Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

HEVANS VINICIUS PEREIRA

Estudo de Preditores da Evolução de Pacientes Internados por COVID-19

MARINGÁ, PR
2023

Sumário

Introdução **2**
1 Referencial Teórico **4**
1.1 Regressão Logística 4
1.2 Floresta Aleatória 5
1.3 Rede Neural 6
2 Metodologia **8**
2.0.1 Critérios de Seleção da Amostra 8
2.0.2 Tratamento dos Dados 9
2.0.3 Análise Exploratória de Dados 10
3 Resultados Preliminares **13**

Referências **14**

Introdução

A pandemia de COVID-19 é um evento histórico sem precedentes que afetou a saúde e a vida de milhões de pessoas em todo o mundo. Desde o seu surgimento em dezembro de 2019 (HUANG C.; WANG, 2020), o novo coronavírus se espalhou rapidamente pelo globo, levando a medidas de contenção sem precedentes, como o distanciamento social, o fechamento de fronteiras e a interrupção de atividades econômicas. O impacto da COVID-19 foi sentido em todas as esferas da sociedade, desde a saúde pública até a economia, a educação e as relações sociais.

A pandemia de COVID-19 chegou ao Brasil em fevereiro de 2020, quando o país registrou seu primeiro caso confirmado da doença. Desde então, a COVID-19 se espalhou rapidamente por todo o território nacional, levando o país a se tornar um dos epicentros mundiais da pandemia.

Em março de 2020, o governo federal decretou estado de emergência em saúde pública de importância nacional em razão da pandemia de COVID-19. Desde então, várias medidas foram tomadas em todo o país para conter a disseminação do vírus, como a imposição de medidas de distanciamento social, o fechamento de escolas e comércios não essenciais, a restrição de viagens e a proibição de aglomerações.

No entanto, a implementação dessas medidas foi desigual em todo o país, com alguns estados e municípios adotando estratégias mais rigorosas e outros sendo mais lenientes. Além disso, o país enfrentou uma série de desafios na gestão da pandemia, como a escassez de equipamentos de proteção individual para profissionais de saúde, a falta de testes em massa e a falta de coordenação entre os governos federal, estaduais e municipais.

Neste trabalho iremos utilizar dados públicos, obtidos no Open Data SUS, sobre Síndrome Respiratória Aguda Grave e COVID para criar modelos de classificação que nos permitam estimar quais pacientes tem mais chance de vir a óbito com base em diversas características coletadas.

Para isso, consideramos apenas os casos de pacientes adultos, isto é, com 18 anos ou mais, hospitalizados por Covid-19 e que foram notificados no ano de 2022.

O questionário que monitora a Síndrome Respiratória Aguda Grave já existia antes da COVID, mas foi alterado em virtude desta. Essa mudança nos obriga a tomar um cuidado adicional na limpeza dos dados.

Os modelos serão criados usandos técnicas de aprendizado supervisionado com o objetivo de predizer o desfecho de um paciente internado com COVID-19. Com o objetivo de avaliar a capacidade preditiva dos modelos usaremos métricas de avaliações que envolvam sensibilidade e especificidade, como a acurácia, f1 score e área sob a curva ROC, entre outras que forem julgadas pertinentes.

Além do modelo preditivo em si, haverá um ganho de informação se conseguirmos extrair do modelo as informações referentes à importância das variáveis, a fim de identificar os fatores que mais influenciam no desfecho dos pacientes. Trabalhos semelhantes estão se tornando mais comuns (DABBAGH, 2023) (SILVADARCY RISOMARIO; NETO, 2022) e um entendimento dos modelos aumentará a usabilidade dos mesmos nas mais diversas áreas, trazendo benefícios e aumentando a responsabilidade de seus usos.

Capítulo 1

Referencial Teórico

Vejamos um breve histórico de alguns modelos de classificação bem como alguma explicação sobre o funcionamento dos mesmos. Neste trabalho iremos abordar a regressão logística, floresta aleatória e redes neurais.

1.1 Regressão Logística

A regressão logística foi desenvolvida como um modelo de crescimento populacional em uma série de artigos de Pierre François Verhulst, um matemático belga, que foram publicados entre 1838 e 1845.

Primeiramente Verhulst introduziu o conceito de função logística (VERHULST, 1838) e posteriormente mostrou como ajustar a curva à um conjunto de pontos (VERHULST, 1845).

A função logística, também chamada de sigmóide por seu gráfico lembrar um "S", é definida por $f : \mathbb{R} \to (0, 1)$ em que $f(t) = \dfrac{1}{1 + e^{-t}}$ e foi redescoberta posteriormente e de forma independente por outras pessoas.

A regressão logística foi usada originalmente em problemas de classificação binária, isto é, quanto temos uma variável que assume apenas dois valores, habitualmente chamados de 0 (zero) representando fracasso e 1 (um) representando sucesso; embora possa ser adaptada para classificação com mais de dois valores e para outros cenários. Embora tenha originalmente aparecido nos trabalhos de Verhulst a regressão logística se popularizou como modelo estatístico com Joseph Berkson (BERKSON, 1944).

Na função logística, se t é combinação linear de alguma outra variável independente x, então $t = \beta_0 + \beta_1 x$, logo $p(x) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$. Neste caso, $p(x)$ é interpretado como a probabilidade da variável dependente Y ser igual a 1. Para o caso de múltiplas variáveis

independentes temos $t = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$ e $p(x) = \dfrac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{m} \beta_i x_i)}}$

A estimação dos coeficientes β_i é feita via Máxima Verossimilhança e a vantagem deste modelo é que é possível realizar testes de hipóteses para verificar a significância dos valores estimados, bem como obter intervalos de confiança e interpretabilidade dos coeficientes, embora a interpretação dos coeficientes não seja tão direta como na regressão linear.

Uma das vantagens da regressão logística é que está permite interpretabilidade do modelo. Para interpretar os coeficientes de regressão, utilizamos a Razão de Chances (RC), que é a razão entre a chance de ter um efeito sobre a chance de não ter o efeito, isto é, ela calcula quantas vezes a chance de ter um efeito é maior do que a chance de não ter esse efeito. Lambrando que a chance é calculada como a probabilidade de ocorrência sobre a probabilidade de não ocorrência, isto é, calcula quantas vezes a probabilidade de ter um efeito é maior ou menor do que a probabilidade de não ter o efeito. Valores de chance próximos de 1 indicam que as probabilidades de ter efeito e de não ter efeito são praticamente as mesmas.

Na regressão logística e^{β_i} representa a mudança na razão de chances do sucesso para uma mudança unitária no x_i. Por exemplo, se $e^{\beta_i} = 0,5$ e $x_i = 2$, então as chances de ocorre sucesso é $0,5$ vezes a chance de sucesso para $x_i = 1$.

1.2 Floresta Aleatória

Proposto em 1995 (HO, 1995) esse modelo apresenta o nome de floresta pois é construído com base em vários modelos chamados árvores de decisão. Vamos ter um primeiro entendimento de como funcionam as árvores de decisão e então passaremos para a floresta aleatória.

Uma árvore de decisão é um modelo construído com base em condicionais, isto é, se algo acontecer decide-se por uma ação, caso contrário decide-se por outra. Por exemplo, se uma pessoa quer ir à praia apenas em dias ensolarados e sem vento e encontra-se em um momento de tomada de decisão sobre ir ou não à praia. Neste caso, se não estiver ensolarado a pessoa não irá à praia; se estiver ensolarado então deve-se observar se há vento, pois não havendo vento irá à praia e havendo vento não irá à praia.

Costumamos tomar decisões deste tipo todos dias, e a forma da tomada de decisão pode ser apresentada de forma gráfica, lembrando uma árvore. Em geral, desenha-se um nó a partir do qual partem dois caminhos e cada caminho se ramifica novamente e assim por diante.

Qual é a diferença do nosso processo de tomada de decisão para uma árvore de decisão? Quando tomamos decisões nós costumamos estabelecer os critérios em que haverá bifurcações

com relação à tomada de decisão, e quando usamos uma árvore de decisão os critérios são escolhidos pelo modelo de forma a conseguir fazer a melhor separação de classe possível com base em algum critério como Índice de Gini ou entropia.

Agora que entendemos intuitivamente como funciona uma árvore de decisão vamos partir para a construção da floresta. Como o nome sugere iremos usar várias árvores, mas as árvores não podem ser iguais, caso contrário todas as árvores iriam ter o mesmo comportamento e o conceito de floresta deixaria de fazer sentido. Como podemos criar árvores diferentes para um mesmo conjunto de dados? Há mais de uma forma de fazer isso e todas costumam ser utilizadas.

Primeiramente, cada árvores será criada usando apenas algumas das variáveis e essa escolha é feita aleatoriamente. Existe a possibilidade de aleatoriamente duas árvores usarem as mesma variáveis, principalmente se houverem muitas árvores, então cada árvore é construída usando apenas parte das observações e esse processo também é construído aleatoriamente. Dessa forma, é pouco provável que duas árvores sejam exatamente iguais e, mesmo se forem, seriam apenas duas dentre muitas e, portanto, não afetando o desempenho da floresta como um todo.

A classificação feita por uma árvore é um procedimento relativamente simples. Basicamente a árvore usa as condições que definem os nós para tomar decisões do tipo "se condição, então ação"de modo a partir do nós central e ir descendo a árvore até as folhas. Quando se chega em uma folha a árvore classifica de acordo com a classe que compõe a maioria naquela folha.

Agora que entendemos como ela é construída precisamos entender como a decisão é tomada. Em problemas de classificação pode-se atribuir classificação à uma variável tomando a classificação mais frequente para cada árvore, numa espécie de votação, e em problemas de regressão pode-se considerar as médias dos valores de cada árvore.

1.3 Rede Neural

Criada para imitar o funcionamento do cérebro humano por Warren McCulloch e Walter Pitts em 1943 (MCCULLOCH, 1943) a rede neural considera vários neurônios artificiais agrupados em camada que são interligadas. A figura 1.3.1 ilustra uma rede neural do tipo *Multi Layer Perceptron*, semelhante à que será usada neste trabalho.

Na figura 1.3.1, temos a camada de entrada à esquerda e a camada de saída à direita, todas as demais são chamadas de camadas ocultas. O termo Aprendizagem Profunda (*Deep Learning*, em inglês) refere-se à redes neurais com muitas camadas ocultas, em geral mais do que três, embora os modelos que deram popularidade às redes neurais e estão muito conhecidos

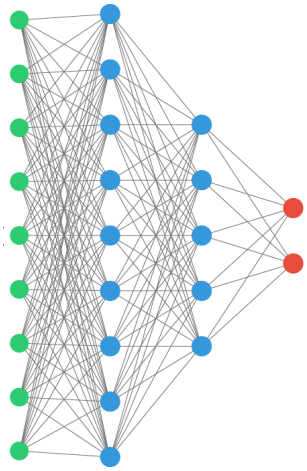

Figura 1.3.1 – Rede Neural

atualmente, como o ChatGPT (BASTIAN, 2023), tenham dezenas de camadas, cada uma com centenas de neurônios, chegando a mais de 175 bilhões de parâmetros.

Cada neurônio da camada de entrada vai assumir um valor de uma variável, os neurônios seguintes recebem combinações lineares dos valores de entrada vezes os pesos de cada neurônio da camada anterior e aplica-se a esse valor uma função conhecida como função de ativação. Também há uma constante chamada viés que é somada à combinação linear.

Esses pesos e o viés são números que são ajustados de modo que a previsão da rede seja a mais próxima possível do desejado. Os pesos são inicializados de forma alatória e são os parâmetros que devem ser aprendidos. Para que os melhores valores possíveis sejam encontrados para os parâmetros da rede, parte dos dados são apresentados à rede e então é feita uma previsão, a qual é comparada ao resultado esperado. A partir desse erro de previsão é usada uma técnica conhecida como gradiente descendente para otimizar uma função de erro, também chamada de função custo, e então atuailiza-se os parâmetros da rede. Com os novos parâmetros o processo é repetido até que a função de erro atinja um mínimo. Esse processo é chamado de treinamento da rede e cada vez que a rede utiliza todo o conjunto de dados para atualizar seus pesos é chamado de época de treinamento.

Usualmente, uma rede é treinada com muitas épocas e há ainda muitos parâmetros que poderiam ser acrescentados e/ou alterados para criar uma rede neural.

Capítulo 2

Metodologia

Neste trabalho usou-se a linguagem Python (versão 3.11.3) para analisar dados públicos sobre Síndrome Respiratória Aguda Grave disponibilizado no site do Open Data SUS.

A base de dados utilizada contém inúmeras variáveis, conforme dicionário de dados disponibilizado no site do Open Data SUS. As variáveis utilizadas serão apresentadas na Tabela 1.

Tais variáveis contém apenas informações sobre o paciente, comorbidades e sintomas observados.

A variável EVOLUCAO é a variável alvo, pois ela contém informação sobre óbito ou cura do paciente.

2.0.1 Critérios de Seleção da Amostra

Antes de qualquer outra coisa, precisamos fazer um trabalho inicial de limpeza e pré processamento dos dados, além de uma análise exploratória. Para a variável EVOLUCAO descartou-se as observações que contenham "9 - Ignorado" pois não queremos imputar a variável alvo. Além disso, tratou-se as observações "2-Óbito" e "3-Óbito por outras causas" como "Óbito" Desse modo, o desfecho do estudo é uma variável binária, indicando a cura ou óbito do paciente.

Decidiu-se por analisar apenas pacientes internados, portanto a variável HOSPITAL inicialmente selecionada será desconsiderada, pois todos os pacientes analisados terão o valor 1 para essa variável tornando irrelevante em análises posteriores.

Iremos trabalhar somente com adultos, por isso selecionaremos a faixa de idade acima de 18 anos e iremos desconsiderar a variável TP_IDADE pois esta indica apenas se a idade

é medida em anos, dias ou meses. Decidiu-se também trabalhar somente com pacientes que foram diagnosticados com COVID-19, portanto descartaremos a variável CLASSI_FIN após a seleção adequada.

2.0.2 Tratamento dos Dados

A base de dados possui muitos valores faltantes e muitas observações marcadas com "9 - Ignorado"para fatores associados, portanto decidiu-se considerar que tais observações indicam ausência do fator observado. Dessa forma, substituiu-se dados faltantes e as observações não registradas para todas as variáveis de fatores associados (puérpera, doença cardiovascular crônica, doença hematológica crônica, síndrome de Down, doença hepática crônica, asma, diabetes mellitus, doença neurológica crônica, outra pneumopatia crônica, imodeficiência ou imunosupressão, doença renal crônica, obesidade ou outros fatores) pelo valor 0.

Mesmo com tais escolhas para preencher dados faltantes, não temos como imputar valores para as variáveis CS_ESCOL_N (escolaridade), CS_SEXO, CS_RACA (raça), CS_ZONA (zona urbana/rural) e VACINA_COV (se a pessoa se vacinou para a COVID-19) então vamos trocar observações marcadas como ignorado para valores nulos.

A variável SUPORT_VEN possuia originalmente três categorias (”1-Sim, invasivo”, ”2-Sim, não invasivo”, ”3-Não”). Vamos juntar as observações ”1-Sim, invasivo” e ”2-Sim, não invasivo” em uma única categoria.

Também mudou-se a variável SG_UF_INTE que indica unidade federativa da internação do paciente para trabalharmos apenas com as cinco regiões do país, de acordo com a classificação do IBGE.

Após esses tratamentos iniciais verificou-se que na variável CS_ZONA não há observações para Rural, pois provavelmente essas observações foram descartadas em algum tratamento anterior, portanto vamos descartar toda essa variável.

Para a variável CS_RACA vamos considerar pretos e pardos como uma única categoria. Para a variável VACINA_COV vamos considerar que valores ignorados indicam falta de vacinação.

Por fim, vamos descartar todas as observações que ainda possuem dados faltantes. Isso nos deixa com pouco mais de 57000 observações, sendo um número suficiente para objetivos de encontrar modelos de classificação em momentos posteriores deste trabalho.

2.0.3 Análise Exploratória de Dados

Após essa primeira limpeza dos dados, contatou-se que a maioria das variáveis está desbalanceada quando consideradas em relação à variável EVOLUCAO e realizou-se o teste Qui Quadrado de Pearson para verificar associação entre as variáveis categóricas e a variável EVOLUCAO.

Na Tabela 1 encontram-se a quantidade de de valores para cada variável, considerando se a variável EVOLUCAO foi 0 ou 1, e também o p valor associado ao teste de hipótese qui quadrado de cada variável.

Conforme pode ser visto na Tabela 1, com exceção de Síndrome de Down, obesidade e fadiga, todos as demais variáveis apresentam-se relevantes ao nível de 5% de significância.

Tabela 1 – Frequências absolutas (relativas, %) e p-valor do teste qui-quadrado de Pearson.

Variável	Categoria	Cura	Óbito	Total	p valor
Sexo	Masc.	17762 (62)	10804 (38)	28566	<0,001
	Fem.	19351 (68)	9099 (32)	28450	
Raça	Branca	25600 (66)	13362 (34)	38962	<0,001
	Preta/Parda	11094 (64)	6315 (36)	17409	
	Amarela	355 (65)	193 (35)	548	
	Indígena	64 (66)	33 (34)	97	
Escolaridade	Analfabeto	10573 (62)	6559 (38)	17132	<0,001
	Fundamental	12237 (61)	7811 (39)	20048	
	Médio	9586 (71)	3902 (29)	13470	
	Superior	4717 (74)	1631 (26)	6348	
Nosocomial	Não	35410 (65)	18751 (35)	54161	<0,001
	Sim	1703 (60)	1152 (40)	2855	
Febre	Não	19921 (63)	11535 (37)	31456	<0,001
	Sim	17192 (67)	8368 (33)	25560	
Tosse	Não	11548 (59)	8112 (41)	19660	<0,001
	Sim	25565 (68)	11791 (32)	37356	
Evolução	Cura	28704 (62)	17408 (38)	46112	<0,001
	Óbito	8409 (77)	2495 (23)	10904	
Dispneia	Não	16478 (76)	5116 (24)	21594	<0,001
	Sim	20635 (58)	14787 (42)	35422	
Desconforto Respiratório	Não	20798 (73)	7589 (27)	28387	<0,001
	Sim	16315 (57)	12314 (43)	28629	
Saturação $O_2 < 95\%$	Não	18202 (77)	5337 (23)	23539	<0,001
	Sim	18911 (56)	14566 (44)	33477	
Diarreia	Não	33568 (65)	18129 (35)	51697	0.013
	Sim	3545 (67)	1774 (33)	5319	
Vômito	Não	33640 (65)	18393 (35)	52033	<0,001
	Sim	3473 (70)	1510 (30)	4983	
Outros Sintomas	Não	24897 (64)	14257 (36)	39154	<0,001
	Sim	12216 (68)	5646 (32)	17862	
Puérpera	Não	36691 (65)	19859 (35)	56550	<0,001
	Sim	422 (91)	44 (9)	466	
Doença Cardiovascular Crônica	Não	23427 (68)	10802 (32)	34229	<0,001
	Sim	13686 (60)	9101 (40)	22787	
Doença Hematológica Crônica	Não	36636 (65)	19541 (35)	56177	<0,001
	Sim	477 (57)	362 (43)	839	

Variável	Categoria	Cura	Óbito	Total	p valor
Síndrome de Down	Não	36990 (65)	19831 (35)	56821	0.61
	Sim	123 (63)	72 (37)	195	
Doença Hepática Crônica	Não	36716 (65)	19421 (35)	56137	<0,001
	Sim	397 (45)	482 (55)	879	
Asma	Não	35938 (65)	19454 (35)	55392	<0,001
	Sim	1175 (72)	449 (28)	1624	
Diabetes mellitus	Não	28112 (67)	14072 (33)	42184	<0,001
	Sim	9001 (61)	5831 (39)	14832	
Doença Neurológica Crônica	Não	34488 (66)	17590 (34)	52078	<0,001
	Sim	2625 (53)	2313 (47)	4938	
Outra Pneumopatia Crônica	Não	37777 (66)	17966 (34)	55743	<0,001
	Sim	2336 (55)	1937 (45)	4273	
Imunodepressão	Não	35203 (66)	18200 (34)	53403	<0,001
	Sim	1910 (53)	1703 (47)	3613	
Doença Renal Crônica	Nâo	35111 (66)	17978 (34)	53089	<0,001
	Sim	2002 (51)	1925 (49)	3927	
Obesidade	Não	35058 (65)	18734 (35)	53792	0.10
	Sim	2055 (64)	1169 (36)	3224	
Outros Fatores de Risco	Não	24626 (68)	11429 (32)	36055	<0,001
	Sim	12487 (60)	8474 (40)	20961	
Região do Brasil	Sul	9245 (70)	3830 (30)	13075	<0,001
	Sudeste	18861 (63)	11225 (37)	30086	
	Centro Oeste	2886 (70)	1246 (30)	4132	
	Norte	2578 (66)	1309 (34)	3887	
	Nordeste	3543 (61)	2293 (39)	5836	
Suporte Ventilatório	Não	35321 (74)	12413 (26)	47734	<0,001
	Sim	1792 (20)	7490 (80)	9282	
Dor Abdominal	Não	34063 (65)	18621 (35)	52684	<0,001
	Sim	3050 (70)	1282 (30)	4332	
Fadiga	Não	27527 (65)	14860 (35)	42387	0.20
	Sim	9586 (66)	5043 (34)	14629	
Perda do Olfato	Não	35694 (65)	19409 (35)	55103	<0,001
	Sim	1419 (74)	494 (26)	1913	
Perda do Paladar	Não	35559 (65)	19360 (35)	54919	<0,001
	Sim	1554 (74)	543 (26)	2097	
Recebeu Vacina COVID-19	Não	6487 (61)	4077 (39)	10564	<0,001
	Sim	30626 (66)	15826 (34)	46452	

Capítulo 3

Resultados Preliminares

Utilizou-se os modelos de classificação regressão logística, floresta aleatória e rede neural. As métricas utilizadas e os valores obtidos para cada modelo podem ser encontrados na tabela 2. Resultados similares foram encontrados por (SILVADARCY RISOMARIO; NETO, 2022) em uma base de dados diferente sobre a COVID-19 no Brasil.

Tabela 2 – Métricas de Avaliação dos Modelos

Modelo	Acurácia	Recall	Precision	f1 score	ROC-AUC
Regressão Logística	0,73	0,70	0,75	0,72	0,73
Floresta Aleatória	0,74	0,69	0,76	0,72	0,74
Rede Neural	0,74	0,73	0,74	0,73	0,74

Notamos, na Tabela 2, que as três técnicas utilizadas apresentam um poder preditivo muito similar, com a acurácia acima de 70%. Para a sequência deste trabalho, consideraremos os seguintes pontos:

- mudar a categorização de algumas covariáveis (idade, uso de suporte ventilatório, entre outros) a fim de obter um maior poder preditivo;

- avaliar o efeito dos fatores associados na probabilidade de óbito dos pacientes;

- utilizar outras técnicas de aprendizado supervisionado.

Referências

BASTIAN, M. *GPT-4 has more than a trillion parameters - Report*. [S.l.], 2023. Disponível em: <https://the-decoder.com/gpt-4-has-a-trillion-parameters/>. Acesso em: 06 jun. 2023.

BERKSON, J. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, n. 39, p. 357–365, 1944.

DABBAGH, R. e. a. Harnessing machine learning in early covid-19 detection and prognosis: A comprehensive systematic review. *Cureus*, v. 15, 2023.

HO, T. K. Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, p. 278–282, 1995.

HUANG C.; WANG, Y. e. a. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, v. 395, n. 10223, p. 497–506, 2020.

MCCULLOCH, W. W. P. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, v. 5, p. 115–133, 1943.

SILVADARCY RISOMARIO; NETO, R. d. S. Inteligência artificial e previsão de óbito por covid-19 no brasil: uma análise comparativa entre os algoritmos logistic regression, decision tree e random forest. *Saúde debate*, v. 46, 2022.

VERHULST, P.-F. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance Mathématique et Physique*, n. 10, p. 113–121, 1838.

VERHULST, P.-F. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, n. 18, 1845.