Análise comparativa de múltiplos classificadores na detecção de diabetes

Gustavo Portela Rautenberg[1], **Ronaldo D. F. Pachico**[1]

[1]Centro de Ciências Exatas e da Terra – Universidade Estadual do Oeste do Paraná (UNIOESTE)
Cascavel – PR – Brazil

{gprautenberg,ronaldodreckslerfp}@gmail.com

Abstract. *This paper presents the first project from the 2024 Machine Learning course (Csc3040), focusing on comparing the accuracy of various classification algorithms applied to Diabetes Classification. Models such as K-Nearest Neighbors (KNN), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) were analyzed, as well as classifier combination techniques like sum, majority voting, and Borda Count. Statistical analysis using Kruskal-Wallis and Mann-Whitney tests assessed the significance of accuracy differences between the models.*

Resumo. *Este artigo apresenta o primeiro projeto da disciplina de Aprendizado de Máquina (Csc3040) de 2024, focado na comparação de acurácia entre diferentes algoritmos classificadores aplicados à Classificação de Diabetes. Foram analisados modelos como K-Nearest Neighbors (KNN), Árvore de Decisão (AD), Naive Bayes (NB), Support Vector Machine (SVM) e Multilayer Perceptron (MLP), além de técnicas de combinação de classificadores, como soma, voto majoritário e Borda Count. A análise estatística, utilizando Kruskal-Wallis e Mann-Whitney, verificou a significância das diferenças na acurácia entre os modelos.*

1. Introdução

A diabetes é uma das doenças crônicas mais prevalentes no mundo, impactando uma grande parcela da população global e gerando consequências significativas tanto para a saúde pública quanto para a economia [Hu 2011]. Trata-se de uma condição em que os portadores de diabetes perdem a capacidade de regular eficazmente os níveis de glicose no sangue, o que pode levar a complicações graves, como doenças cardíacas, perda de visão, amputação de membros e doenças renais, resultando em uma redução significativa da qualidade e da expectativa de vida [Deshpande et al. 2008]. A importância do diagnóstico precoce é evidente, pois ele permite intervenções mais eficazes, tanto em termos de mudanças no estilo de vida quanto de tratamentos médicos, tornando os modelos preditivos para o risco de diabetes ferramentas essenciais para os profissionais de saúde pública.

Com base nessas considerações, o presente trabalho foca na análise de qual técnica de aprendizado de máquina consegue alcançar a melhor acurácia entre diversos algoritmos classificadores aplicados a um conjunto de dados relacionado à diabetes. O conjunto de dados utilizado possui três classes: diabético, pré-diabético e não diabético. Assim, o objetivo é identificar qual estratégia atinge as melhores métricas para a correta classificação dessas categorias.

2. Dataset

O conjunto de dados é composto pela pesquisa telefônica anual realizado pelo CDC (Centers of Disease Control and Prevention) nos Estados Unidos, que coleta informações sobre comportamentos de risco à saúde, condições crônicas e uso de serviços preventivos entre os habitantes do país. A pesquisa tem o nome de ”Behavioral Risk Factor Surveillance System (BRFSS) e compreende a respostas de 441.455 indivíduos e com mais de 300 variáveis. [Teboul 2021]

Entretanto, para este projeto foi utilizado um dataset disponível no Kaggle, referente ao ano de 2015. Este conjunto de dados, corresponde somente 253.680 respostas da pesquisa presente, o qual foi normalidade, todas as variáveis, com exceção a classe, foram normalizadas. A variável alvo possui três classes: 0 para não-diabético, 1 para pré-diabético e 2 para diabético, pode se observar a distribuição das classes na Figura 1 além de contemplar somente 21 variáveis, uma descrição delas pode ser visualizada na Tabela 1, a Figura 2, revela-nos que a correlação entre as variáveis é baixa, sendo que alguns dos poucos casos onde ela tem uma maior correlação é na pressão alta e idade, e dificuldade de andar, saúde e a auto avaliação da saúde.

Para esse trabalho foi usado apenas uma subamostra do dataset, correspondente a dez porcento da base original, mantendo a proporcionalidade das classes conforma ilustrado na Figura 1. Sendo divididos a cada interação em 50% para treino, 25% para validação e 25% para teste, os três subconjuntos mantendo a proporcionalidade das classes.

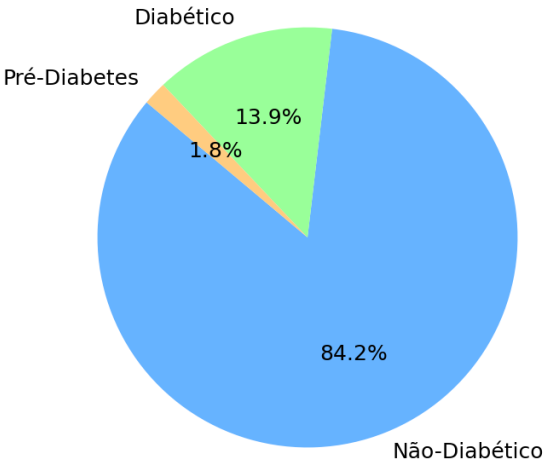

Figura 1. Gráfico de distribuição das variáveis no Datasets.

3. Problema

O objetivo principal é construir e avaliar diversos modelos de classificação para prever a condição de diabetes dos indivíduos com base nessas variáveis. Para isso, foram em-

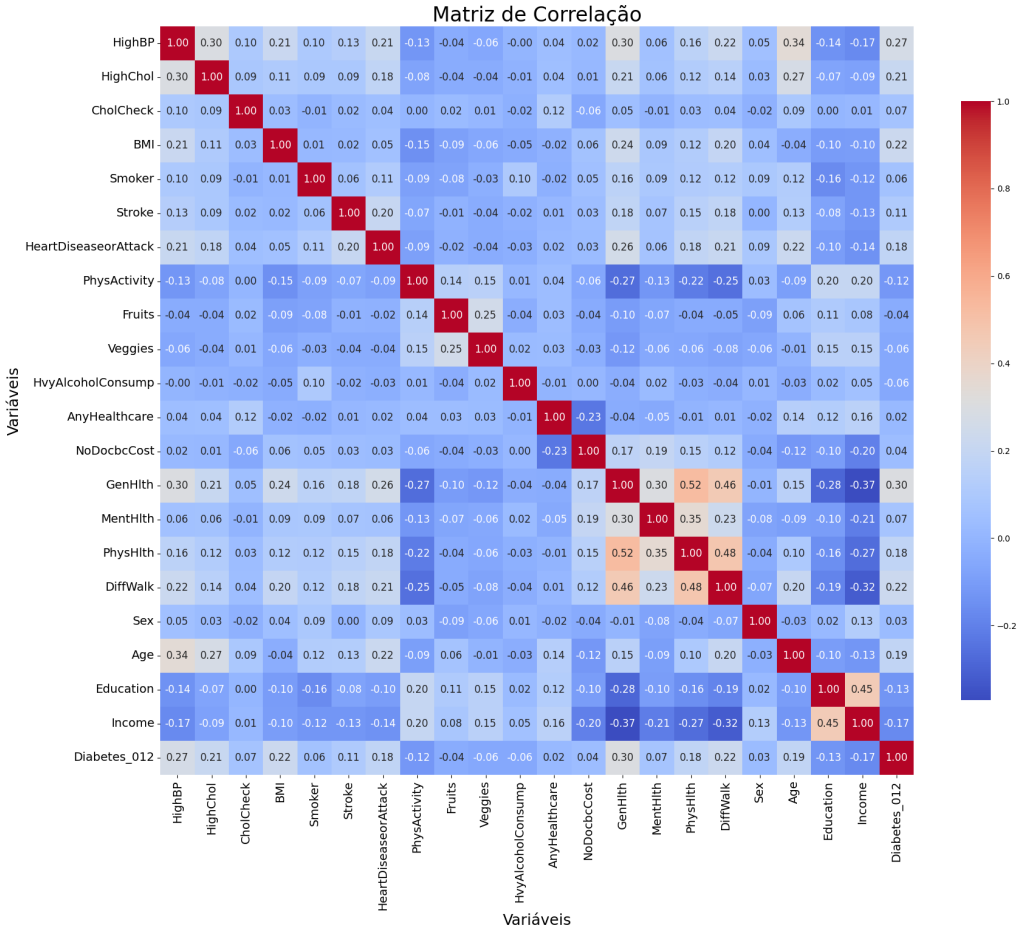

Figura 2. Matriz de correlação entre as variáveis do Dataset.

pregados vários classificadores, incluindo K-Nearest Neighbors (KNN), Árvores de Decisão(AD), Naive Bayes (NB), Multlayer Perceptron (MLP) e Support Vector Machines (SVM). Foi verificado o comportamento de multiclassificadores, sendo as regras da soma, voto majoritário e da borda count.

4. Classificadores

Esse trabalho avaliou 5 classificadores monólitos: KNN, AD, NB, MLP e SVM [Pedregosa et al. 2011], foi realizado o processo de `GridSeach` – busca em grade –, o qual testa todas as possíveis variações das dos hiperparâmetros de cada classificador, visando otimizar a acurácias no conjunto de dados de validação, os parâmetros que foram submetidos aos testes, assim como os valores que eles puderam assumir estão descritos nas sub-sessões adiante.

4.1. K-Nearest Neighbors (KNN)

O classificador KNN faz previsões baseadas na proximidade dos dados, seus vizinhos, em um espaço de características. A busca em grade ajusta os seguintes parâmetros:

- **Número de Vizinhos (`n_neighbors`):** Variando de 1 a 50.
- **Métrica de Ponderação (`weights`):** ”distance”(pondera com base na distância) ou ”uniform”(ponderação uniforme, não importando a distância do vizinho).

Tabela 1. Descrição das variáveis

Variável	Descrição	Valores
HighBP	Pressão alta	0 = Sem pressão alta, 1 = Pressão alta
HighChol	Presença de problema com colesterol	0 = Sem colesterol alto, 1 = Colesterol alto
CholCheck	Checagem do colesterol nos últimos 5 anos	0 = Não, 1 = Sim
BMI	Índice de massa corporal (IMC)	Valor float
Smoker	Fumou ao menos 100 cigarros durante a vida	0 = Não, 1 = Sim
Stroke	Teve AVC	0 = Não, 1 = Sim
HeartDisAttck	Doença cardíaca coronária (DCC) ou infarto do miocárdio (IM)	0 = Não, 1 = Sim
PhysActivity	Atividade física nos últimos 30 dias, sem incluir o trabalho	0 = Não, 1 = Sim
Fruits	Consumo de frutas 1 ou mais vezes por dia	0 = Não, 1 = Sim
Veggies	Consumo de vegetais 1 ou mais vezes por dia	0 = Não, 1 = Sim
HvyAlcoholCon	Adultos que consomem mais de 14 doses por semana e mulheres adultas que consomem mais de 7 doses por semana	0 = Não, 1 = Sim
AnyHealthcare	Possui qualquer tipo de cobertura de saúde, incluindo seguro de saúde, planos pré-pagos	0 = Não, 1 = Sim
NoDocbcCost	Algum momento nos últimos 12 meses em que precisou consultar um médico, mas não pôde devido ao custo	0 = Não, 1 = Sim
GenHlth	Como classificaria sua saúde em uma escala de 1-5	1 = Excelente, 2 = Muito bom, 3 = Bom, 4 = Justa, 5 = Precária
MentHlth	Saúde mental nos últimos 30 dias (estresse, depressão, emoções)	Escala de 1 a 30 dias
PhysHlth	Saúde física nos últimos 30 dias (doenças, lesões)	Escala de 1 a 30 dias
DiffWalk	Dificuldade de caminhar ou subir escadas	0 = Não, 1 = Sim
Sex	Sexo	0 = Feminino, 1 = Masculino
Age	Categoria de idade de 13 níveis	1 a 13
Education	Nível de educação	Escala de 1 a 6
Income	Escala de renda	Escala de 1 a 8

4.2. Árvore de Decisão

A Árvore de Decisão divide os dados em várias regiões baseadas em regras de decisão. A busca em grade ajusta os seguintes parâmetros:

- **Critério (`criterion`):** ”entropy”ou ”gini”.
- **Profundidade Máxima (`max_depth`):** Limita a profundidade da árvore. Variando de 1 à 11.
- **Número Mínimo de Amostras por Folha (`min_samples_leaf`):** Número mínimo de amostras em uma folha da árvore. Variando de 1 à 11.
- **Número Mínimo de Amostras para Divisão (`min_samples_split`):** Número mínimo de amostras necessárias para dividir um nó. Variando de 2 à 16.
- **Estratégia de Divisão (`splitter`):** ”best”ou ”random”.

4.3. Naive Bayes

O classificador Naive Bayes é um modelo probabilístico baseado no Teorema de Bayes, assumindo independência entre características. O modelo é treinado com os parâmetros padrão e avaliado no conjunto de teste.

4.4. Perceptron Multi-camadas (MLP)

O MLP é uma rede neural que pode aprender representações complexas dos dados. A busca em grade ajusta os seguintes parâmetros:

- **Arquitetura (`hidden_layer_sizes`):** Tamanhos das camadas ocultas, foi testado todas as possíveis combinações entre uma e duas camadas ocultas, cada uma podendo ter 21, 27, 33 e 39 neurônios.
- **Taxa de Aprendizado (`learning_rate`):** Pode ser ”constant”, ”invscaling”ou ”adaptive”.
- **Número Máximo de Épocas (`max_iter`):** Número máximo de épocas para o treinamento. Variando de 50, 100, 150 e 200.
- **Função de Ativação (`activation`):** Pode ser ”identity”, ”logistic”, ”tanh”ou ”relu”.

4.5. Support Vector Machine (SVM)

O SVM busca encontrar o melhor hiperplano que separa as classes. A busca em grade ajusta os seguintes parâmetros:

- **Função do Kernel (`kernel`):** Pode ser ”linear”, ”poly”, ”rbf”ou ”sigmoid”.
- **Parâmetro de Regularização (`C`):** Controla o trade-off entre maximizar a margem e minimizar o erro de treinamento. Variando de 0.1, 1 e 10.

5. Agregação das Previsões

Após treinar e avaliar os modelos, o código realiza três tipos de agregação das previsões:

- **Soma das Probabilidades:** As probabilidades previstas por cada classificador são somadas e a classe com a maior soma é escolhida.
- **Voto Majoritário:** A classe mais votada entre os classificadores é escolhida como a previsão final.
- **Borda Count:** As previsões de cada classificador são ranqueadas e a classe com o maior total de ranks é escolhida.

Tabela 2. Acurácia dos Classificadores Monolíticos

Repetição	KNN	NB	MLP	AD	SVM
1	0.8423	0.0946	0.8360	0.8312	0.8423
2	0.8502	0.3691	0.8454	0.8360	0.8265
3	0.8391	0.1120	0.8407	0.8423	0.8360
4	0.8454	0.0899	0.8249	0.8312	0.8391
5	0.8407	0.1593	0.8517	0.8344	0.8423
6	0.8438	0.0883	0.8533	0.8391	0.8328
7	0.8407	0.1751	0.8312	0.8233	0.8423
8	0.8438	0.7713	0.8486	0.8344	0.7902
9	0.8438	0.0868	0.8391	0.8360	0.8423
10	0.8407	0.1136	0.8391	0.8423	0.8218
11	0.8438	0.1215	0.8454	0.8265	0.8423
12	0.8407	0.2050	0.8360	0.8249	0.8454
13	0.8407	0.2271	0.8423	0.8233	0.8344
14	0.8423	0.0804	0.8328	0.8312	0.8438
15	0.8438	0.1088	0.8454	0.8375	0.8344
16	0.8486	0.1909	0.8438	0.8407	0.8407
17	0.8438	0.3675	0.8360	0.8375	0.8423
18	0.8470	0.1215	0.8391	0.8423	0.8360
19	0.8423	0.3407	0.8391	0.8375	0.8423
20	0.8407	0.1498	0.8249	0.8155	0.8139
Média(DP)	0.843(0.003)	0.199(0.163)	0.840(0.008)	0.833(0.007)	0.835(0.013)

6. Resultados e Discussão

6.1. Modas dos Hiperparâmetros

Além da analise da acurácia, foi realizado uma analise da dos valores que mais apareceram nos hiperparâmetros de cada classificador, de forma a observar a tendencia a qual eles estavam assumindo, o resumo disso pode ser encontrado na Tabela 4. Pode-se observar que o MLP segue um tendencia de um rede altamente complexa, sendo sua arquitetura próxima do máximo testado – 42 neurônios em ambas camadas ocultas.

6.2. Análise Comparativa

Para a análise comparativa, é necessário comparar a média da acurácia obtida em vinte execuções, que pode ser visualizada na Tabela 2 para os classificadores monolíticos e na Tabela 3, os multi, utilizando o teste de Kruskal-Wallis para verificar se há diferença significativa entre as estratégias baseadas em múltiplos classificadores. Caso seja identificada uma diferença significativa (rejeição da hipótese nula, H0), deve-se aplicar o teste de Mann-Whitney para realizar comparações par-a-par, a fim de determinar quais pares de classificadores apresentam diferenças estatisticamente relevantes.

6.2.1. Monolítica vs Monolítica

A análise estatística conduzida para comparar o desempenho entre diferentes classificadores monolíticos revelou resultados significativos com base no teste Kruskal-Wallis, com

Tabela 3. Acurácia dos Multiclassificadores

Repetição	Soma	Voto Majoritário	Borda Count
1	0.8391	0.8423	0.7918
2	0.8375	0.8391	0.4937
3	0.8470	0.8454	0.7792
4	0.8438	0.8454	0.7839
5	0.8454	0.8470	0.7382
6	0.8517	0.8454	0.7729
7	0.8375	0.8407	0.6924
8	0.8438	0.8423	0.1798
9	0.8438	0.8438	0.7839
10	0.8375	0.8423	0.7461
11	0.8502	0.8438	0.7539
12	0.8391	0.8470	0.6672
13	0.8344	0.8423	0.6498
14	0.8407	0.8423	0.8028
15	0.8407	0.8438	0.7634
16	0.8407	0.8486	0.6909
17	0.8423	0.8423	0.4968
18	0.8438	0.8438	0.7603
19	0.8423	0.8423	0.5237
20	0.8375	0.8328	0.7129
Média(DP)	0.842(0.004)	0.843(0.003)	0.679(0.152)

Tabela 4. Modas dos parâmetros dos classificadores ao longo das 20 interações

Classificador	Parâmetros e Modas
KNN	K: 36 distance: uniform
AD	criterion: entropy max_depth: 5 min_samples_leaf: 3 min_samples_split: 10 splitter: random
SVM	kernel: poly C: 1
NB	- -
MLP	hidden_layer_sizes: [39, 33] activation: relu max_iter: 50 learning rate: constant

uma Estatística H de 61.42 e um Valor-p extremamente baixo (1.46e-12). Estes resultados indicam que rejeitamos a hipótese nula (H0), sugerindo que há pelo menos um classificador com desempenho significativamente diferente dos demais. O teste de Mann-Whitney foi então utilizado para identificar pares de modelos com diferenças estatisticamente sig-

nificativas, que pode ser visualizado na Tabela 5, os resultados revelam que, para a maioria das comparações, como KNN vs AD, KNN vs NB, KNN vs SVM, AD vs NB, NB vs SVM, e NB vs MLP, as diferenças são estatisticamente significativas, com valores de p muito baixos (0.000). Esses resultados sugerem que os desempenhos desses classificadores são significativamente distintos entre si, o que é corroborado com o resultado da média da acurácia disponível na Tabela 2, a qual revela que o NB tem o pior resultado – 0.199 de acurácia –, enquanto todos os demais então em todos dos 0.84, sendo o melhor o KNN, obtendo 0.843 de precisão com um desvio padrão de apenas 0.003.

Por outro lado, comparações como KNN vs MLP, AD vs SVM e SVM vs MLP não apresentaram diferenças significativas, com valores de p superiores ao nível de significância adotado (0.05), indicando que, para esses pares, os desempenhos dos classificadores não diferem de forma significativa.

Tabela 5. Testes Mann-Whitney dos classificadores monolíticos

Comparação	Estatística U	Valor-p	Significativa
KNN vs AD	370.0	0.000	Sim
KNN vs NB	400.0	0.000	Sim
KNN vs SVM	306.5	0.004	Sim
KNN vs MLP	264.5	0.081	Não
AD vs NB	400.0	0.000	Sim
AD vs SVM	145.0	0.137	Não
AD vs MLP	108.5	0.014	Sim
NB vs SVM	0.0	0.000	Sim
NB vs MLP	0.0	0.000	Sim
SVM vs MLP	161.5	0.301	Não

6.3. Multiclassificadores vs Multiclassificadores

A comparação entre diferentes estratégias de combinação de classificadores foi realizada utilizando-se o teste de Kruskal-Wallis, o que revelou a existência de pelo menos uma estratégia com desempenho significativamente diferente das outras. Esta descoberta levou à aplicação do teste de Mann-Whitney para comparações par-a-par, aprofundando a análise sobre quais estratégias realmente se destacam em termos de acurácia.

Primeiramente, a comparação entre a estratégia de ”Soma”e o ”Voto Majoritário”não indicou uma diferença significativa (valor-p: 0.1756), sugerindo que ambas as abordagens podem ser igualmente eficazes no contexto analisado. Isso implica que, dependendo das necessidades específicas do problema, qualquer uma dessas estratégias poderia ser utilizada sem comprometer a performance.

Por outro lado, a comparação entre a estratégia de ”Soma”e a de ”Borda Count”mostrou uma diferença altamente significativa (valor-p: 6.54e-08), indicando uma diferença do ”Borda Count”em relação à ”Soma”.

O teste de Kruskal-Wallis seguido pelo teste de Mann-Whitney identificou que a estratégia ”Borda Count”é significativamente diferente das outras, mas não afirma que ela é a melhor em termos de acurácia. A significância estatística indica que os resultados obtidos com ”Borda Count”são consistentes e distintamente diferentes das outras estratégias, observando a Tabela 3, vê se que o resultado média de sua acurácia é o pior do três.

Tabela 6. Testes Mann-Whitney para diferentes comparações

Comparação	Estatística U	Valor-p	Diferença Significativa
Soma vs Voto Majoritário	150.0	1.755955e-01	Não
Soma vs Borda Count	400.0	6.541131e-08	Sim
Voto Majoritário vs Borda Count	400.0	6.135377e-08	Sim

6.3.1. Monolítica vs Multiclassificadores

A análise comparativa entre o melhor modelo monolítico e a melhor estratégia de múltiplos classificadores mostra que modelo monolítico com o melhor desempenho foi o KNN, com uma média de acurácia de 0.843, enquanto a melhor estratégia de múltiplos classificadores foi o Voto Majoritário, com uma média de acurácia muito próxima de 0.843. Para avaliar a significância das diferenças de desempenho entre esses dois modelos, foi realizado o teste estatístico de Mann-Whitney. O valor da estatística U foi 181.0, com um valor-p de 0.610. Como o valor-p é superior ao nível de significância de 0.05, não há evidências suficientes para rejeitar a hipótese nula. Isso indica que não há diferenças estatisticamente significativas entre o desempenho do KNN e do Voto Majoritário, sugerindo que ambos os métodos oferecem desempenhos comparáveis em termos de acurácia.

7. Conclusão

Os resultados mostraram que os classificadores KNN e SVM apresentaram a melhor acurácia média, seguidos pelo MLP, enquanto o Naive Bayes obteve o pior desempenho, já entro os múltiplos o voto majoritário, obteve a maior acurária, 0.843, estaticamente igual ao do KNN, apesar do desse resultado não apresentar ainda um grande taxa de falhas, os modelos poderiam ser usado para uma indicação inicial da necessidade da realização de exames laboratoriais.

Continuações naturais do presente trabalho são a realização um seleção melhor das variáveis usadas, assim como um ampliação da quantia neurônios no MLP, visto que esse ficou quase no limite do testado, podendo ser que em números maiores o resultado poderia melhor. Além de outra alternativa, aumentar a quantidade de amostras nas classes menos representadas como a pré-diabetes, seria uma possível estratégia para o equilíbrio de aprendizado em cenários de desequilíbrio de classes. Ao fornecer mais exemplos das classes com menor frequência, o modelo pode ou não se tornar mais sensível e capaz de reconhecer padrões relevantes dentro dessa classe, por sua vez, melhorando a capacidade de generalização do classificador. A utilização de técnicas de oversampling, pode contribuir para a equilibrar a distribuição das classes.

Referências

Deshpande, A., Harris-Hayes, M., and Schootman, M. (2008). Epidemiology of diabetes and diabetes-related complications. *Physical Therapy*, 88:1254 – 1264.

Hu, F. (2011). Globalization of diabetes. *Diabetes Care*, 34:1249 – 1257.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Teboul, A. (2021). Diabetes health indicators dataset.