FCTUC – DEI
Mestrado em Engenharia Informática

PA 2 Data Analysis

Segurança em Tecnologias da Informação

Vasco Rodrigues Nr.º 2024169097

05/06/2025

Índice

Índice Figuras

Índice Tabelas

1. Introdução

Este segundo projeto, realizado no âmbito da unidade curricular de Sistemas de Gestão de Dados, tem como objetivo principal efetuar uma análise detalhada de um dataset construído manualmente. Este dataset será elaborado a partir de múltiplas fontes de informação.

Ao longo deste projeto, será especificado todo o tratamento dos dados e as técnicas utilizadas para extrair informações relevantes dos mesmos.

2. Motivação e problema

O objetivo inicial definido para este trabalho centrava-se na análise da relevância que os hábitos alimentares e a atividade física têm para os índices de obesidade em cada país. Pretendia-se, assim, avaliar de que forma estes dois fatores contribuem para a prevalência da obesidade em diferentes nações. No entanto, devido à dificuldade significativa em encontrar dados compatíveis em termos de escalas e medidas utilizadas, este objetivo inicial revelou-se difícil de concretizar.

Diante deste desafio, optou-se por redefinir o foco do estudo, alargando o seu âmbito para incluir não apenas os hábitos alimentares e a atividade física, mas também outros fatores potencialmente relevantes, mas mudando o escopo para apenas os países europeus. Assim, o novo objetivo do trabalho passou a ser a análise da influência que os restaurantes de fast food, os hábitos alimentares, a atividade física e o Produto Interno Bruto (PIB) per capita têm nos índices de obesidade dos países europeus.

Tendo em conta isto foram definidas 9 perguntas para responder com a análise feita aos dados:

- Há relação entre o consumo de bebidas alcoólicas e a obesidade nos países europeus?
- Países com maior nível de atividade física têm menores taxas de obesidade?
- O consumo de alimentos ultraprocessados, como snacks e sobremesas, está positivamente correlacionado com a obesidade?
- Quais grupos alimentares têm correlação negativa com a obesidade, ou seja, podem estar ligados a hábitos alimentares mais saudáveis?
- A densidade populacional influência a concentração de restaurantes fast food?
- Há diferença clara de perfil alimentar entre países com altas e baixas taxas de obesidade?
- A riqueza de um país (PIB) está correlacionada com a taxa de obesidade?
- A densidade populacional está associada com a taxa de obesidade?

Após responder a estas questões procura-se desenvolver um modelo capaz de prever o índice de obesidade de um novo país, para escolher o melhor modelo serão comparados os resultados de 3 modelos de regressão escolhidos.

2.1. Porque é relevante

O estudo analisa a obesidade e os seus fatores influenciadores. Após uma reavaliação dos dados disponíveis o estudo inicial foi redefinido para incluir outros fatores como a densidade de restaurantes fast food e o PIB, restringindo a análise à Europa.

A inclusão destes fatores adicionais permite uma análise mais abrangente. As perguntas de investigação abordam várias dimensões do problema, permitindo assim identificar correlações entre as variáveis.

Para além disto o estudo visa desenvolver um modelo preditivo para a obesidade, comparando 3 modelos de regressão. Que na falta de dados ajudaria a prever o índices de obesidade para qu

O estudo é relevante pois amplia o entendimento sobre os determinantes da obesidade. Adicionalmente, traz nova informação, uma vez que nenhuma pesquisa anterior foi encontrada tentando relacionar todos estes fatores simultaneamente.

3. Configuração dos dados e sistema

3.1. Tratamento e criação do dataset

O dataset final e completamente tratado é composto por vários países da europa e os seus respetivos valores para as colunas definidas que são:

Tabela 3-1 - Features do dataset tratado

Feature	Descrição
Population	Corresponde à população de cada país
Area_km2	Corresponde à área em quilómetros de cada país
BurgerKing_Count	Corresponde ao número de restaurantes Burger King presentes em cada país
McDonalds_Count	Corresponde ao número de restaurantes McDonalds presentes em cada país
Total_FastFood_Count	Corresponde ao número total de restaurantes fast food (BK e MAC) presentes em cada país
FastFood_per_100k	Corresponde ao número de restaurantes fast food por cada 100 mil habitantes de cada país
FastFood_density_per_1000km2	Corresponde à densidade de restaurantes fast food por cada mil quilómetros quadrados de cada país
Pop_Density	Corresponde à densidade populacional
Consumo_Eggs_and_egg_products	Quantidade de ovos e produtos derivados de ovos consumidos por cada país
Consumo_Fish_and_other_seafood_including_amphibians_rept	Consumo de peixes e outros frutos do mar, incluindo anfíbios e répteis por cada país
Consumo_Fruit_and_fruit_products	Quantidade de frutas e produtos à base de frutas consumidos por cada país
Consumo_Fruit_and_vegetable_juices	Consumo de sumos de frutas e vegetais por cada país
Consumo_Grains_and_grain-based_products	Quantidade de grãos e produtos à base de grãos consumidos por cada país

Consumo_Herbs_spices_and_condiments	Consumo de ervas, especiarias e condimentos por cada país
Consumo_Legumes_nuts_and_oilseeds	Quantidade de legumes, nozes e sementes oleaginosas consumidos por cada país
Consumo_Meat_and_meat_products_including_edible_offal	Consumo de carne e produtos cárneos por cada país
Consumo_Milk_and_dairy_products	Quantidade de leite e produtos lácteos consumidos por cada país
Consumo_Non-alcoholic_beverages_excepting_milk_based_beve	Consumo de bebidas não alcoólicas, excluindo as à base de leite por cada país
Consumo_Snacks_desserts_and_other_foods	Quantidade de snacks, sobremesas e outros alimentos consumidos por cada país
Consumo_Starchy_roots_and_tubers	Consumo de raízes e tubérculos por cada país
Consumo_Sugar_and_confectionary	Quantidade de açúcar e produtos de confeitaria consumidos por cada país
Consumo_Vegetables_and_vegetable_products_including_fungi	Consumo de vegetais e produtos à base de vegetais, incluindo fungos por cada país
Consumo_Alcoholic_beverages	Quantidade de bebidas alcoólicas consumidas por cada país
Consumo_Animal_and_vegetable_fats_and_oils	Consumo de gorduras e óleos de origem animal e vegetal por cada país
Consumo_Composite_food_including_frozen_products	Quantidade de alimentos compostos, incluindo produtos congelados, consumidos por cada país
Consumo_Drinking_water_water_without_any_additives_except	Consumo de água potável sem aditivos por cada país
GDP_2017	PIB de cada país
Obesity_Rate_Europe	Índice de obesidade de cada país
Physical_Activity_4plus_times	Índice de atividade física de cada país

Foi aqui que foi gasto grande parte do tempo dedicado a este projeto no tratamento e agrupamento de todos as informações e datasets num só, foi também por causa deste processo que foi redefinido o objetivo e escopo do trabalho.

As informações relativas aos restaurantes MacDonalds e Burger King foram extraídas diretamente das seguintes fontes *[1][2]*, não foi preciso grande tratamento dos dados para estes a não ser a remoção de colunas extra que não eram necessárias, apenas foram mantidas as colunas referentes à contagem de restaurantes por cada país. Estes dados são mais recentes relativamente com os restantes que são todos referentes a 2017, isto porque não foram encontrados dados mais antigos, no entanto foi possível verificar que apesar de os números terem aumentado não foi significativamente.

Os dados relativos à população, densidade populacional e área do país foram extraídos das fontes *[3]*, foram tomados os devidos cuidados relativamente à data destes dados pois são de 2017.

Foi através dos dados de contagem de restaurantes de MacDonals e Burger King e dos dados populacionais e de área de quadrada que foram calculadas as features: FastFood_density_per_1000km2, FastFood_per_100k e Total_FastFood_Count.

Todos os dados relativos à dieta dos países foram extraídos da seguinte fonte *[4]*. Este conjunto de dados exigiu um tratamento significativo, uma vez que continha muitas colunas que não eram relevantes para o âmbito deste trabalho. Além disso, existiam diversos questionários de diferentes anos para cada país, o que obrigou a uma seleção manual daqueles que faziam sentido utilizar. Dado que o nome desses questionários não era padronizado, optou-se por selecionar os dados referentes ao ano de 2017 e extrair apenas as linhas correspondentes.

Neste ficheiro, a coluna "Foodex L1" continha as categorias de alimentos que podem ser vistas nas colunas do conjunto de dados final. Esta coluna era padronizada, mas, como todas as categorias estavam numa única coluna, existiam diversas entradas para cada país. Para resolver este problema, estas classes foram pivotadas para se tornarem colunas no conjunto de dados final, eliminando assim a redundância e permitindo uma melhor análise.

Todos os dados relativos à PIB de cada país foram obtidos a partir da fonte *[5]*. Novamente foram extraídos apenas os dados referentes ao ano de 2017, para além disso foi necessário remover caracteres “:” que representavam dados em falta e foram removidas linhas que não faziam sentido, ou seja, que não eram referentes a países.

Em relação aos dados de índice de obesidade foram todos extraídos a partir da fonte *[6]*, novamente tendo o cuidado de extrair apenas os dados de 2017 e neste caso foi apenas extraída a coluna de *overweight*, esta coluna de dados tendo em conta o objetivo definido anteriormente representará a variável target do dataset.

Em relação aos dados de índices de atividade física estes foram extraídos a partir da seguinte fonte *[7]*, neste caso tiveram de ser usados dados mais recentes de 2022 pois não existiam dados referentes a 2017. Daqui foi extraída a coluna referente às pessoas que fazem exercício físico no mínimo 4 vezes por semana.

Por fim, o agrupamento de todas estas informações exigiu um grande esforço e a realização de várias etapas. Primeiramente, foram identificados e consolidados todos os países únicos presentes nos diferentes conjuntos de dados. De seguida, foram transformadas as categorias alimentares em colunas através de uma tabela dinâmica, como já fora referido anteriormente. Os nomes das colunas foram limpos e padronizados para facilitar a análise.

Posteriormente, os dados de consumo alimentar foram inseridos no dataset final. Foram adicionados também os dados do PIB, os dados relativos aos restaurantes de fast food, as taxas de obesidade na Europa e os níveis de atividade física ao dataset final.

Foram adicionados ainda os dados populacionais completos, incluindo a densidade populacional, população e a área em quilómetros quadrados. Por fim, o dataset foi reorganizado para colocar as colunas mais relevantes no início e foram removidos os países sem dados de obesidade ou com valores nulos. Países sem dados suficientes foram identificados e removidos. O dataset final foi ordenado por nome de país e guardado num ficheiro CSV para análise posterior.

Uma vez com o dataset completamente tratado foi iniciada a análise dos dados.

3.2. Sistema e métodos

Para obter estes resultados foi utilizado um ambiente de programação python, que realizou todas as operações de tratamento e agrupamento de dados na secção anterior. Além disto foi também através deste ambiente que foi feita a análise de dados e a criação dos modelos de regressão.

Foram utilizadas várias tecnologias, de entre elas principalmente as bibliotecas:

- Pandas: Esta biblioteca foi utilizada principalmente para a leitura dos diferentes datasets escolhidos para terem os seus dados extraídos. Para além disto foram utilizados métodos desta biblioteca para agrupar e concatenar os diferentes dados;
- Seaborn: Esta biblioteca foi apenas utilizada para uma função que foi a criação do heatmat através da matriz de correlação do dataset;
- Matplotlib: Assim como a biblioteca anterior esta apenas teve uma função que foi demonstrar o heatmap criado pelo seaborn;
- Scikit-learn: Esta biblioteca teve um grande papel no desenvolvimento dos modelos, pois forneceu métricas para classificar os diferentes modelos escolhidos, como R2 score, mean absolute error e root mean absolute error. Foi através desta biblioteca que os dados foram normalizados, e foi esta biblioteca que forneceu os modelos de regressão: Linear Regression Ridge Regression e Random Forest Regressor.

Após o tratamento dos dados como foi referido serão aplicados modelos de regressão preditivos da variável obesidade os escolhidos foram os seguintes:

Tabela 3-2 - Desrição de modelos utilizados

Modelo	Descrição
Linear Regression	A regressão linear é um modelo estatístico que tenta estabelecer uma relação linear entre uma variável dependente e uma ou mais variáveis independentes. É um dos métodos mais simples e amplamente utilizados para modelar a relação entre duas variáveis.
Ridge Regression	A regressão Ridge é uma técnica de regressão linear que inclui uma penalização de regularização. Esta penalização ajuda a reduzir a

	complexidade do modelo e a evitar overfitting, sendo especialmente útil quando há multicolinearidade entre as variáveis independentes.
Random Forest Regressor	O Random Forest Regressor é um modelo de ensemble learning que utiliza múltiplas árvores de decisão para melhorar a precisão da previsão e controlar o overfitting. Cada árvore é treinada em uma amostra aleatória dos dados, e a previsão final é a média das previsões de todas as árvores.

Para treinar estes modelos o dataset tratado foi dividido em dois, dataset de treino e dataset de teste, todos estes modelos tiveram as mesmas condições de treino e teste para as comparações serem justas.

4. Resultados

4.1. Análise de dados

Após a criação do dataset a partir das diversas fontes, foi iniciada a análise de resultados.

Para tal iniciou-se com o cálculo da matriz de correlação entre todas as variáveis presentes no mesmo e os resultados estão presentes na seguinte *Figura 4-1*.

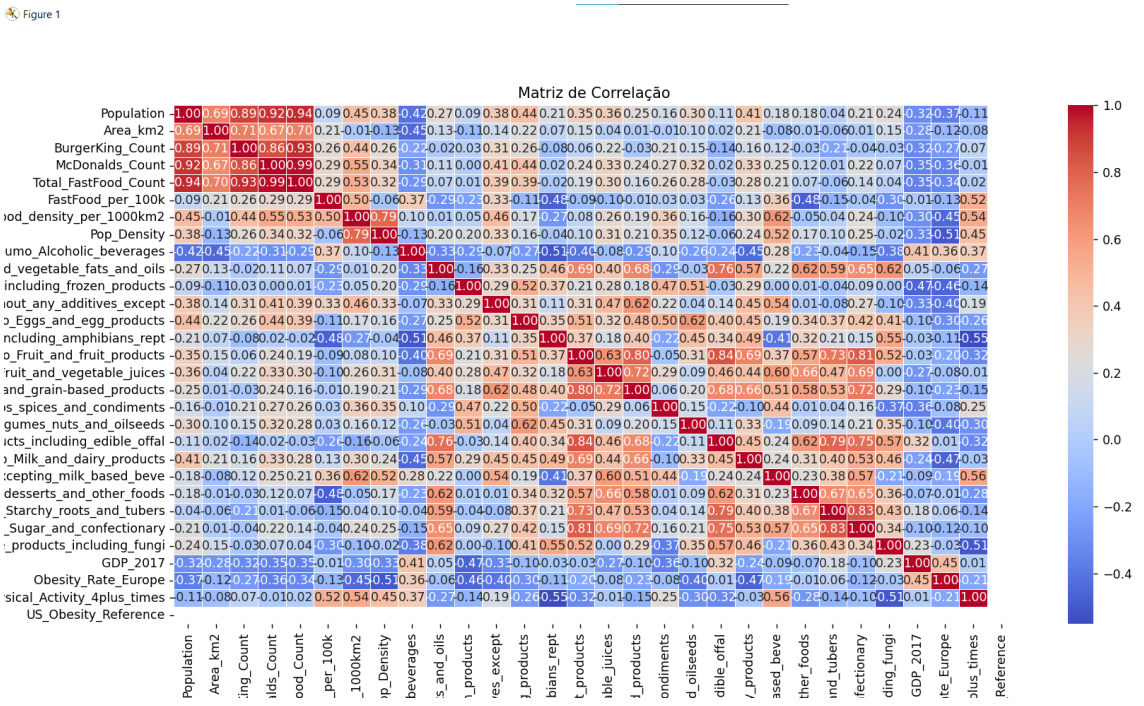

Figura 4-1 - Mapa de correlação

E foi possível encontrar que existem várias correlações entre as diferentes variáveis do dataset, e, portanto, serão destacadas as seguintes:

- Em relação ao índice de obesidade é possível verificar que os fatores que apresentam uma maior correlação são:
    - A população e consequentemente a densidade populacional sendo que este último tem uma moderada correlação (-0.51) com a diminuição destes índices e é o maior índice de correlação negativo presente no dataset.
    - A dieta alimentar de cada país pode sim ter uma influencia na diminuição destes índices como pode ser verificado pelo consumo de “Legumes nuts and oilseeds” que tem uma correlação negativa de -0.40, de “Drinking water, water without any additives” que tem uma correlação negativa de -0.40, de “Milk and dairy products” que tem a maior correlação negativa de -0.47 e por fim “Composite food including frozen products” com uma correlação negativa de -0.46
    - Foi ainda possível verificar as features que mais contribuíram para o aumento do índice de obesidade foram o PIB com um valor de correlação positiva de 0.45 e o consumo de “Alcoholic beverages”.
- Em relação ao PIB foi possível verificar as seguintes correlações:

    - O consumo de “Alcoholic Beverages” está correlacionado positivamente com esta feature, ou seja quanto mais rico for um país maior serão os índices de consumo de álcool que por sua vez também aumenta o índice de obesidade com observado na análise anterior.
- Em relação à atividade física foi possível verificar as seguintes correlações:
    - O consumo de “Non-alcoholic beverages excepting milk based beverages” tem uma correlação positiva com estes índices de 0.56. No entanto houve também outras bebidas que quando os índices de atividade física são altos também tendem a aumentar que são “Alcoholic beverages” com uma correlação de 0.37
    - No lado negativo foi possível verificar países que tendem a consumir mais “Fish and other seafood including amphibians reptils” tendem a ter índices de atividade física mais baixos, como pode ser verificado pela sua correlação de -0.55.
- Em relação ao restaurantes fast food, não foram encontradas correlações significativas a não ser obviamente com as métricas calculadas através destes dados, no entanto, foi possível verificar que sim o número de restaurantes fast food é correlacionado com a densidade populacional com uma correlação de 0.53.

Em baixo encontram-se diversos gráficos que permitem visualizar estas correlações encontradas no mapa de correlação.

- Figura 4-2 - Distribuição da taxa de obesidade : Esta figura demonstra a distribuição dos níveis de obesidade presentes no dataset, que só por si só já são bastante inquietantes uma vez que são tão altos e que apesar de haver índices mais baixos nas casas dos 40 a maior parte dos índices de obesidade encontram-se no lado direito do gráfico na casa dos 55;
- Figura 4-3 - Tipo de dietas vs Indice de obesidade : Neste gráfico é possível visualizar a correlação moderada que as diferentes dietas identificadas anteriormente têm com a taxa de obesidade e é mesmo possível verificar que sim conforme aumenta o consumo destas a taxa de obesidade tem alguma tendência a baixar;
- Figura 4-4 - Densidade populacional vs indice de obesidade : Aqui podemos verificar que sim a densidade populacional à medida que aumenta vai descendo também o índice de obesidade isto vai de acordo com os casos extremos que são vistos no dia a dia nas notícias que países com grandes densidades tendem a ter mais fome e consequentemente menos obesidade;
- Figura 4-5 - PIB vs consumo de bebidas alcoólicas : Aqui podemos verificar que consoante o PIB vai aumentando também vai aumentando, apesar de levemente, o consumo de bebidas alcoólicas, oque confirma a correlação moderada anteriormente discutida;
- Figura 4-6 - Atividade fisica vs tipo de dietas : Aqui podemos verificar que quanto mais atividades física são praticadas mais líquidos sejam alcoólicos ou não são ingeridos, para além disto é possível verificar que países que têm maior consumos de dietas baseadas em peixes tendem a fazer menos exercício físico;
- Figura 4-7 - Densidade populacional vs densidade de restaurantes Fast Food : Nesta figura é possível verificar que sim a densidade de restaurantes de Fast Food estará correlacionada moderadamente com a densidade populacional de um país, poís estes restaurantes procuram fazer dinheiro, inserindo mais restaurantes em países mais densos em termos de população.

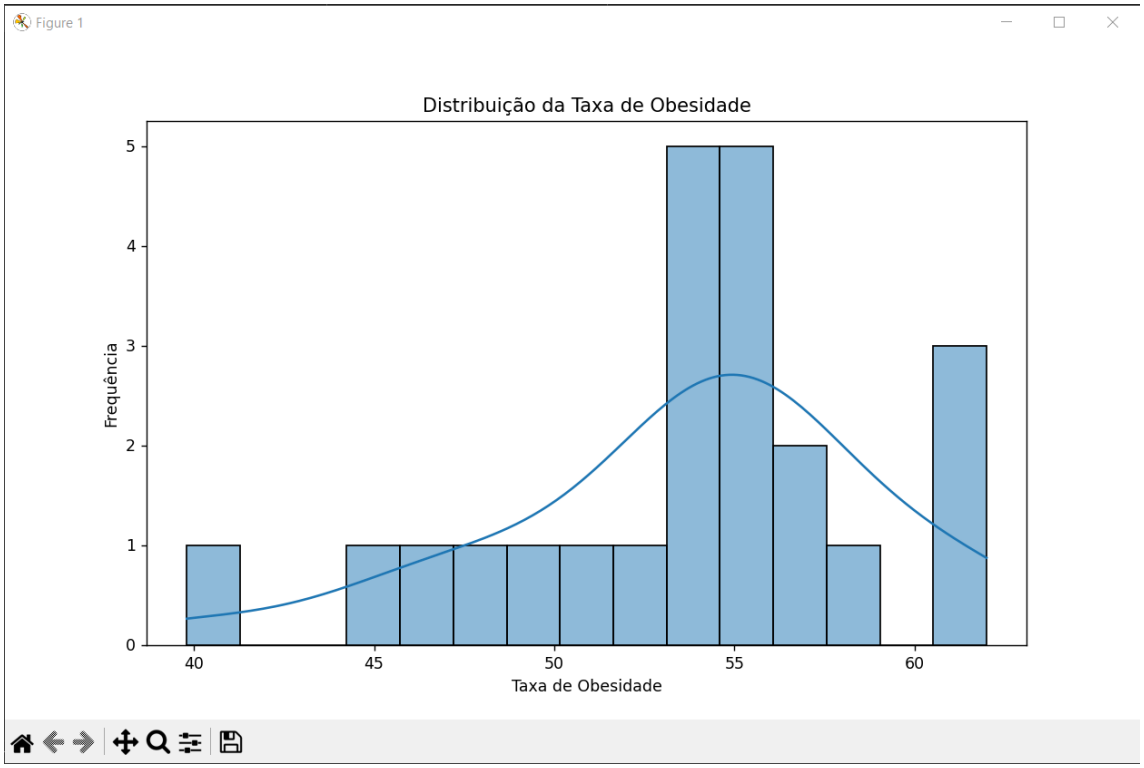

Figura 4-2 - Distribuição da taxa de obesidade

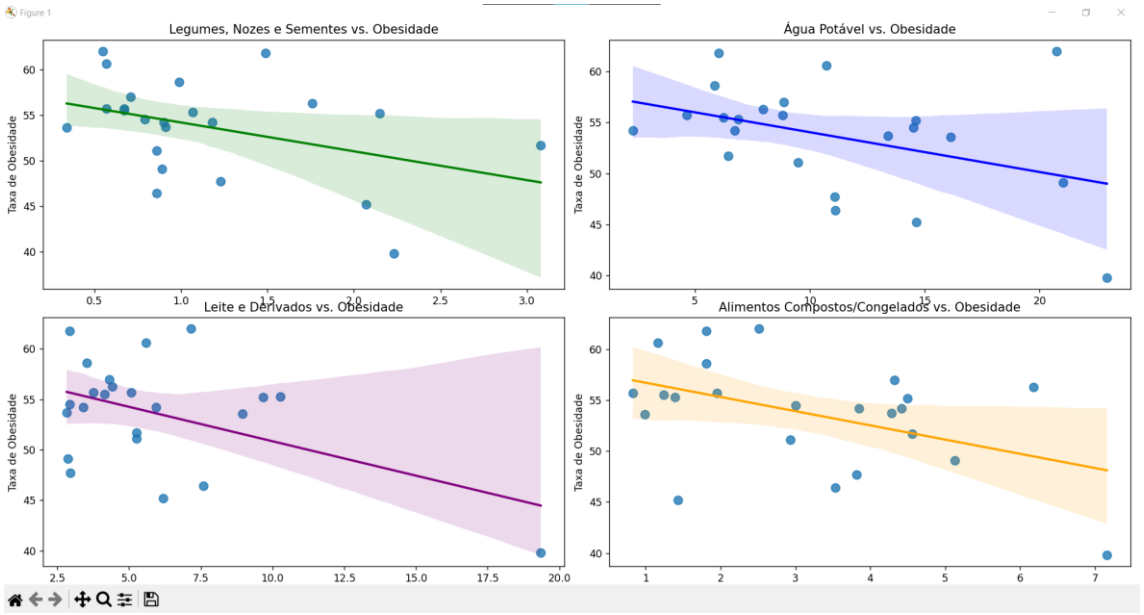

Figura 4-3 - Tipo de dietas vs Indice de obesidade

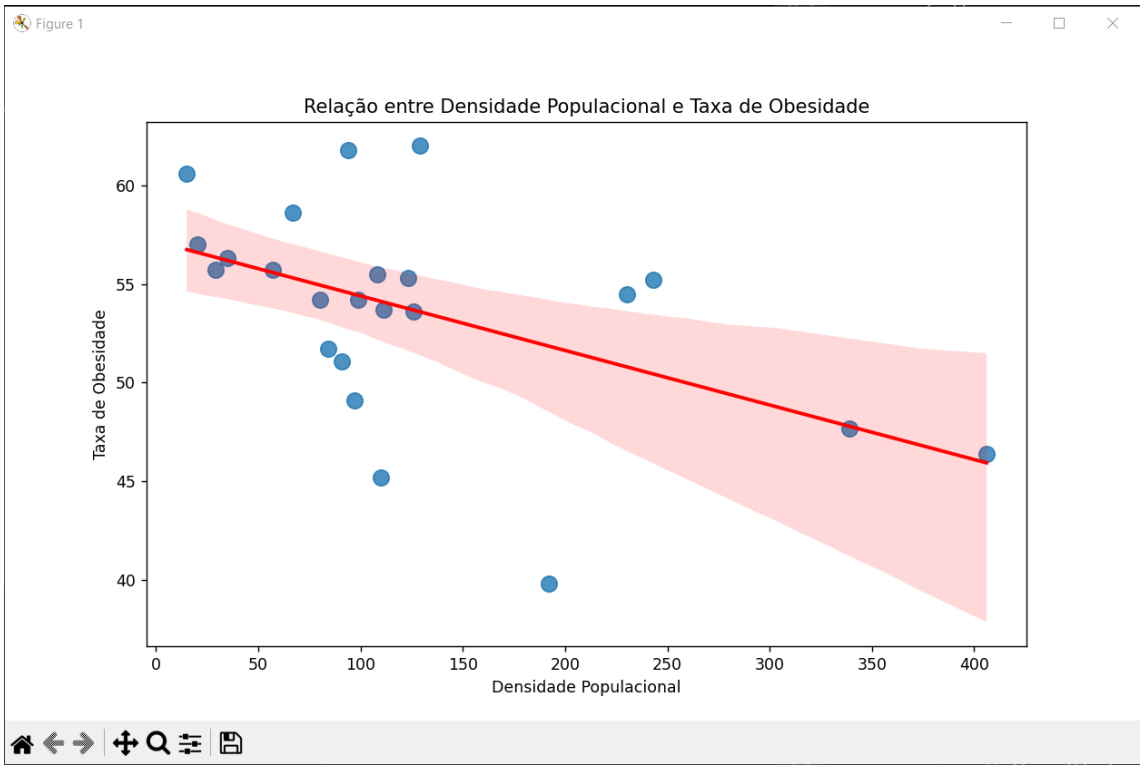

Figura 4-4 - Densidade populacional vs indice de obesidade

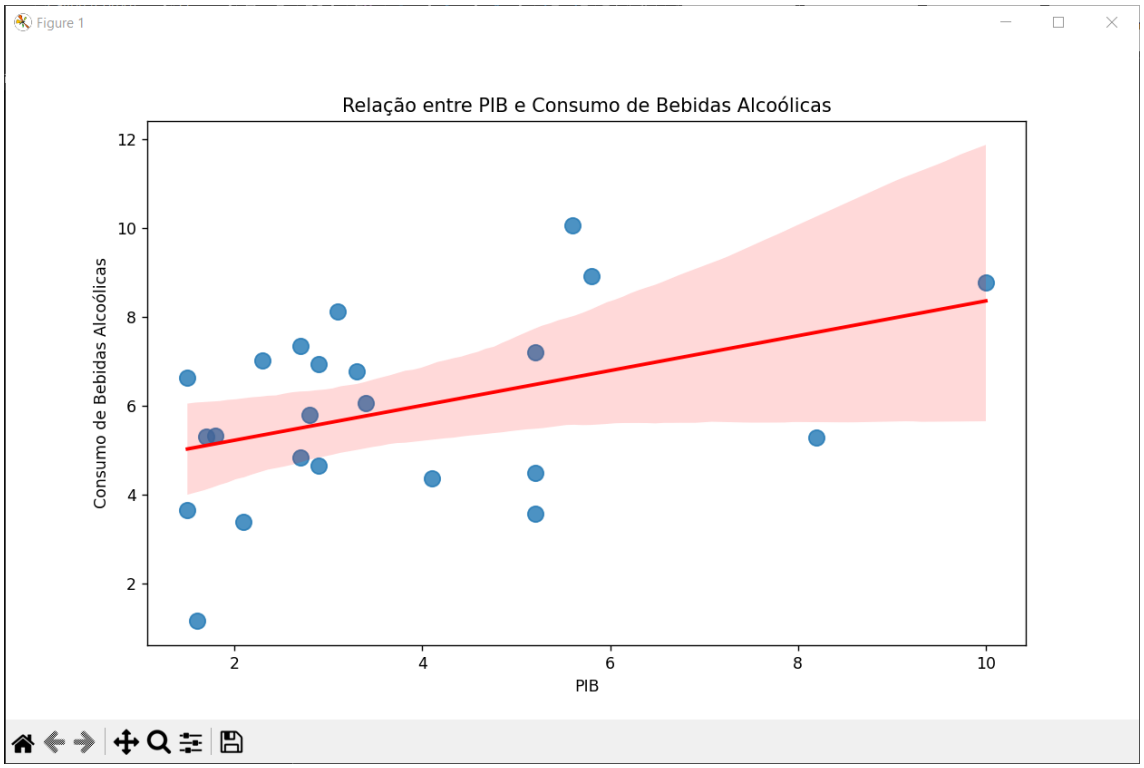

Figura 4-5 - PIB vs consumo de bebidas alcoólicas

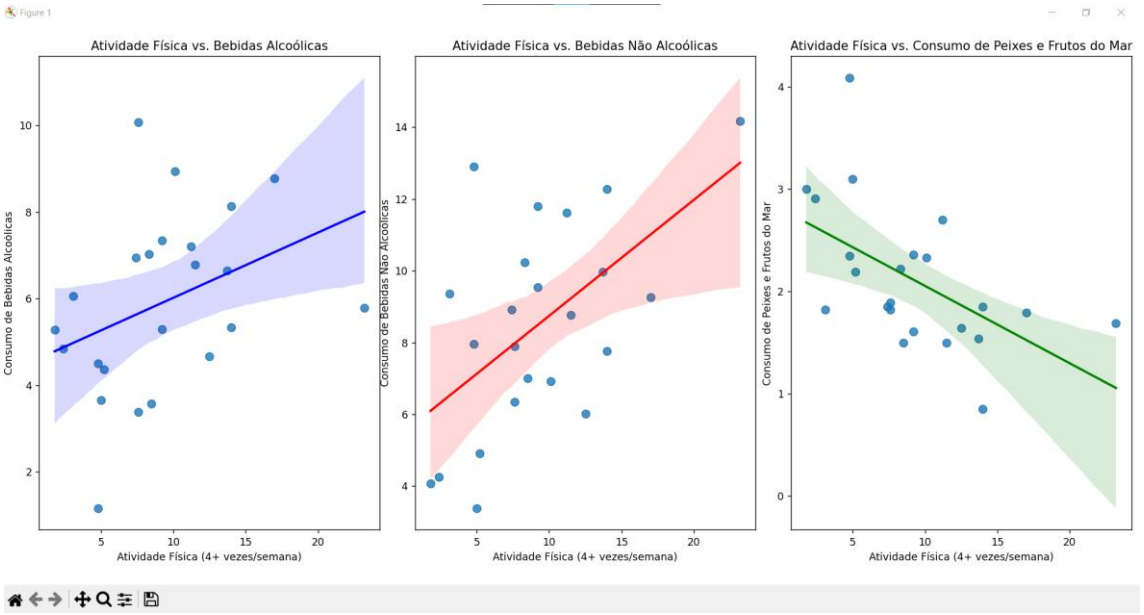

Figura 4-6 - Atividade física vs tipo de dietas

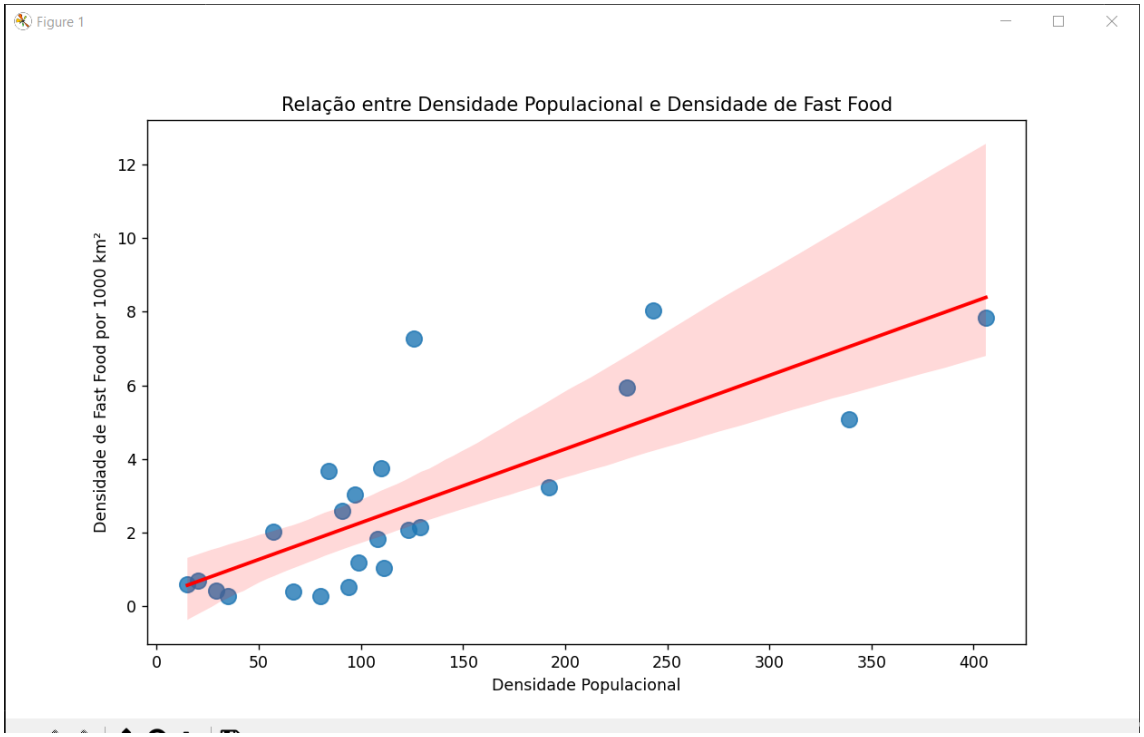

Figura 4-7 - Densidade populacional vs densidade de restaurantes Fast Food

4.2. Modelos preditivos

Foram então como já referidos anteriormente utilizados 3 modelos para comparar os seus resultados e definir o melhor a utilizar caso seja pretendido prever dados para um novo país.

Os dados foram então os seguintes:

Tabela 4-1 - Comparação de modelos preditivos

Métrica	Regressão Linear	Regressão ridge	Regressor Random Forests

R2 score	-0.77	-0.76	-0.48
MAE	7.20	7.20	6.43
RMAE	8.51	8.50	7.79

Como é possível verificar o modelo que melhor se saí por uma margem significativa é o modelo Random Forests Regressor. Obtendo em todas as métricas utilizadas valores superiores. Para uma melhor visualização do quanto este modelo erra *Figura 4-8*. Como é possível verificar existe ainda um grande erro em alguns casos, no entanto noutros houve um erro mínimo ou pequeno, oque significa que o modelo foi sim capaz de aprender alguns padrões nos dados. O facto de o modelo não ter uma performance boa como pode ser visto nas métricas utilizadas, pode dever-se ao facto de o dataset ter poucos dados, não sendo os suficientes para aprender melhor estes padrões presentes neles.

	Real	Previsto	Erro (Previsto - Real)	Erro Absoluto
0	46.4	55.904	9.504	9.504
1	54.5	55.263	0.763	0.763
2	49.1	54.833	5.733	5.733
3	45.2	54.735	9.535	9.535
4	61.8	54.956	-6.844	6.844
5	55.7	55.444	-0.256	0.256
6	47.7	55.469	7.769	7.769
7	39.8	55.680	15.880	15.880
8	53.6	57.648	4.048	4.048
9	58.6	54.640	-3.960	3.960

Figura 4-8 - Comparação de dados previstos vs reais, Random Forests Regressor

Para fins de comparação em baixo podem se encontrar as mesmas tabelas, mas para os outros dois modelos utilizados, Linear Regression e Ridge Regression respetivamente, *Figura 4-9* e *Figura 4-10*.

	Real	Previsto	Erro (Previsto - Real)	Erro Absoluto
0	46.4	57.683231	11.283231	11.283231
1	54.5	53.580345	-0.919655	0.919655
2	49.1	58.451483	9.351483	9.351483
3	45.2	52.505702	7.305702	7.305702
4	61.8	52.768010	-9.031990	9.031990
5	55.7	60.661657	4.961657	4.961657
6	47.7	54.173634	6.473634	6.473634
7	39.8	56.539046	16.739046	16.739046
8	53.6	58.747040	5.147040	5.147040
9	58.6	57.829483	-0.770517	0.770517

Figura 4-9 - Comparação de dados previstos vs reais, Linear Regression

	Real	Previsto	Erro (Previsto - Real)	Erro Absoluto
0	46.4	57.619582	11.219582	11.219582
1	54.5	54.289899	-0.210101	0.210101
2	49.1	58.401751	9.301751	9.301751
3	45.2	53.170455	7.970455	7.970455
4	61.8	52.240887	-9.559113	9.559113
5	55.7	59.298251	3.598251	3.598251
6	47.7	54.922623	7.222623	7.222623
7	39.8	55.910387	16.110387	16.110387
8	53.6	58.935674	5.335674	5.335674
9	58.6	57.128831	-1.471169	1.471169

Figura 4-10 - Comparação de dados previstos vs reais, Ridge Regression

5. Discussão

Neste capítulo serão respondidas todas as perguntas definidas anteriormente, todas as respostas serão feitas com base nos resultados obtidos a partir da análise feita.

Há relação entre o consumo de bebidas alcoólicas e a obesidade nos países europeus?

Sim como foi visto na análise dos resultados o consumo de bebidas alcoólicas está moderadamente relacionado com o nível de obesidade do país, este consumo está ainda correlacionado com o GDP de cada país, ou seja, quanto mais rico o país for maior será o consumo de álcool que aumentará por sua vez o índice de obesidade.

Países com maior nível de atividade física têm menores taxas de obesidade?

Não foram encontradas correlações significativas entre estes dois dados, no entanto, foi possível verificar que quanto mais atividade física é realizada mais bebidas alcoólicas e não alcoólicas são ingeridas estas por sua vez têm correlações positivas e negativas moderadas com o índice de obesidade, ou seja, não existem indícios nestes dados de que a maiores índices de atividade física diminuam a obesidade.

O consumo de alimentos ultraprocessados, como snacks e sobremesas, está positivamente correlacionado com a obesidade?

Não foi possível identificar quaisquer correlações significativas neste quesito, ou seja, tendo em contas os dados extraídos dos diferentes datasets europeus a maior ingestão deste tipo de dieta não está diretamente relacionado com os índices de obesidade.

Quais grupos alimentares têm correlação negativa com a obesidade, ou seja, podem estar ligados a hábitos alimentares mais saudáveis?

Após a análise foi possível verificar que os principais grupos alimentares que contribuem para a diminuição dos índices de obesidade são os seguintes: “Legumes, Nozes e sementes”, “Água potável”, “Leite e derivados” e “Alimentos compostos/congelados”. Ou seja, países com dietas alimentares mais saudáveis terão índices de obesidade mais baixos como comprovado em cima.

A densidade populacional influência a concentração de restaurantes fast food?

Sim a densidade populacional influência a concentração destes tipos de restaurantes, como foi explicado anteriormente. Isto acontece, pois, os restaurantes são empresas que procurar ganhar o máximo de dinheiro possível então em países onde existe muita densidade populacional serão onde estas empresas apostarão mais.

Há diferença clara de perfil alimentar entre países com altas e baixas taxas de obesidade?

Para verificar isto podemos pegar em dois exemplos de países com índices de obesidade baixos e altos e então foram escolhidos os seguintes, Roménia e Itália, e os aspetos que podemos destacar são os seguintes, a Roménia tem um consumo de álcool muito maior em comparação com o da Itália que como foi visto anteriormente contribuí para um índice de obesidade maior, já a Itália tem um consumo de Água, Frutas, sumos vegetais, grãos lacticínios e bebidas não alcoólica muito maiores que implica que a Itália tem uma dieta mais saudável e equilibrada, oque contribuí para os seus índices de obesidade mais baixos.

Tabela 5-1 - Comparação Romania vs Itália

Dieta	Roménia	Itália
Alc. Cons.	5.28	1.16
Anim/Veg Fats	0.75	0.92
Comp. Food	1.81	7.16
Water	6.05	22.92
Eggs	0.96	1.34
Fish	3.00	4.09
Fruit	3.36	7.27
Fruit/Veg Juices	0.68	6.36
Grains	2.96	6.23
Herbs	0.12	0.25
Legumes	1.49	2.23
Meat	2.93	3.33
Milk	2.93	19.35
Non-Alc.	4.07	7.96
Snacks	1.42	2.00
Starchy	2.88	3.22
Sugar	0.45	0.84
Veg	5.58	4.37
Obesity Rate (%)	**61.8**	**39.8**

A riqueza de um país (PIB) está correlacionada com a taxa de obesidade?

Sim na realidade é o fator que mais contribui para os índices de obesidade neste dataset. Oque demonstra que sim quanto mais rico um país é mais obeso os seus cidadãos tendem também a o ser. Isto pode ser justificado pelo mais fácil acesso a alimentos mais calóricos.

A densidade populacional está associada com a taxa de obesidade?

Sim a densidade populacional está correlacionada com a taxa de obesidade de forma negativa. Oque significa que quanto mais denso em nível de população o país é, menores são as chances de obesidade.

6. Conclusão

Após toda a análise dos dados e toda a discussão dos resultados considero que sim existem vários fatores neste dataset, que foi criado, que estão correlacionados e influenciam os índices de obesidade dos diferentes países.

Foi descoberto que existem hábitos alimentares estão moderadamente correlacionados com a diminuição destes índices de obesidade, nomeadamente o consumo de água, lacticínios e legumes contribuem para uma diminuição destes valores, por outro lado o consumo de bebidas alcoólicas está relacionado com o aumento da obesidade.

Foi possível verificar que países com um PIB maior tendem também a ter maiores índices de obesidade, isto pode acontecer pois as suas populações terão acesso a maios alimentos calóricos.

Foi possível verificar que a atividade física não está diretamente correlacionada com a descida da obesidade, isto pois aumenta a ingestão de bebidas alcoólicas e não alcoólicas que por sua vez têm correlações positivas e negativas, respetivamente.

Para além disto foi verificado que quanto maior for a densidade populacional de um país menor serão os índices de obesidade do mesmo.

Foi verificado que dos modelos comparados aquele que melhor resultado apresenta é o “Random Forests Regressor”, no entanto dado a pequena dimensão dos dados ele não conseguiu aprender suficientemente rápido os padrões presentes nos dados. Pois como foi visto anteriormente houve instâncias que este modelo classificou bastante bem, mas outras que obteve um erro bastante grande.

Este trabalho mostrou-me a importância da pesquisa e tratamento de dados, se pretendemos obter um trabalho significativo e com um bom produto final. Fazendo com que eu aprendesse competências na área da pesquisa e tratamento dos mesmos.

Fui obrigado a redefinir objetivos e datasets de onde pretendia extrair os dados por haver incompatibilidade entre eles, eu pretendia inicialmente acrescentar os Estados Unidos nesta análise, mas os dados de dieta que estes disponibilizam eram muito diferentes dos europeus que mesmo apesar de todo o esforço feito para converter as classes de alimentação europeias em nutrientes (formato utilizado pelos estados unidos) não foi possível o realizar com sucesso por causa das unidades utilizadas.

Caso seja necessário verificar os datasets utilizados estes estarão todos no seguinte repositório “https://github.com/McDoritos/SGD_Data-Analysis”

Por fim penso que foram obtidos conhecimentos e dados interessantes e relevantes para a área com este dataset e a sua ánalise. Considerando então que o trabalho foi um sucesso.

7. Referências

[1] 'Most McDonald's by Country 2025'. Accessed 6 June 2025. https://worldpopulationreview.com/country-rankings/most-mcdonalds-by-country#sources.

[2] 'List of Countries with Burger King Franchises'. In *Wikipedia*, 1 June 2025. https://en.wikipedia.org/w/index.php?title=List_of_countries_with_Burger_King_franchises&oldid=1293441477.

[3] World Bank Open Data. 'World Bank Open Data'. Accessed 6 June 2025. https://data.worldbank.org.

[4] 'Data.Europa.Eu'. Accessed 6 June 2025. https://data.europa.eu/data/datasets/the-efsa-comprehensive-european-food-consumption-database?locale=en.

[5] '[Tec00115] Real GDP Growth Rate - Volume'. Accessed 6 June 2025. https://ec.europa.eu/eurostat/databrowser/view/tec00115/default/table?lang=en.

[6] '[Ilc_hch10] Person Distribution by Body Mass Index, Educational Attainment Level, Sex and Age'. Accessed 6 June 2025. https://ec.europa.eu/eurostat/databrowser/view/ilc_hch10__custom_16964456/default/table?lang=en&page=time:2022.

[7] '[Ilc_scp03] Persons Participating in Cultural or Sport Activities in the Last 12 Months by Sex, Age, Educational Attainment, Activity Type and Frequency'. Accessed 6 June 2025. https://ec.europa.eu/eurostat/databrowser/view/ilc_scp03$dv_551/default/table?lang=en&category=sprt.sprt_pcs.sprt_pcs_ilc.