INSTITUTO UNIVERSITÁRIO DE LISBOA

Ano Letivo 2022/2023

15 de junho de 2022

PISA 2018

Educação na Finlândia

UC Métodos de Aprendizagem Não Supervisionada

Licenciatura Ciência de Dados

Grupo Finlândia, CDB1 e CDB2

Docentes

Jorge Sinval

José Gonçalves Dias

André Silvestre Nº104532

Diogo Catarino Nº104745

Francisco Gomes Nº104944

Maria Margarida Pereira Nº 105877

Rita Matos Nº104936

Índice

Introdução

Educação na Finlândia

O sistema educativo da Finlândia é amplamente reconhecido como um modelo de inovação inigualável comparativamente ao que é praticado a nível mundial.

Ao contrário de grande parte dos sistemas de educação ao redor do mundo, na Finlândia todas as escolas são de ensino gratuito, isto é, apesar da existência de escolas privadas estas são financiadas pelo estado e pelos municípios, garantindo que os alunos não tenham quaisquer encargos financeiros relacionados à sua educação. Além disso, os professores são altamente qualificados, tendo uma formação contínua ao longo da sua carreira. Aliando o facto de a carga horária ser muito inferior aos restantes países e inexistência de trabalhos de casa, é culturalmente habitual que os estudantes tenham várias atividades extracurriculares, permitindo o desenvolvimento de competências transversais.

Apesar dessa abordagem pouco ortodoxa, onde os alunos passam menos horas na escola e não investem o seu tempo (fora dela) em trabalhos de casa, os estudantes finlandeses continuam a alcançar resultados de excelência, como evidenciado pelos resultados obtidos no PISA. [1]

PISA 2018 | OCDE

O *Programa Internacional de Avaliação de Alunos* (PISA) da OCDE procura investigar o conhecimento dos estudantes nas áreas da leitura, matemática e ciências, bem como a sua capacidade de aplicar esse conhecimento de forma prática.

Este programa aprovisiona a avaliação internacional mais inclusiva e rigorosa da educação dos alunos de diferentes países até à data. Os resultados que dele provêm do PISA indicam a qualidade e a equidade da educação mundial e permite que pedagogos e decisores políticos instruam-se com as políticas e práticas aplicadas noutros países. [2]

No âmbito do PISA 2018, foi atribuído um enfase especial à avaliação da componente de leitura, visando aferir a capacidade dos estudantes em compreender, utilizar e refletir sobre textos escritos em variados contextos. [3]

PISA 2018 | Finlândia

Com o propósito de aprofundar o conhecimento acerca do sistema educativo finlandês, contactámos o NPM *(National Project Manager)* encarregue do PISA deste país. Após contactado, Arto Ahonen, *Senior Researcher* e *NPM* da Finlândia, confirmou que não ocorreram grandes mudanças no seu sistema educativo de 2012 a 2018.

Adicionalmente, apurámos que, a média do tamanho das turmas para estudantes de 15 anos é relativamente pequena, o que pode favorecer um ambiente de aprendizagem mais individualizado. [4]

Além disso, apresenta uma das menores percentagens de estudantes que relatam sofrer *bullying* regularmente, abrangendo todas as formas de *bullying*. Isto indica um ambiente escolar mais seguro e um clima social favorável ao bem-estar dos alunos.

É ainda importante destacar que a Finlândia tem uma cobertura significativa da população de 15 anos no PISA 2018, o que indica uma amostra representativa e confiável dos alunos finlandeses. Isso contribui para a validade e a relevância dos resultados obtidos. [4]

Base de Dados

Neste sentido, de modo a fazer um estudo desta temática, a base de dados fornecida diz respeito ao PISA 2018. Esta é composta por dados de vários países, sendo que apenas nos cingiremos à Finlândia, o correspondente a **5649** observações do *dataset* completo. [4][5]

Questão Problema do Trabalho

A principal questão que tencionamos responder prende-se com: *Quais são os fatores-chave fundamentais que contribuem para o sucesso dos alunos finlandeses no PISA 2018?*

Para responder a este objetivo utilizaremos técnicas de *PCA*, de modo a sintetizar a informação das variáveis, criando componentes; e *clustering,* de forma a analisar perfis de alunos. Neste sentido, faremos uma breve análise aos dados, dividiremos em variáveis PROFILE e INPUT, identificaremos as dimensões da análise utilizando a técnica de *PCA*, e avaliaremos a heterogeneidade dos alunos através do *clustering.*

Análise dos Dados

Descrição dos Dados e Pré-Processamento

Nesta fase procurámos enquadrar e analisar descritivamente as variáveis, compreendendo os seus significados no *dataset*, e limpar os dados de modo a puderem ser aplicados nas fases seguintes.

Tendo em vista o referido, começámos por importar a base de dados e estudar o significado de cada variável de modo a compreendermos o que é analisado no PISA. Verificámos ainda possíveis duplicados, não se tendo verificado nenhum caso.

Posteriormente, através de uma análise inicial aos dados e após restringimos apenas aos estudantes inquiridos da Finlândia, o equivalente a **`5649 observações`**, podemos observar um número elevado de valores omissos quer por linha, quer por coluna. Para tratar estes casos decidimos optar por uma heurística que elimina as colunas com mais de 60% de valores omissos e depois as linhas com mais de 20 variáveis omissas de um total de 71, de modo a não imputarmos mais dados do que aqueles que estão preenchidos, o que poderia enviesar em demasia os resultados.

No fim dessa limpeza foram eliminadas 33 colunas de 104 totais e 400 observações (≈7% dos inquiridos da Finlândia).

De modo a trabalhar com os restantes casos que ainda apresentam alguns valores omissos, imputámos de duas formas distintas esses valores, uma usando a *Regressão Linear*, e outra usando a *Random Forest*, pudendo assim utilizar as **`5249 observações`** que integram o *dataset*. Adicionalmente, criámos ainda um *DataFrame* apenas com as observações completas, isto é, eliminámos todas as observações que continham valores omissos, ficando apenas **`3096 observações`**. Optámos por fazer estas três abordagens de lidar com valores omissos, para avaliar se os resultados das fases seguintes mudam conforme esta limpeza.

Escolha das Variáveis PROFILE e INPUT

De seguida, dividimos as restantes variáveis do *dataset* em variáveis de PROFILE (para caracterização dos *clusters*) e em variáveis de INPUT (utilizadas no PCA e posterior *Clustering*).

O critério de seleção destas variáveis provem do conhecimento prévio que adquirimos do seu significado, pelo que optámos por separar em variáveis que medem o desempenho escolar e possíveis fatores que o influenciam para variáveis de INPUT; e, por outro lado, as variáveis de PROFILE com os elementos de caráter socioeconómico do estudante e dos pais, que o permite caracterizar. É possível encontrar estas variáveis e a sua descrição no Anexo 1.

Análise Exploratória de Dados – Caracterização dos Estudantes

Após a preparação dos dados, procedemos à análise de gráficos e tabelas para caracterizar os estudantes finlandeses inquiridos no PISA 2018. Todos os elementos de visualização produzidos podem ser obtidos através do código desenvolvido e enviado, destacando os seguintes.

Analisando os dados da amostra que *a priori* representa a população de estudantes, podemos inferir através da variável **`IMMIG`** que a maioria (94.8%) dos estudantes são nativos, ou seja, pelo menos um dos pais nasceu na Finlândia; e através da variável **`ST004D01T`** que ambos os géneros dos estudantes (feminino e masculino) estão equitativamente representados na amostra.

Relativamente à variável **`ESCS`** podemos inferir que um valor mais elevado indica um contexto socioeconómico e cultural mais favorável. Esta variável varia entre -3 e 3, observando-se que a maior parte dos casos encontra-se acima de 0, ou seja, existem mais alunos que vivem num contexto socioeconómico e cultural favorável. [6]

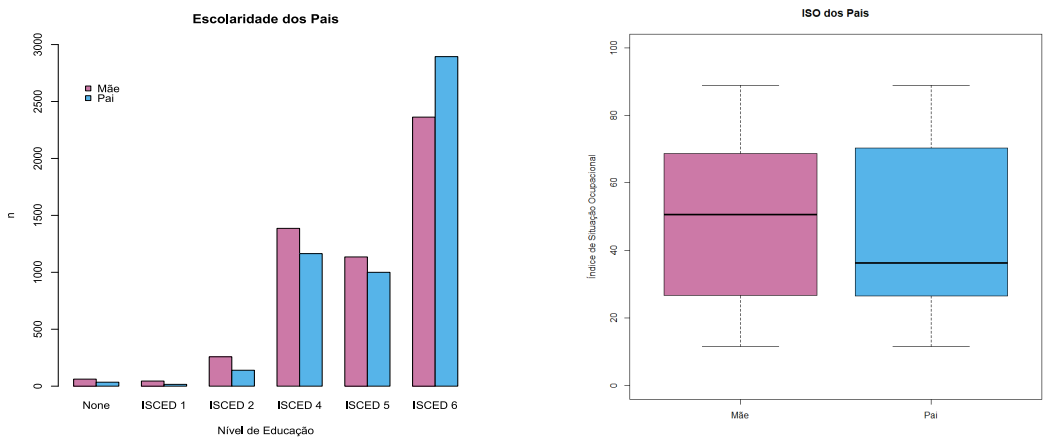


Figura 1 | Gráfico de Barras e *Boxplot* das variáveis relativas à escolaridade e ISO dos pais.

O primeiro gráfico (**Fig.1**), relativo ao grau de escolaridade dos pais dos alunos inquiridos, indicado através do ISCED (*International Standart Classification of Education*), evidencia que os pais dos estudantes finlandeses têm um grau de escolaridade elevado.

Já o seguinte gráfico (**Fig.1**), representativo do *Índice de Situação Ocupacional* revela que os valores medianos são mais elevados na figura materna, sendo mais baixo para a figura paterna. Este índice é calculado a partir de respostas a perguntas abertas que é, posteriormente, codificado, indicado quais são, por exemplo, as *"Elementary Occupations"*.

Análises de PCA

De modo a cumprir o objetivo proposto, começámos por recorrer à técnica de redução de dimensionalidade da Análise de Componentes Principais (PCA).

Teoricamente, esta técnica prende-se com a extração dos principais construtos subjacentes aos dados observados, através de combinações lineares, permitindo uma representação mais concisa e significativa das variáveis originais, minimizando a perda de informação e maximizando a informação distinta em cada componente. [7]

Adequabilidade da Amostra

Primeiramente, procedeu-se à verificação da adequabilidade da amostra que contém as variáveis de INPUT pré-selecionadas para o PCA. Para tal, averiguámos as correlações de *Pearson* de todas as variáveis através da matriz ilustrada no Anexo 2.

Verificámos a existência de padrões de associação entre as 58 variáveis diferentes. Alguns pares de variáveis mostram uma correlação positiva significativa, indicando uma relação direta, enquanto outros apresentam uma correlação negativa, sugerindo uma relação inversa.

Além destas, eliminámos as variáveis que possuem correlação fraca com as demais e as que apresentam correlação perfeita, para evitar haver informação dispensável e redundante, respetivamente, razão que antevemos piorar os resultados do PCA.

Após essa análise, testámos o *dataset* com o Teste de *Bartlett* que tem como hipótese nula as variáveis não estarem correlacionadas na população. Deste modo, rejeitámos H_0 dado que $p - value < 0.05$ (considerando $\alpha = 5\%$), verificando-se a adequabilidade da amostra para fazer o PCA.

Por fim, calculámos o índice KMO (Critério de *Kaiser–Meyer–Olkin*) que compara a magnitude dos coeficientes de correlação observados com a magnitude dos coeficientes de correlação parcial, e para o qual queremos obter valores acima de 0.6, de modo que a matriz seja favorável. O valor global do KMO na amostra foi de 0.74, porém dado existirem variáveis com KMO inferior a 0.6, optámos por eliminá-las. (Anexo 3)

Número de Componentes

Para escolher o número de componentes mais adequado para o PCA, utilizámos os critérios: Critério de *Kaiser*, *ScreePlot* e Variância Explicada Acumulada.

Iniciando pelo Critério de *Kaiser*, que determina a seleção exclusiva de fatores com valores próprios superiores a 1, tomaríamos a opção de reter 9 componentes. Já com a visualização do *ScreePlot* (Anexo 4), constata-se a indicação de 7 ou 9 componentes. E, por último, com a Variância Explicada Acumulada, estabelecendo a retenção de, no mínimo, 60% da informação, teríamos de fazer 9 componentes.

Assim, e por unanimidade dos critérios, realizámos o PCA utilizando 7 e 9 componentes principais (Anexo 5), sendo que a opção mais adequada e relevante para o contexto em questão foi com 9 componentes. Esta escolha permitiu a retenção de maior quantidade de informação face ao PCA com 7 PCs e, simultaneamente, proporcionou uma interpretabilidade aprimorada dos componentes obtidos.

Rotação *Varimax*

Após definir o número de componentes a reter, e observar os resultados sem rotação, decidimos analisar o PCA com rotação *Varimax*. Esta técnica de rotação é amplamente adotada no PCA e visa aprimorar a interpretação dos resultados, conferindo maior relevância aos *loadings* dos fatores, o que facilita substancialmente a compreensão e a interpretabilidade dos mesmos.

Resultados

Após analisar os componentes, rotulámo-los observando a sua correlação com as variáveis através dos *loadings*, obtendo-se os seguintes componentes:

PCs	Designação da Componente	Descrição da Componente
1	Teacher's Engagement	Interação professor-aluno e clima escolar
2	Academic Performance	Notas dos alunos em literatura, matemática e ciências
3	Student's Mental Well-being	Bem-estar mental, resiliência e satisfação com a vida
4	Student-ICT Relation	Familiaridade e competências dos alunos com os recursos TIC
5	Digital Access and Resources	Bens materiais e recursos TIC disponíveis em casa do aluno
6	ICT Use	Uso de TIC na escola e fora da escola
7	Attitudes towards School and Learning	Envolvimento em atividades de aprendizagem
8	Competition in School	Perceção de competição na escola
9	Digital Learning Enrichment	Relação entre o uso de TIC na aprendizagem

***TIC -** Tecnologias da Informação e Comunicação

Análises de Clustering

Para a avaliação da heterogeneidade dos alunos, iremos seguir a abordagem do *clustering* que visa agrupar os alunos em *clusters* com características semelhantes, com o intuito de identificar e explorar padrões. Para tal, recorreu-se a diferentes métodos de *clustering,* tais como, *Clustering* Hierárquico, *K-means*, PAM *(Partition Around Medoids)* e GMM *(Gaussian Mixture Models).*

É de notar que os *scores* produzidos na redução de dimensionalidade através da técnica de PCA já se encontram estandardizadas, pelo que não se procede à estandardização no *clustering*.

Clustering Hierárquico

A abordagem de *Clustering* Hierárquico cria uma estrutura de árvore que representa as relações entre os dados, podendo ser visualizada num dendrograma que regista as sequências de fusões ou divisões das observações. Este método oferece flexibilidade na escolha do critério de agrupamento. [8]

Seguindo este método, tentámos duas abordagens de critérios distintas utilizando o algoritmo de *Ward* e, posteriormente, o *complete-linkage*. Para avaliar a sua performance utilizámos o coeficiente de *Silhouette*, que varia de -1 a 1, onde valores próximos a 1 indicam *clusters* bem definidos e valores próximos a -1 indicam atribuições incorretas dos objetos aos *clusters*.

Primeiramente, utilizando o algoritmo de *Ward*, que se baseia na minimização da soma dos quadrados das diferenças dentro dos *clusters*, é possível agrupar os dados de forma a combinar *clusters* que apresentem a menor variância possível após a união. Assim, analisando o dendrograma produzido (Anexo 6) tentou-se a criação de 5 *clusters*. Este obteve uma *Average Silhouette Width* de 0.02 que simboliza uma estrutura quase inexistente e fraca.

Posteriormente, adotando como critério de semelhança entre *clusters* o *complete-linkage*, no qual a semelhança de dois *clusters* tem por base os dois pontos menos semelhantes (mais distantes) em *clusters* diferentes, optámos por cortar em 3 *clusters* (Anexo 6), obtendo-se uma *Average Silhouette Width* de 0.09.

K-means

O *K-means* é um algoritmo de *clustering* amplamente utilizado nas abordagens de particionamento. Este algoritmo é projetado para dividir um conjunto de dados em k grupos distintos, onde k é um número pré-definido.

Primeiro seleciona-se aleatoriamente os centroides iniciais e, em seguida, atribui-se pontos de dados aos centroides mais próximos. Posteriormente, recalcula-se os centroides com base nos pontos atribuídos e repete-se este processo até que ocorra a convergência, ou seja, até que não haja mais alterações nos centroides ou nas atribuições dos pontos. [8]

De modo a definir o número de *clusters* recorreu-se à visualização do gráfico WSS *(Within-Cluster Sum of Squares)* que apresenta a soma dos quadrados das distâncias de cada ponto dentro de um *cluster* em relação ao seu centroide. Desta forma, optámos por utilizar $k = 5$, obtendo-se uma *Average Silhouette Width* de 0.07.

PAM - Partition Around Medoids

Ao contrário do *K-means* que utiliza pontos médios calculados (centroides), o PAM é um algoritmo de particionamento que usa medoides, pontos de dados que pertencem ao conjunto de dados, como representantes dos grupos.

Inicia-se escolhendo aleatoriamente k medoides e, em seguida, atribui-se pontos de dados aos medoides mais próximos. O algoritmo procura melhorar a qualidade dos grupos trocando medoides e avaliando o impacto na função objetivo. Este processo é repetido até que não haja mais melhorias. [8]

Utilizando o mesmo k, testámos o método PAM, como visível no Anexo 8, sendo que dele não é possível retirar conclusões significativas sobre os dados. Porém, dele denota-se o facto de que as 2 componentes principais (PC1 e PC2) explicam cerca de 22.22% da variabilidade do *dataset*.

GMM - Clustering Probabilístico

Por último, o *clustering* probabilístico é uma abordagem de agrupamento que atribui probabilidades a cada ponto de dados pertencer a diferentes grupos.

Como modelo desta abordagem utilizámos o GMM (*Gaussian Mixture Model*), o qual assume que os dados são gerados a partir de uma combinação de várias distribuições Gaussianas (distribuições normais).

Ao contrário dos métodos tradicionais, esta abordagem permite que um ponto pertença a múltiplos grupos simultaneamente, modelando a distribuição de probabilidade dos dados e ajustando os parâmetros do modelo para maximizar a verossimilhança dos dados observados.

O GMM, em particular, representa cada cluster por uma distribuição gaussiana, sendo capaz de modelar *clusters* com diferentes formas e tamanhos. Além disso, fornece estimativas de densidade probabilística para cada ponto de dados, o que é útil em várias aplicações. [8]

Para selecionar o modelo probabilístico a aplicar no GMM, usámos o critério BIC (*Bayesian Information Criterion*) que produz o gráfico presente no Anexo 9. Através da sua análise, verifica-se que o melhor modelo apresenta 6 *clusters* uma estrutura do tipo VVE, isto é, os *clusters* serão elipsoides cujo volume e forma varia, mas estão alinhados numa mesma orientação.

Os *clusters* apresentados (Anexo 10) são difusos, tendo-se maior dificuldade a relacionar e separá-los, podendo-se constatar que as cores estão concentradas e não é possível discernir claramente *clusters* distintos de forma significativa. Assim, evidencia-se a dificuldade em identificar perfis bem definidos de alunos através deste método.

Resultados

Após avaliar todas as abordagens realizadas, optámos por utilizar os *clusters* resultantes do algoritmo de *K-means*. Esta decisão fundamenta-se na sua maior interpretabilidade, decorrente da capacidade deste método em produzir agrupamentos que apresentam uma estrutura e separação mais evidentes, facilitando assim a compreensão e análise dos padrões subjacentes aos dados. (Anexo 11)

Segue-se numa tabela a designação e descrição dos 5 *clusters* produzidos:

	Designação	Descrição do Cluster
Cluster 1 (596 observações)	Digital Students	Estudantes que exibem uma boa atitude em relação à escola e ensino (PC7_ATSL), ao acesso a tecnologias (PC9_DLE) e uma boa componente digital (PC6_IU); no entanto, exibem uma menor performance académica (PC2_AP).
Cluster 2 (928 observações)	Disengaged Students	Estudantes com um nível relativamente baixo de desempenho académico (PC2_AP), uma forte competência digital (PC6_IU) e acesso a recursos digitais (PC5_DAR).
Cluster 3 (1342 observações)	Non-tech Students	Estudantes que demonstram um nível moderado de bem-estar mental (PC3_SMW), mas apresentam baixo envolvimento com as tecnologias da informação e comunicação (PC6_IU) e recursos digitais (PC5_DAR).
Cluster 4 (950 observações)	Disconnected Students	Estudantes que apresentam um baixo nível de bem-estar mental (PC3_SMW), uma interação positiva com as tecnologias da informação e comunicação (PC4_SIR), porém uma utilização relativamente baixa das TIC em sala de aula e em casa (PC6_IU).
Cluster 5 (1433 observações)	High Achievers	Estudantes que mostram um desempenho académico destacado (PC2_AP) e envolvimento em atividades de aprendizagem (PC7_ATSL), têm baixa familiaridade e competências com recursos TIC (PC4_SIR).

De seguida, com o intuito de padronizar os alunos pelas suas características, visualizámos uma relação entre os *clusters* com as variáveis PROFILE.

Ao segmentar os *clusters* por género é possível verificar que nos *clusters* 1 *(Digital Students)* e 4 *(Disconnected Students)* não há discriminação por género. Contrariamente, em relação ao *cluster* 2 *(Disengaged Students)* e cluster 5 *(High Achievers)* há grandes discrepâncias, onde 64% dos estudantes do cluster 2 são do sexo masculino e 68% do cluster 5 são do sexo feminino. Com isto, podemos observar que os estudantes do sexo feminino têm maior tendência a obter bons resultados escolares.

Gender	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Female	0.5	0.36	0.43	0.47	0.68
Male	0.5	0.64	0.57	0.53	0.32

Figura 2 | Tabela com as proporções do género dos alunos por *cluster*.

Através da análise do gráfico de dispersão que relaciona as variáveis relativas ao *Índice de Situação Económica, Social e Cultura da Família* (ISESC), o *Índice de Situação Ocupacional Parental Mais Alto* (HISEI) e os clusters apurados (Anexo 12) é possível verificar que os alunos que apresentam pais com o menor status ocupacional e o menor ISESC são os "*Digital Students*", o que pode ser uma explicação para a sua menor performance académica, mas com uma boa atitude em relação à escola.

Por contrapartida, os alunos com elevados valores de ambos os índices são os "*Non-tech Students*", caracterizando-se por um melhor bem-estar mentalmente, mas não se envolvem muito com as tecnologias.

Ao considerarmos a totalidade dos *clusters*, é notável a tendência progressiva e favorável do ISESC em consonância com o índice ocupacional dos progenitores, suscitando, desta forma, uma segmentação mais precisa dos estudantes.

Conclusão

Respondendo à questão problema que nos propusemos a dar resposta, "*Quais são os fatores-chave fundamentais que contribuem para o sucesso dos alunos finlandeses no PISA 2018"*, podemos aferir que a situação económica, social e cultural, o bem-estar mental e o envolvimento com tecnologias de informação e comunicação do aluno representam as características mais significativas para o seu sucesso escolar e consequentemente bons resultados no PISA.

Considerando a imprescindibilidade de avaliar o sistema educativo da Finlândia a fim de extrair conclusões e as ideias fundamentais que possibilita alcançar os resultados de excelência do PISA 2018, elaborámos o presente estudo.

Através da redução de dimensionalidade e posterior perfilização dos estudantes constatámos que os estudantes finlandeses demonstram uma atitude positiva em relação à escola.

Importa salientar que comparando os resultados obtidos com os que se obtêm removendo os valores omissos, ou imputando-os com a regressão linear produzem-se resultados com ligeiras variações quantitativas que não comprometem as conclusões finais a serem retidas.

Como já referido anteriormente, há segmentação dos *clusters* por sexo, a qual constatou-se que o sexo feminino tende a obter melhores resultados escolares, sendo predominante no cluster *High Achievers*.

É de notar a relação peculiar que alunos com um nível moderado de bem-estar mental e um desempenho académico destacado têm baixa interação com recursos TIC.

Em suma, consideramos que cumprimos com sucesso o objetivo pretendido neste projeto, corroborando a excelência de educação finlandesa em relação aos demais sistemas educacionais mundiais.

Bibliografia

[1] Colagrossi, M. (2018, September 10). *10 reasons why Finland's education system is the best in the world*. World Economic Forum; Big Think. https://www.weforum.org/agenda/2018/09/10-reasons-why-finlands-education-system-is-the-best-in-the-world

[2] OECD. (2018). *PISA 2018 Results*. OECD.org. https://www.oecd.org/pisa/publications/pisa-2018-results.htm

[3] OECD (2020), *PISA 2018 Results (Volume V): Effective Policies, Successful Schools*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/ca768d40-en.

[4] PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT (PISA) RESULTS FROM PISA 2018 | FINLAND. (2019). In OECD. https://www.oecd.org/pisa/publications/PISA2018_CN_FIN.pdf

[5] OECD. (2021). The Future of Education and Skills 2030: OECD Learning Compass 2030. https://www.oecd-ilibrary.org/sites/0a428b07-en/index.html?itemId=/content/component/0a428b07-en

[6] Eurostat. (n.d.). *International Standard Classification of Education (ISCED)*. Ec.europa.eu. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCED)

[7] Hair, J. F., Black, W. C., & Babin, B. J. (2018). *Multivariate data analysis (8th ed.)*. Cengage Learning Emea. Copyright.

[8] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), An Introduction to Statistical Learning: with applications in R, New York: Springer.

Anexos

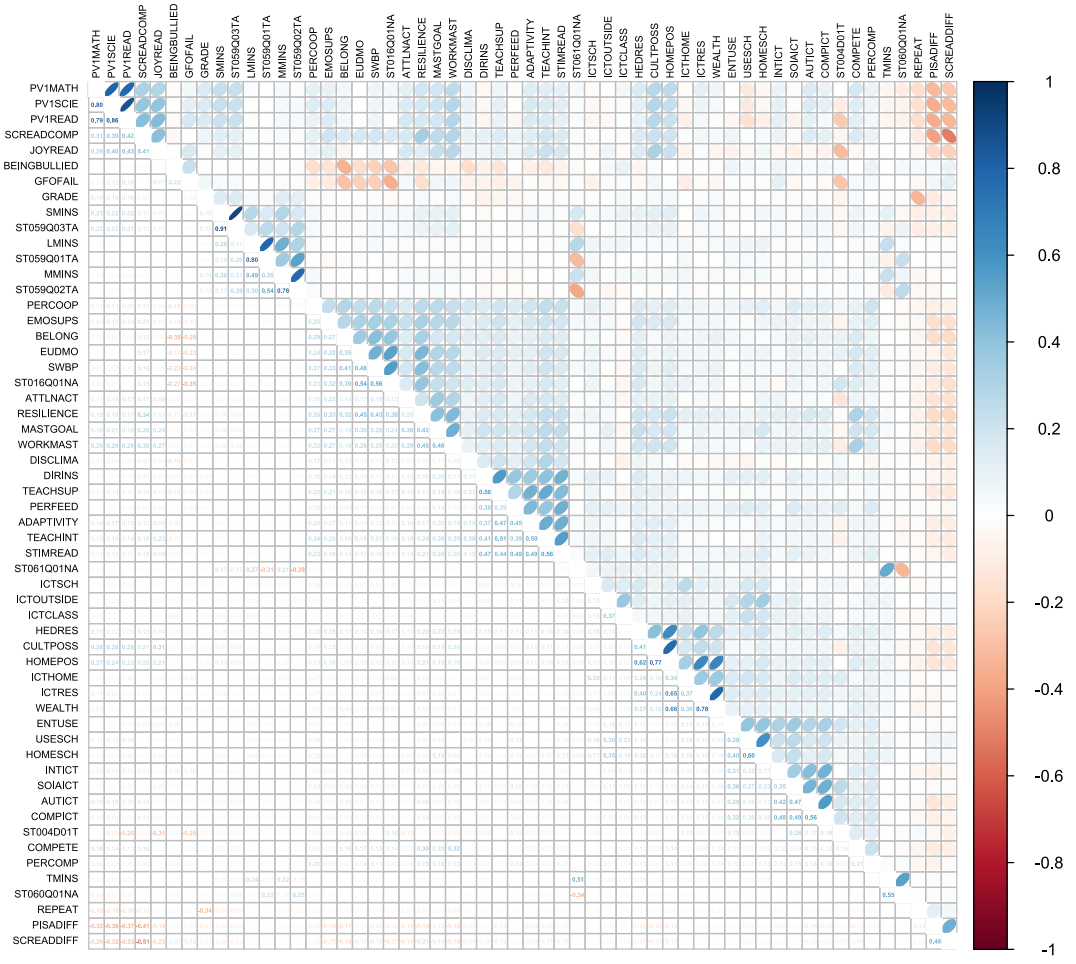
Anexo 1 | Tabela com a descrição das variáveis

	Variáveis	Descrição
Variáveis de **PROFILE**	AGE	Idade
	IMMIG	Índice de situação de imigração
	ESCS	Índice de situação económica, social e cultural
	HISEI	Índice de situação ocupacional parental mais alto
	BFMJ2	Índice de situação ocupacional do pai
	BFMJ1	Índice de situação ocupacional da mãe
	HISCED	Maior escolaridade dos pais
	FISCED	Educação do pai
	MISCED	Educação da mãe
	ISCEDL	Nível de ISCED
	PROGN	Código único do programa de estudos nacional
	OCOD1	Codificação dos dados ocupacionais dos pais do aluno
	OCOD2	Codificação mais detalhada dos dados ocupacionais dos pais dos alunos
	OCOD3	ISCO-08 código de ocupação do próprio
	ST004D01T	Sexo dos estudantes (**1**=*Female* | **2**=*Male*)
Variáveis de **INPUT**	PV1SCIE	Nível de proficiência do estudante em ciências
	PV1READ	Nível de proficiência do estudante em leitura
	PV1MATH	Nível de proficiência do estudante em matemática
	ICTOUTSIDE	Uso de recursos TIC fora das aulas
	ICTCLASS	Uso de recursos TIC durante as aulas
	SOIAICT	TIC como um tópico na interação social
	AUTICT	Autonomia com recursos TIC
	COMPICT	Competências TIC
	INTICT	Interesse do estudante em recursos TIC
	USESCH	Uso de recursos TIC na escola
	HOMESCH	Uso de recursos TIC fora da escola para atividades de trabalho escolar
	ENTUSE	Uso de recursos TIC fora da escola para lazer
	BEINGBULLIED	Experiência dos alunos de sofrerem *bullying*
	BELONG	Bem-estar subjetivo: integração do aluno na escola

MASTGOAL	Orientação para metas de domínio
RESILIENCE	Resiliência do estudante
SWBP	Bem-estar subjetivo: afeto positivo
EUDMO	Eudaimonia: significado na vida
GFOFAIL	Medo geral de falhar
WORKMAST	Domínio no trabalho
COMPETE	Competitividade
ATTLNACT	Atitude em relação à escola: atividades de aprendizagem
PERCOOP	Perceção de cooperação na escola
PERCOMP	Perceção de competição na escola
PISADIFF	Perceção da dificuldade do teste PISA
SCREADDIFF	Autoconceito de leitura: perceção de dificuldade
SCREADCOMP	Autoconceito de leitura: perceção de competência
JOYREAD	Gosto pela leitura
TEACHINT	Perceção de interesse do professor
ADAPTIVITY	Adaptação do ensino
STIMREAD	Estímulo do professor ao envolvimento com a leitura percebido pelo aluno
EMOSUPS	Perceção de apoio emocional dos seus pais
PERFEED	Avaliar o feedback percebido pelo professor
DIRINS	Instrução dirigida pelo professor
TEACHSUP	Apoio do professor em aulas de teste de idioma
DISCLIMA	Clima disciplinar nas aulas de línguas
ICTREES	Recursos TIC
WEALTH	Riqueza familiar
HEDRES	Recursos educacionais domésticos
CULTPOSS	Bens culturais em casa
HOMEPOS	Bens de casa
ICTSCH	TIC disponível na escola
ICTHOME	TIC disponível em casa
TMINS	Tempo de aprendizagem (minutos por semana) no total
SMINS	Tempo de aprendizagem (minutos por semana) de ciências
LMINS	Tempo de aprendizagem (minutos por semana) de línguas
MMINS	Tempo de aprendizagem (minutos por semana) de matemática

GRADE	Nota em comparação com a nota modal no país
ST061Q01NA	Minutos, em média, num período de aula
ST060Q01NA	Numa semana normal e completa na escola, o número de períodos de aula que é necessário assistir no total
ST059Q03TA	Número de períodos de aula por semana em ciências
ST059Q02TA	Número de períodos de aula por semana em matemática
ST059Q01TA	Número de períodos de aula por semana em aulas de teste de idioma
ST016Q01NA	No geral, quão satisfeito está com a sua vida como um todo atualmente
ST001D01T	Nota internacional do aluno
REPEAT	Repetição de um ou mais anos

Anexo 2 | Matriz de Correlação de *Pearson* das variáveis de INPUT

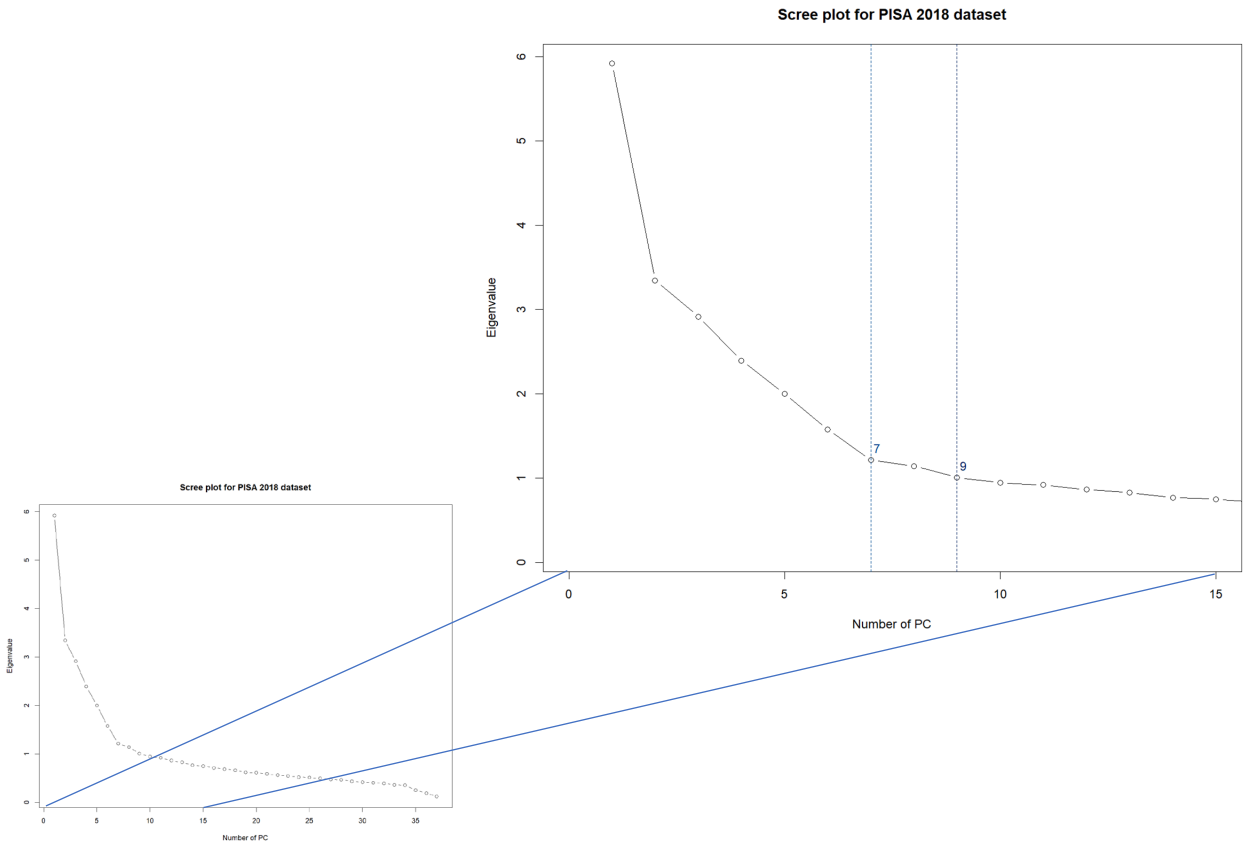


Anexo 3 | KMO

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO (r = correlation)
## Overall MSA = 0.74
## MSA for each item =
##    PV1SCIE    PV1READ    PV1MATH ICTOUTSIDE   ICTCLASS    SOIAICT     AUTICT
##       0.83       0.81       0.86       0.76       0.71       0.87       0.84
##    COMPICT     INTICT     USESCH    HOMESCH     ENTUSE     BELONG   MASTGOAL
##       0.83       0.84       0.79       0.78       0.89       0.91       0.91
## RESILIENCE       SWBP      EUDMO   WORKMAST    COMPETE   ATTLNACT    PERCOMP
##       0.92       0.88       0.89       0.90       0.84       0.89       0.88
## SCREADCOMP    JOYREAD   TEACHINT ADAPTIVITY   STIMREAD    EMOSUPS    PERFEED
##       0.90       0.85       0.90       0.90       0.90       0.94       0.86
##     DIRINS   TEACHSUP   DISCLIMA     ICTRES     WEALTH     HEDRES   CULTPOSS
##       0.85       0.84       0.88       0.86       0.55       0.68       0.55
##    HOMEPOS    ICTHOME      TMINS      SMINS      LMINS      MMINS ST061Q01NA
##       0.62       0.95       0.92       0.56       0.46       0.48       0.41
## ST059Q03TA ST059Q02TA ST059Q01TA ST016Q01NA
##       0.55       0.49       0.47       0.84
```

Os valores salientados representam os KMO inferiores a **0.6** que foram eliminados

Anexo 4 | Gráfico *Screeplot*

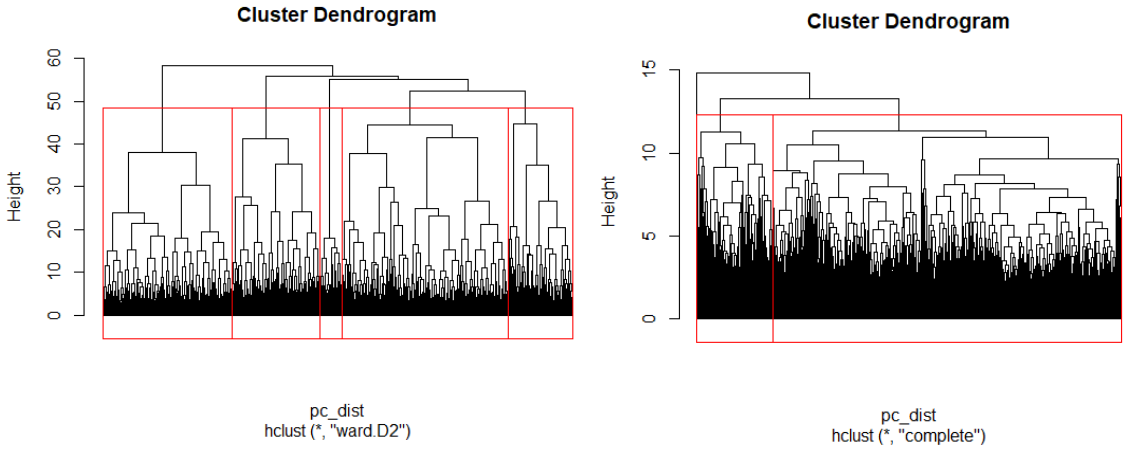


Anexo 5 | Resultados do PCA com 7 e 9 componentes principais

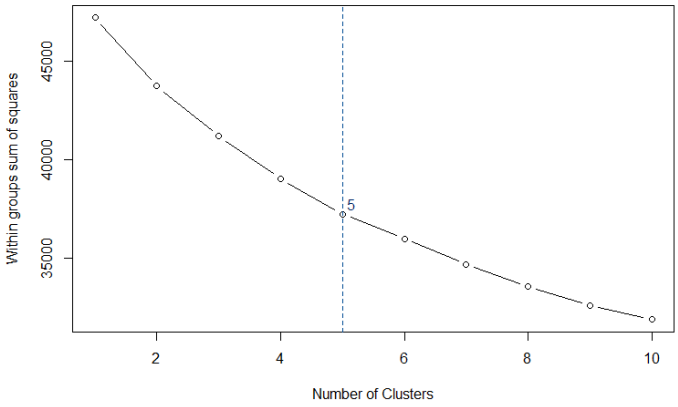
	PCA | 7 Componentes						
	RC1	RC3	RC4	RC2	RC5	RC7	RC6
PV1SCIE		0.903					
PV1READ		0.9188					
PV1MATH		0.852					
ICTOUTSIDE							0.7158
ICTCLASS							0.5917
SOIAICT				0.7031			
AUTICT				0.7706			
COMPICT				0.7887			
INTICT				0.6964			
USESCH							0.681
HOMESCH						0.1962	0.7124
ENTUSE				0.5239			
BELONG			0.6664				
MASTGOAL	0.1908					0.6683	
RESILIENCE			0.5259				
SWBP			0.7617				
EUDMO			0.7128				
WORKMAST		0.2654				0.6646	
COMPETE				0.2025		0.4729	
ATTLNACT						0.5052	
PERCOMP				0.2308		0.4108	
SCREADCOMP		0.5011					
JOYREAD	0.1608	0.5393					
TEACHINT	0.7434						
ADAPTIVITY	0.7209						
STIMREAD	0.7495						
EMOSUPS			0.4698				
PERFEED	0.6528						
DIRINS	0.7055						
TEACHSUP	0.7335						
DISCLIMA							
ICTRES					0.818		
HEDRES					0.6942	0.1568	
HOMEPOS		0.2394			0.8456		
ICTHOME					0.592		
TMINS							0.1588
ST016Q01NA			0.7944				

	PCA | 9 Componentes								
	RC1	RC3	RC4	RC2	RC5	RC6	RC7	RC8	RC9
PV1SCIE		0.9083							
PV1READ		0.9159							
PV1MATH		0.8651							
ICTOUTSIDE						0.4462			0.6019
ICTCLASS						0.2218			0.7163
SOIAICT				0.6964		0.2001			
AUTICT				0.7844					
COMPICT				0.8081					
INTICT				0.7057					
USESCH						0.7669			
HOMESCH						0.794			
ENTUSE				0.4477		0.5403			
BELONG			0.6681						
MASTGOAL	0.1858					0.1654	0.6158	0.2257	
RESILIENCE			0.5521						
SWBP			0.7681						
EUDMO			0.7277						
WORKMAST		0.2548					0.5722	0.3201	
COMPETE								0.6814	
ATTLNACT							0.6198		
PERCOMP								0.6302	
SCREADCOMP		0.4862							
JOYREAD		0.4985							
TEACHINT	0.7342								
ADAPTIVITY	0.7314								
STIMREAD	0.7611								
EMOSUPS			0.4731						
PERFEED	0.6804							0.1656	
DIRINS	0.7026								
TEACHSUP	0.72								
DISCLIMA									
ICTRES					0.8172				
HEDRES					0.7033				
HOMEPOS		0.2309			0.8505				
ICTHOME					0.59				
TMINS									0.5778
ST016Q01NA			0.7986						

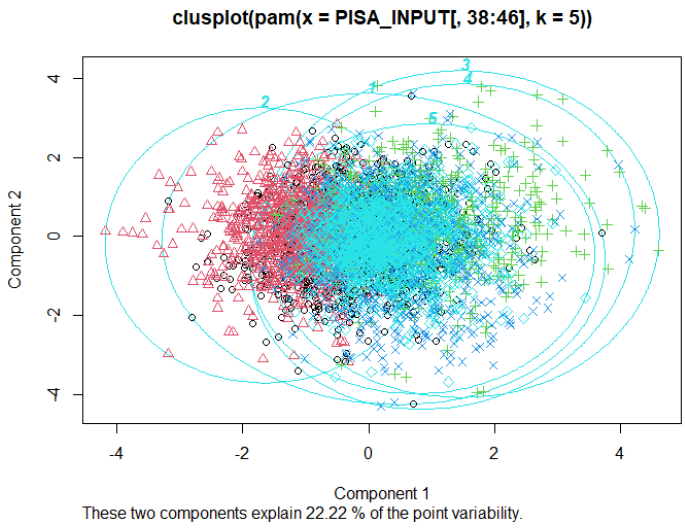
Anexo 6 | Dendrogramas do *Clustering* Hierárquico



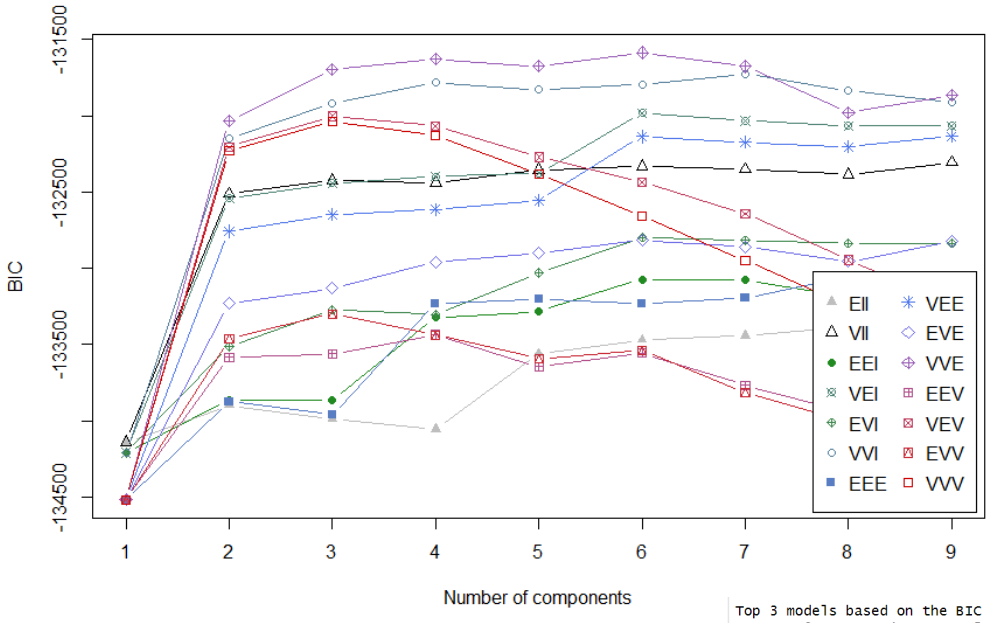
Anexo 7 | Seleção do número de *clusters* ***k*** no *K-means*



Anexo 8 | Gráfico do PCA e Clustering com a técnica de PAM

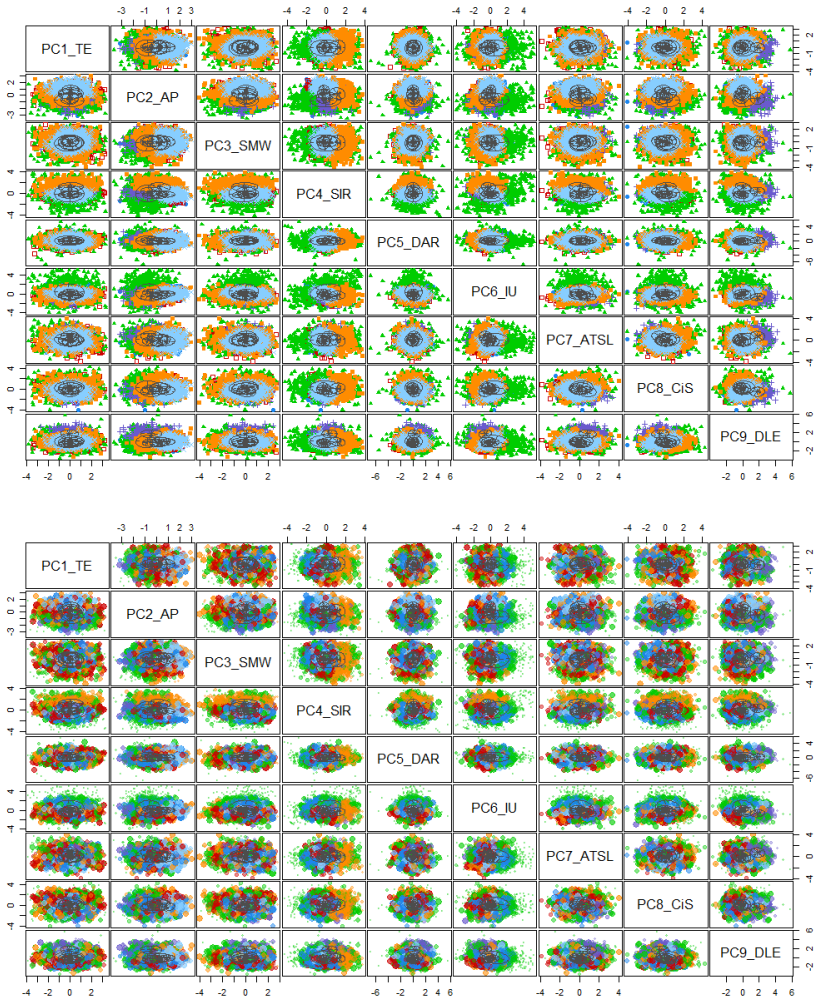


Anexo 9 | Seleção do modelo probabilístico utilizando o critério BIC

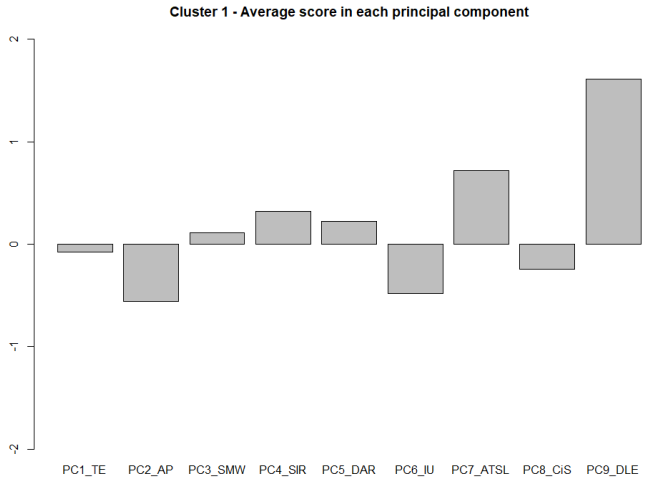


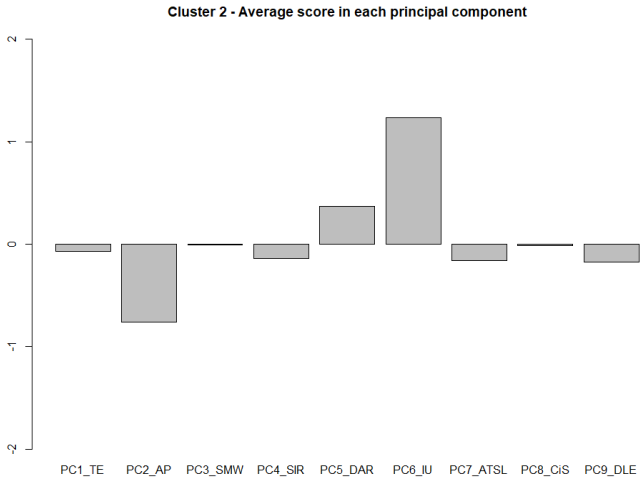

```
Top 3 models based on the BIC criterion:
     VVE,6      VVE,4      VVE,5
-131586.1 -131627.1 -131676.6
```

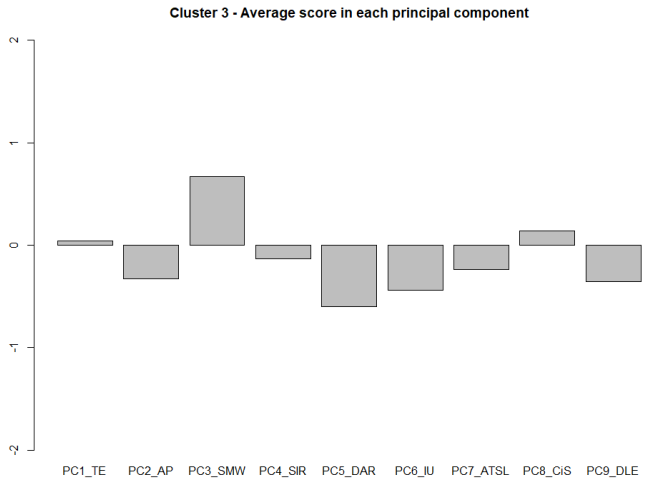
Anexo 10 | Resultados do GMM

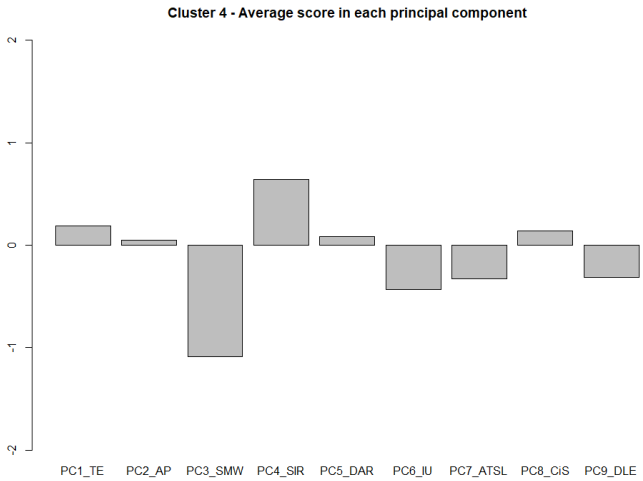


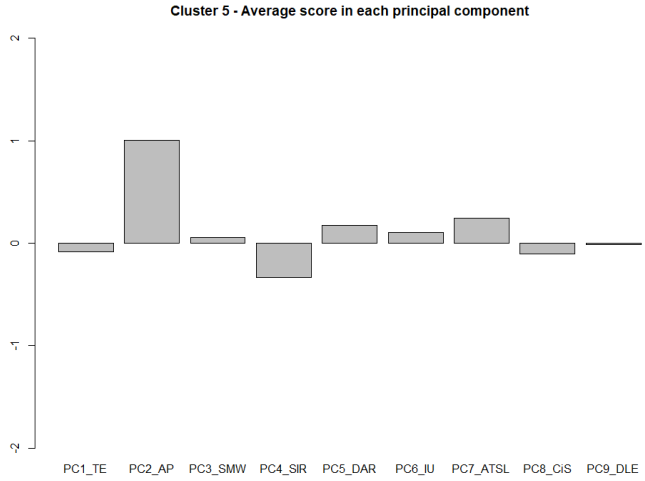
Anexo 11 | Gráficos de barras do *score* médio de cada componente em cada *cluster*











Anexo 12 | Relação entre ESCS e HISEI nos *clusters*