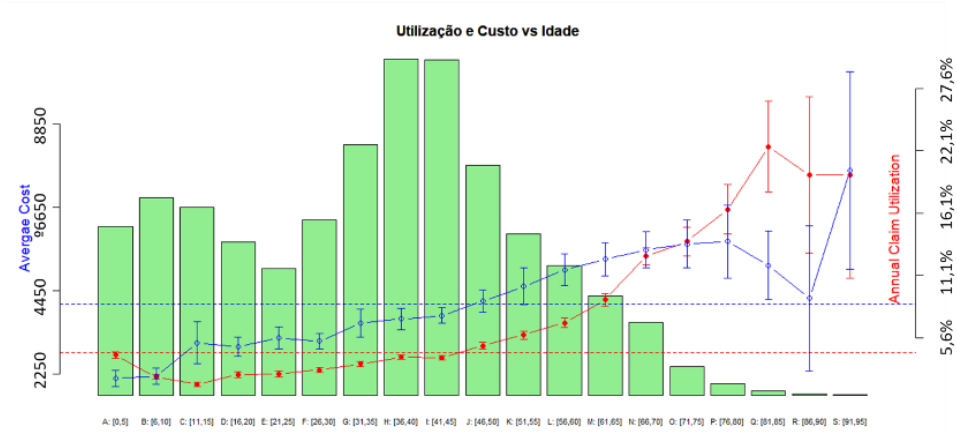
ÍNDICE DE ANEXOS

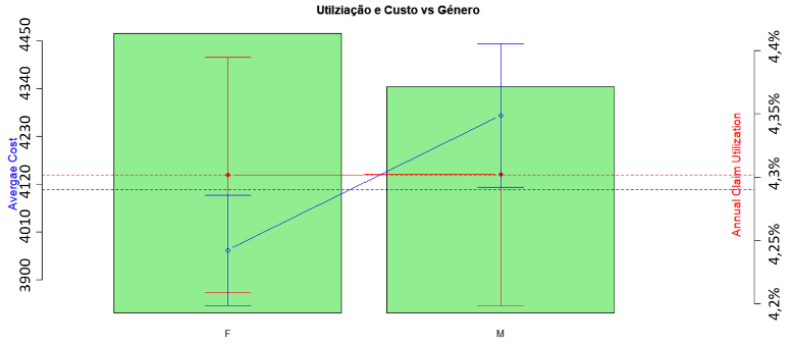
ANEXO 1 - Categoria dos fatores de risco ordenados por peso no número total de DAYLs para Portugal, em 2019

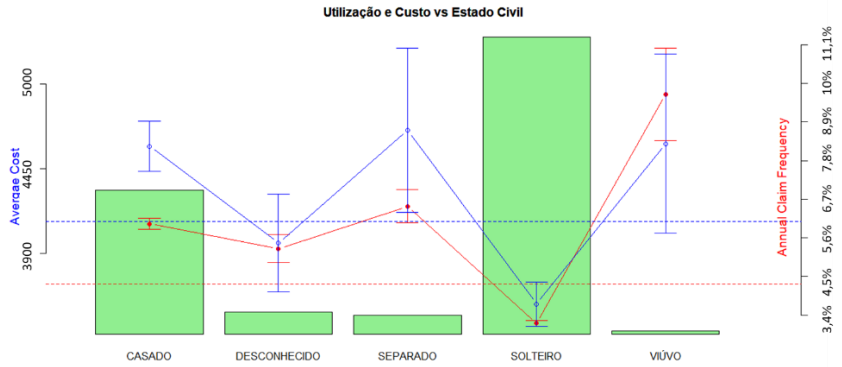


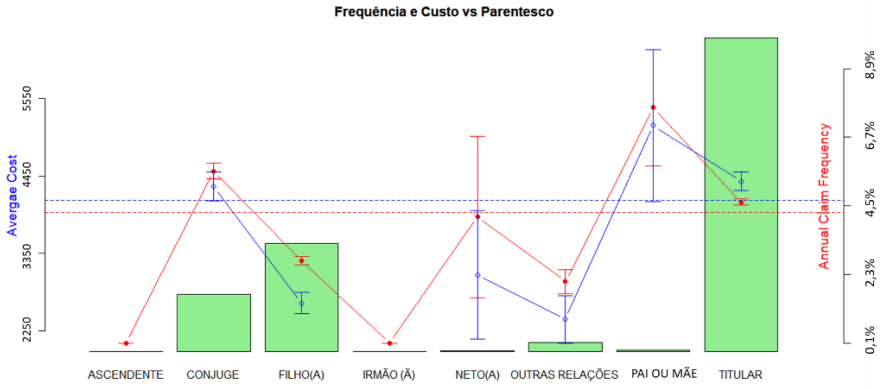
Fonte: Elaborado com base nos dados para Portugal da GBD 2019

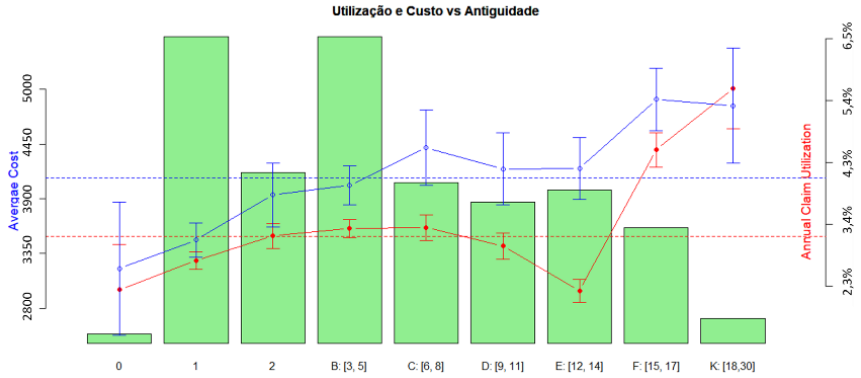
ANEXO 2 - Comportamento do risco e quantidade de pessoas das variáveis internas

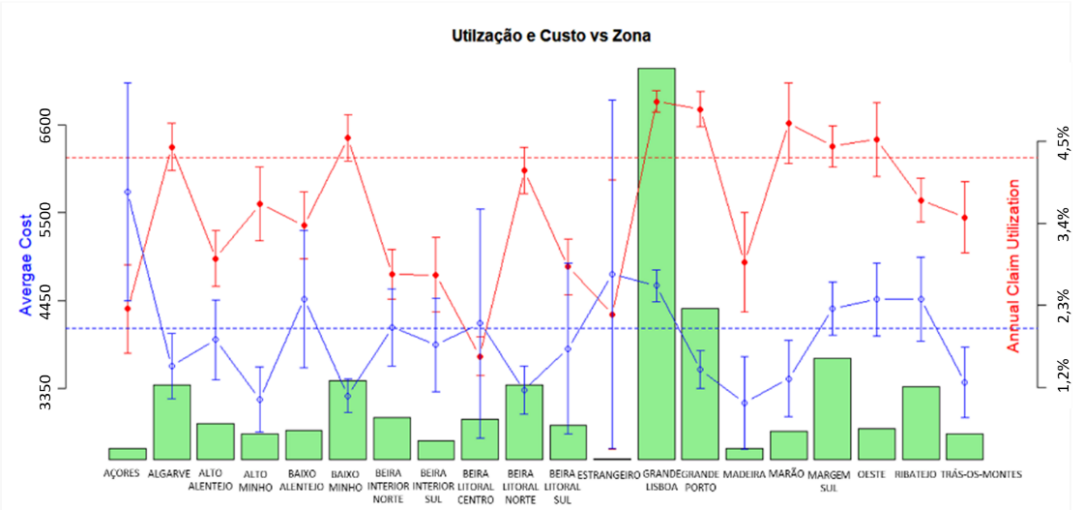


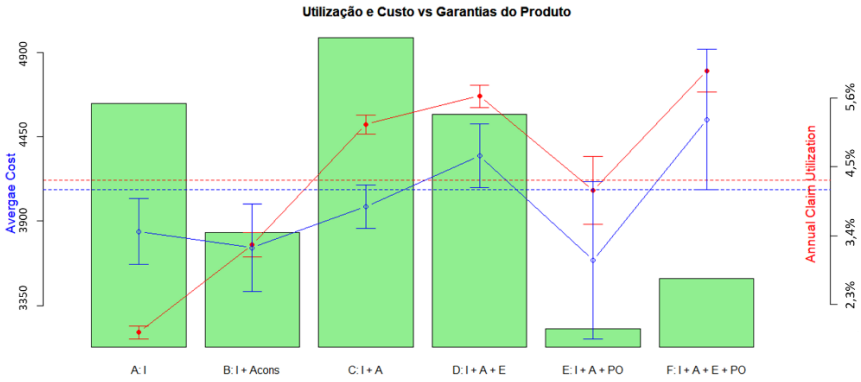


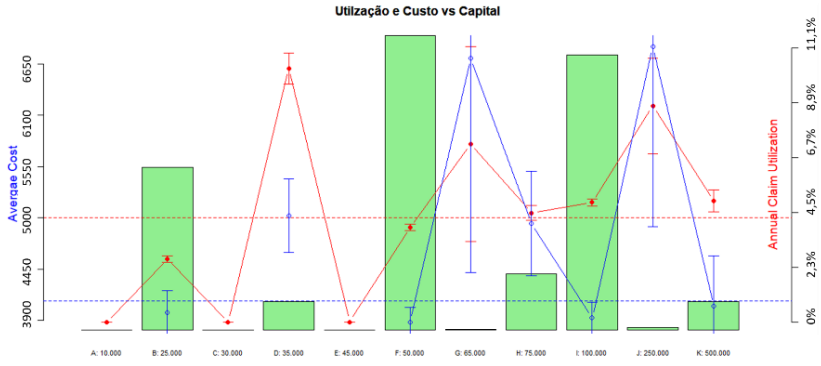


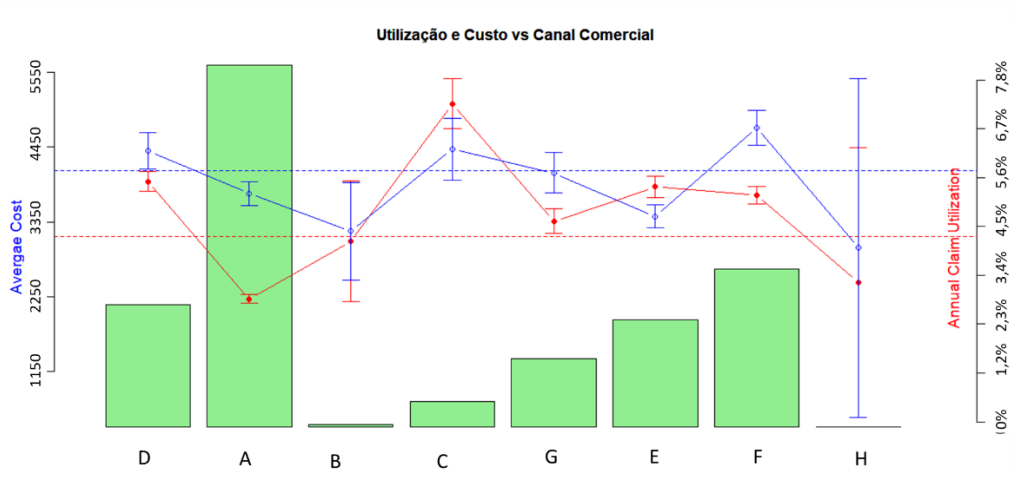


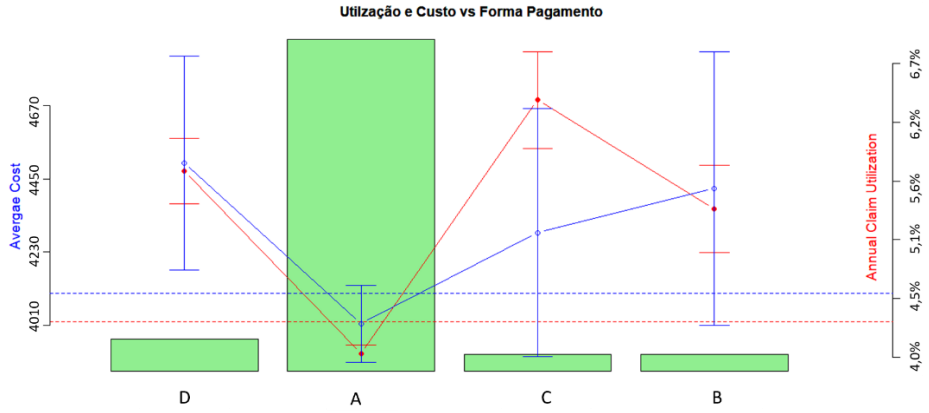

Utilização e Custo vs Antiguidade
Avergae Cost
Annual Claim Utilization
0
1
2
B: [3, 5]
C: [6, 8]
D: [9, 11]
E: [12, 14]
F: [15, 17]
K: [18,30]

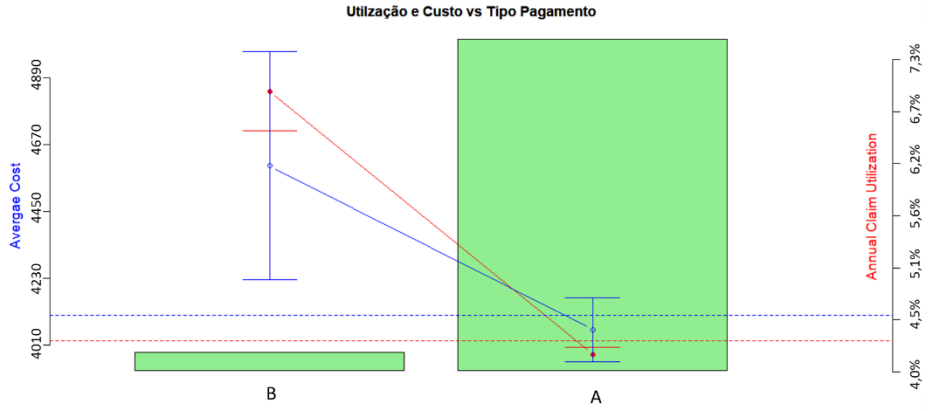

Utilização e Custo vs Zona
Avergae Cost
Annual Claim Utilization
AÇORES
ALGARVE
ALTO ALENTEJO
ALTO MINHO
BAIXO ALENTEJO
BAIXO MINHO
BEIRA INTERIOR NORTE
BEIRA INTERIOR SUL
BEIRA LITORAL CENTRO
BEIRA LITORAL NORTE
BEIRA LITORAL SUL
ESTRANGEIRO
GRANDE LISBOA
GRANDE PORTO
MADEIRA
MARÃO
MARGEM SUL
OESTE
RIBATEJO
TRÁS-OS-MONTES


Utilização e Custo vs Garantias do Produto
Avergae Cost
Annual Claim Utilization
A: I
B: I + Acons
C: I + A
D: I + A + E
E: I + A + PO
F: I + A + E + PO


Utilização e Custo vs Capital
Averaae Cost
Annual Claim Utilization
A: 10.000
B: 25.000
C: 30.000
D: 35.000
E: 45.000
F: 50.000
G: 65.000
H: 75.000
I: 100.000
J: 250.000
K: 500.000


Utilização e Custo vs Canal Comercial
Avergae Cost
1150
2250
3350
4450
5550
Annual Claim Utilization
0%
1,2%
2,3%
3,4%
4,5%
5,6%
6,7%
7,8%
D
A
B
C
G
E
F
H


Utilização e Custo vs Forma Pagamento
Avergae Cost
4010
4230
4450
4670
Annual Claim Utilization
4,0%
4,5%
5,1%
5,6%
6,2%
6,7%
D
A
C
B


Utilização e Custo vs Tipo Pagamento
Avergae Cost
4010
4230
4450
4670
4890
Annual Claim Utilization
4,0%
4,5%
5,1%
5,6%
6,2%
6,7%
7,3%
B
A

ANEXO 3 - Análise de Correlação entre as variáveis significativas

Utilização

Medição da associação entre as variáveis		
Idade	Estado Civil	0,56
Idade	Parentesco	0,68
Idade	Antiguidade	0,34
Idade	Zona	0,07
Idade	Garantias do Produto	0,08
Idade	Capital	0,1
Idade	Canal Comercial	0,07
Idade	Forma de Pagamento	0,11
Idade	Tx_mulheres_mam	0,1
Idade	Prop_1_2_div	0,01
Antiguidade	Zona	0,06
Antiguidade	Garantias do Produto	0,34
Antiguidade	Capital	0,18
Antiguidade	Canal Comercial	0,16
Antiguidade	Forma de Pagamento	0,12
Antiguidade	Tx_mulheres_mam	0,09
Antiguidade	Prop_1_2_div	0,01
Capital	Zona	0,13
Capital	Garantias do Produto	0,64
Capital	Canal Comercial	0,2
Capital	Forma de Pagamento	0,04
Capital	Tx_mulheres_mam	0,15
Capital	Prop_1_2_div	0,01
Zona	Garantias do Produto	0,11
Zona	Canal Comercial	0,26
Zona	Forma de Pagamento	0,02
Zona	Tx_mulheres_mam	0,73
Zona	Prop_1_2_div	0,01
Garantias do Produto	Canal Comercial	0,14
Garantias do Produto	Forma de Pagamento	0,04
Garantias do Produto	Prop_1_2_div	0,02
Canal Comercial	Forma de Pagamento	0,16
Canal Comercial	Prop_1_2_div	0,04
Forma de Pagamento	Prop_1_2_div	0,01

Custos do Tipo I

Medição da associação entre as variáveis		
Idade	Género	0,03
Idade	Zona	0,15
Idade	Canal Comercial	0,16
Género	Zona	0,02
Género	Canal Comercial	0,03
Zona	Canal Comercial	0,28

Probabilidade de ter um dos tipos de custo

Medição da associação entre as variáveis		
Idade	Garantias do Produto	0,03
Idade	Zona	0,15
Garantias do Produto	Zona	0,16
Idade	Val_rent_contract_avg	0,02
Idade	Prop_sem_ativ_econ	0,03
Val_rent_contract_avg	Prop_sem_ativ_econ	0,28

ANEXO 4 - Agrupamento inicial da Zona de residência

Utilização

Zona	
1	Alto Minho, Algarve
2	Grande Porto
3	Beira Litoral Norte
4	Beira Litoral Centro, Marão, Baixo Minho, Beira Litoral Sul
5	Trás-os-Montes, Madeira, Baixo Alentejo, Beira Interior Norte
6	Ribatejo, Alto Alentejo
7	Margem Sul
8	Grande Lisboa
9	Beira Interior Sul, Oeste, Açores, Estrangeiro

Custo do Tipo I

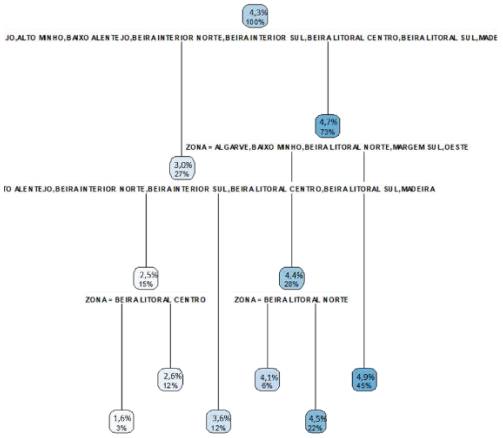
Zona	
1	Açores, Beira Interior Norte, Grande Lisboa, Baixo Alentejo
2	Margem Sul
3	Ribatejo
4	Oeste, Alto Alentejo
5	Beira Litoral Centro, Algarve
6	Grande Porto
7	Beira Litoral Sul, Beira Interior Sul, Alto Minho, Marão
8	Beira Litoral Norte
9	Trás-os-Montes, Baixo Minho, Madeira, Estrangeiro

ANEXO 5 - Resultados Análise de *Clusters* e Árvores de Decisão para a Utilização

Zona

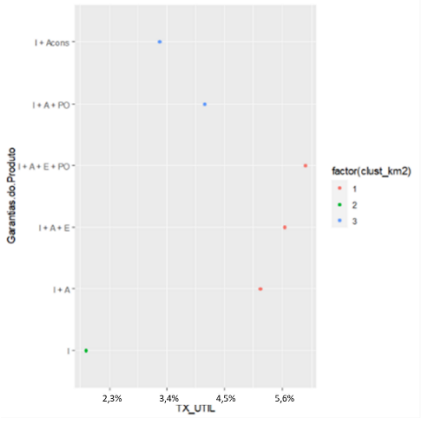
Análise de *Clusters*

Árvore de decisão

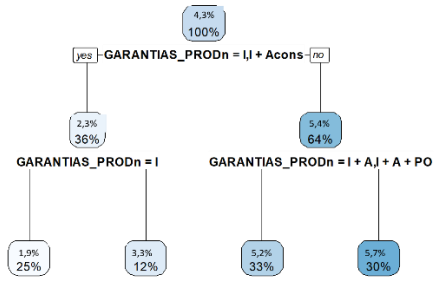


Garantias do Produto

Análise de *Clusters*

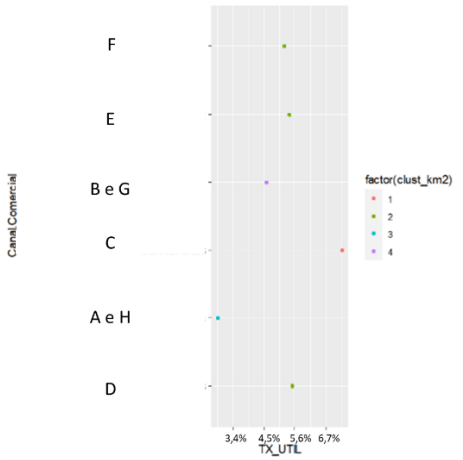


Árvore de decisão

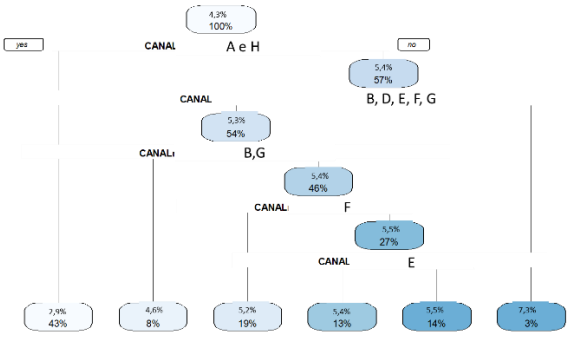


Canal Comercial

Análise de *Clusters*



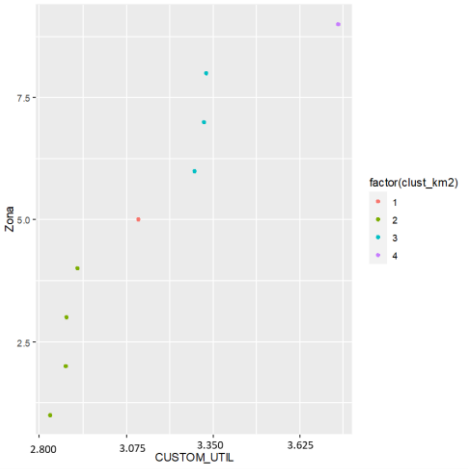
Árvore de decisão



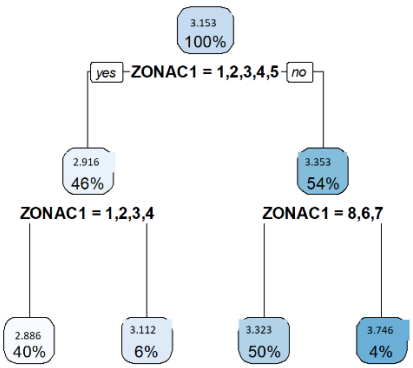
ANEXO 6 - Resultados Análise de *Clusters* e Árvores de Decisão para o Custo do tipo I

Zona

Análise de *Clusters*



Árvore de decisão



ANEXO 7 - Variáveis introduzidas no modelo com as variáveis internas e as variáveis externas

Utilização

Variáveis Internas	Variáveis Externas	
Idade	Valor de renda contratada média/m2	Prop. de pessoas com o ensino superior completo
Género	Prop. de habitações com área até 50 m2	Prop. de pessoas empregados no setor primário
Zona	Prop. de habitações com área entre 50 m2 e 100 m2	Prop. de pessoas empregados no setor secundário
Antiguidade	Prop. de habitações com área entre 100 m2 e 200 m2	Prop. de pessoas pensionistas ou reformados
Garantias do Produto	Prop. de habitações com área superior a 200 m2	Tx. utilização global de consultas médias em um ano
Capital	Prop. de habitações com 1 a 2 divisões	Tx. de urgências
Canal Comercial	Prop. de habitações com 3 a 4 divisões	Tx. de consultas
Forma de Pagamento	Prop. de Analfabetos	Tx. de primeiras consultas realizadas em tempo adequado
Tipo de Pagamento	Prop. de pessoas com o ensino secundário completo	Tx. de mulheres com mamografia realizada nos últimos dois anos
	Prop. de pessoas com o ensino pós-secundário completo	

Custos do Tipo I

Variáveis Internas	Variáveis Externas	
Idade	Valor de renda contratada média/m2	Prop. de pessoas com o ensino superior completo
Género	Prop. de habitações com área até 50 m2	Prop. de pessoas empregados no setor primário
Zona	Prop. de habitações com área entre 50 m2 e 100 m2	Prop. de pessoas empregados no setor secundário
Antiguidade	Prop. de habitações com área entre 100 m2 e 200 m2	Prop. de pessoas pensionistas ou reformados
Garantias do Produto	Prop. de habitações com área superior a 200 m2	Tx. utilização global de consultas médias em um ano
Capital	Prop. de habitações com 1 a 2 divisões	Tx. de urgências
Canal Comercial	Prop. de habitações com 3 a 4 divisões	Tx. de intervenções cirúrgicas
Forma de Pagamento	Prop. de Analfabetos	Tx. de primeiras consultas realizadas em tempo adequado
Tipo de Pagamento	Prop. de pessoas com o ensino secundário completo	Tx. de mulheres com mamografia realizada nos últimos dois anos
	Prop. de pessoas com o ensino pós-secundário completo	

Probabilidade do Tipo de Custos

Variáveis Internas	Variáveis Externas	
Idade	Valor de renda contratada média/m2	Prop. de pessoas com o ensino superior completo
Género	Prop. de habitações com área até 50 m2	Prop. de pessoas empregados no setor primário
Zona	Prop. de habitações com área entre 50 m2 e 100 m2	Prop. de pessoas empregados no setor secundário
Antiguidade	Prop. de habitações com área entre 100 m2 e 200 m2	Prop. de pessoas sem atividade económica
Garantias do Produto	Prop. de habitações com área superior a 200 m2	Tx. utilização global de consultas médias em um ano
Capital	Prop. de habitações com 1 a 2 divisões	Tx. de urgências
Canal Comercial	Prop. de habitações com 3 a 4 divisões	Tx. de intervenções cirúrgicas
Forma de Pagamento	Prop. de Analfabetos	Tx. de primeiras consultas realizadas em tempo adequado
Tipo de Pagamento	Prop. de pessoas com o ensino secundário completo	Tx. de mulheres com mamografia realizada nos últimos dois anos
	Prop. de pessoas com o ensino pós-secundário completo	

ANEXO 8 - Resultados do modelo do Custo do Tipo I com as variáveis internas e as variáveis externas

Idade

	E(X<=10.500|V)
Segurado Padrão	3.026 €
Idade [0,5]	2.142 €
Idade [6,10]	2.142 €
Idade [11,15]	2.142 €
Idade [16,20]	3.026 €
Idade [21,25]	3.026 €
Idade [26,30]	3.026 €
Idade [31,35]	3.026 €
Idade [36,40]	3.026 €
Idade [41,45]	3.313 €
Idade [46,50]	3.313 €

	E(X<=10.500|V)
Segurado Padrão	3.026 €
Idade [51,55]	3.313 €
Idade [56,60]	3.629 €
Idade [61,65]	3.629 €
Idade [66,70]	3.629 €
Idade [71,75]	3.629 €
Idade [76,80]	3.629 €
Idade [81,85]	3.629 €
Idade [86,90]	3.629 €
Idade [91,95]	3.629 €

Género

	E(X<=10.500|V)
Segurado Padrão	3.026 €
Género F	3.026 €
Género M	3.152 €

Canal Comercial

	E(X<=10.500|V)
Segurado Padrão	3.026 €
Canal A	3.026 €
Canal B	3.026 €
Canal E	3.026 €
Canal G	3.026 €

	E(X<=10.500|V)
Segurado Padrão	3.026 €
Canal C	3.026 €
Canal D	3.026 €
Canal E	3.221 €
Canal H	3.221 €

ANEXO 9 - Resultados do modelo da Probabilidade do Tipo de Custos com as variáveis internas e as variáveis externas

Idade

	E(PX1|V)	E(PX2|V)	E(PX3|V)
Segurado Padrão	96,0%	3,8%	0,2%
Idade [0,5]	99,9%	0,0%	0,1%
Idade [6,10]	100,0%	0,0%	0,0%
Idade [11,15]	95,9%	3,7%	0,4%
Idade [16,20]	98,4%	1,6%	0,0%
Idade [21,25]	99,7%	0,0%	0,3%
Idade [26,30]	97,9%	2,1%	0,0%
Idade [31,35]	96,6%	3,3%	0,1%
Idade [36,40]	96,0%	3,7%	0,3%
Idade [41,45]	96,0%	3,8%	0,2%

	E(PX1|V)	E(PX2|V)	E(PX3|V)
Segurado Padrão	96,0%	3,8%	0,2%
Idade [46,50]	93,5%	6,0%	0,5%
Idade [51,55]	93,1%	6,4%	0,5%
Idade [56,60]	90,8%	8,2%	1,0%
Idade [61,65]	91,6%	7,4%	1,0%
Idade [66,70]	89,9%	9,4%	0,7%
Idade [71,75]	88,9%	10,1%	1,0%
Idade [76,80]	88,6%	10,1%	1,3%
Idade [81,85]	90,2%	9,4%	0,4%
Idade [86,90]	90,2%	9,4%	0,4%
Idade [91,95]	90,2%	9,4%	0,4%