UNIVERSIDADE FEDERAL DO CEARÁ - CAMPUS SOBRAL

CURSO DE ENGENHARIA DE COMPUTAÇÃO

DISCIPLINA: TÓPICOS EM PROGRAMAÇÃO

PROFESSOR: WENDLEY S. SILVA

INTRODUÇÃO A MACHINE LEARNING COM SCIKIT-LEARN

ISRAEL DA SILVA PEREIRA - 497145

SOBRAL - CE

2024

SUMÁRIO

1 PARTE 1

1.1 Introdução

Machine Learning (Aprendizado de Máquina) é um subcampo da Inteligência Artificial (IA) que se concentra no uso de dados e algoritmos para imitar a maneira como os humanos aprendem, melhorando gradualmente sua precisão. Ele envolve o desenvolvimento de algoritmos e modelos estatísticos que permitem aos computadores melhorar seu desempenho em tarefas através da experiência. Os algoritmos de Machine Learning são usados para fazer uma previsão ou classificação. Com base em alguns dados de entrada, que podem ser rotulados ou não rotulados, seu algoritmo produzirá uma estimativa sobre um padrão nos dados.

O **scikit-learn** é uma biblioteca Python que fornece uma seleção de algoritmos de aprendizado de máquina eficientes e fáceis de usar. Ele é construído sobre as bibliotecas NumPy, SciPy e matplotlib, que são outras bibliotecas Python que suportam computação numérica e científica. O tutorial em estudo é uma introdução ao aprendizado de máquina usando o scikit-learn que aborda tais tópicos:

- **Carregamento de um conjunto de dados de exemplo**: O scikit-learn vem com alguns conjuntos de dados padrão, como o conjunto de dados iris e dígitos para classificação.
- **Aprendizado e previsão**: Depois de definir o problema e carregar os dados, o próximo passo é treinar um modelo em seus dados e fazer previsões.
- **Convenções**: O tutorial também discute algumas das convenções usadas no scikit-learn.

1.2 Conjunto de dados utilizados

O tutorial usa dois conjuntos de dados padrão que vêm com o scikit-learn:

1. **Conjunto de dados Iris**: Este é talvez o banco de dados mais conhecido que pode ser encontrado na literatura de reconhecimento de padrões. O conjunto de dados contém 3 classes de 50 instâncias cada, onde cada classe se refere a um tipo de planta de íris. As classes são linearmente separáveis umas das outras. Cada exemplo no conjunto de dados especifica as medidas das sépalas e pétalas e a espécie da íris.
2. **Conjunto de dados Digits**: Este conjunto de dados é composto por imagens de dígitos manuscritos. Cada imagem é uma matriz 8x8 de valores de escala de cinza, representando um dígito manuscrito (0-9). O conjunto de dados tem um total de 1797 amostras, e cada

amostra é uma imagem 8x8 de um dígito.

Ambos os conjuntos de dados são usados para problemas de classificação, onde o objetivo é prever a classe de entrada com base em certos atributos.

1.3 Algoritmos utilizados

Para analise foi utilizado os 3 algoritmos abaixo:

- **Support Vector Machines (SVM)**: O SVM é um algoritmo de aprendizado de máquina supervisionado que é usado principalmente para tarefas de classificação, mas também pode ser usado para regressão. O objetivo do SVM é encontrar um hiperplano em um espaço N-dimensional (N - o número de recursos) que classifica claramente os pontos de dados. Para separar as duas classes de pontos de dados, existem muitos hiperplanos possíveis que poderiam ser escolhidos. O SVM busca o hiperplano que tem a distância máxima entre os pontos de dados de ambas as classes. Esses pontos de dados são chamados de vetores de suporte.
- **KNeighborsClassifier**: É um algoritmo de aprendizado baseado em instância, onde você precisa de casos de dados rotulados que você deseja classificar. Um caso de dados não rotulado (ou seja, um teste) é classificado pelo voto majoritário dos rótulos K mais próximos entre seus vizinhos. As distâncias são calculadas usando medidas como a distância euclidiana, a distância de Hamming, a distância de Manhattan e a distância de Minkowski.
- **RandomForestClassifier**: É um método de aprendizado de máquina que opera construindo uma infinidade de árvores de decisão no momento do treinamento e produzindo a classe que é o modo das classes (classificação) ou a média das previsões individuais (regressão) das árvores. RandomForest corrige o hábito das árvores de decisão de se ajustarem demais ao seu conjunto de treinamento.

1.4 Metodologia

Como metodologia foram seguidos, em ordem, os seguintes passos:

1. **Divisão dos dados**: Nessa fase, temos a divisão dos dados em 20% para teste e o restante para treinamento.
2. **Treinamento e teste**: Ademais, se tem a criação dos 3 modelos treinados dos algoritmos

citados acima, e por fim a obtenção das classificações dadas por cada modelo para os dados de teste.

3. **Avaliação**: Por fim, temos as métricas de avaliação de desempenho dos algoritmos, para isso foi utilizado os parâmetros:
   - *Matriz de Confusão*: É uma tabela que mostra as frequências de classificação para cada classe do modelo. Para problemas de várias classes, ela terá dimensões de n x n, onde n é o número de classes. As linhas geralmente representam as classes reais e as colunas representam as classes previstas.
   - *Precisão*: É a proporção de verdadeiros positivos (previsões corretas) em relação ao total de positivos previstos. É uma medida de quão preciso seu modelo é em termos de prever a classe positiva.
   - *Recall (Sensibilidade)*: É a proporção de verdadeiros positivos em relação ao total de amostras reais positivas. É uma medida de quão bom seu modelo é em prever a classe positiva quando a classe real é positiva.
   - *F-Score (ou F1-Score)*: É a média harmônica entre precisão e recall. Ele tenta encontrar o equilíbrio entre precisão e recall. O F1-Score é alto se ambos recall e precisão são altos.
   - *Acurácia*: É a proporção de previsões corretas (verdadeiros positivos e verdadeiros negativos) em relação ao total de amostras. É uma medida de quão bom seu modelo é em prever ambas as classes, positivas e negativas.

1.5 Resultados obtidos

Ao analisar os resultados das Figuras 1 a 14, observamos que, para o banco de dados Digits, os algoritmos SVM e KNN exibiram resultados bastante semelhantes em todas as cinco métricas de avaliação. Ambos se destacaram com uma acurácia aproximada de 99.7%. Por outro lado, o algoritmo Random Forest apresentou desempenho ligeiramente inferior em todas as métricas, com uma acurácia, também notável, de 97.9%. No que diz respeito ao banco de dados Iris, o algoritmo KNN foi o grande destaque, atingindo o máximo em todas as métricas (Precisão, Recall, F-Score e Acurácia), com uma acurácia perfeita de 100%. Por sua vez, os algoritmos SVM e Random Forest apresentaram resultados similares, com uma acurácia em torno de 96.6% para ambos. É importante considerar a possibilidade de que o algoritmo KNN possa ter ficado super ajustado (overfitting) aos dados do banco de dados Iris.

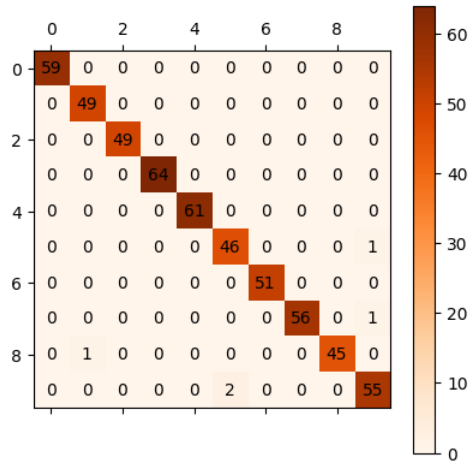

Figura 1 – Matriz de confusão - SVM - Banco de dados Digits

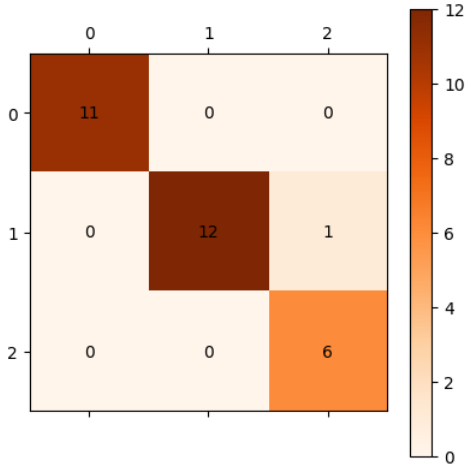

Figura 2 – Matriz de confusão - SVM - Banco de dados Iris

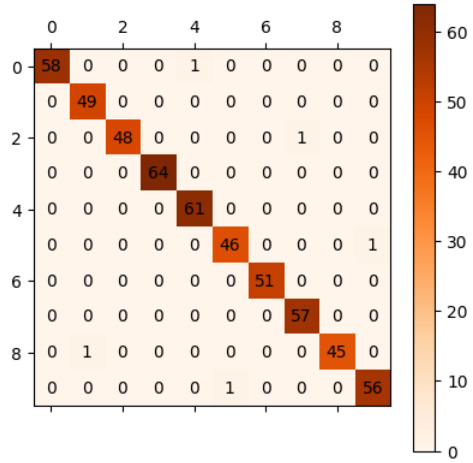

Figura 3 – Matriz de confusão - KNN - Banco de dados Digits

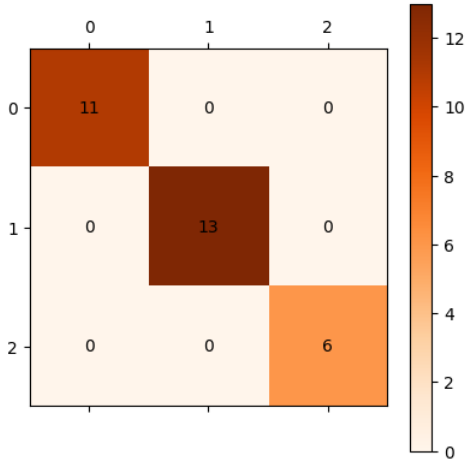

Figura 4 – Matriz de confusão - KNN - Banco de dados Iris

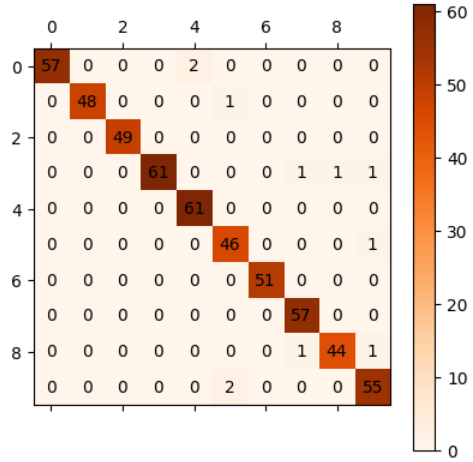

Figura 5 – Matriz de confusão - Random Forest - Banco de dados Digits

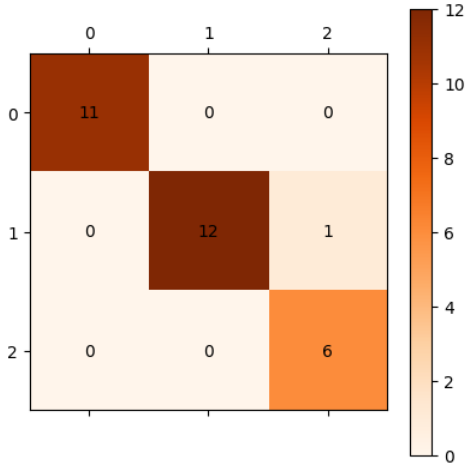

Figura 6 – Matriz de confusão - Random Forest - Banco de dados Iris

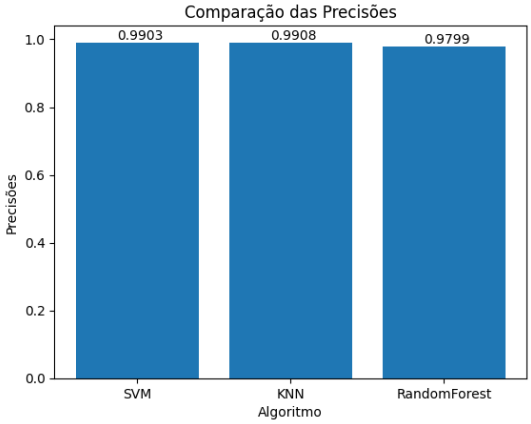

Figura 7 – Precisão - Banco de dados Digits

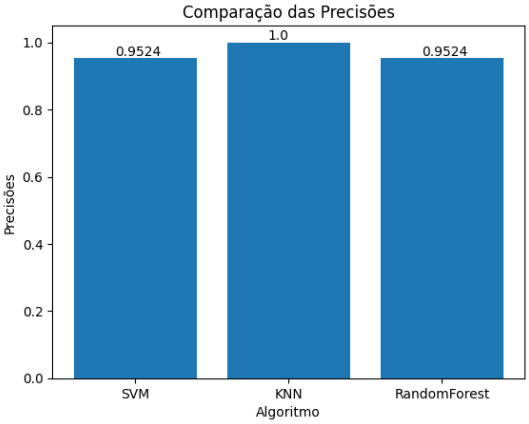

Figura 8 – Precisão - Banco de dados Iris

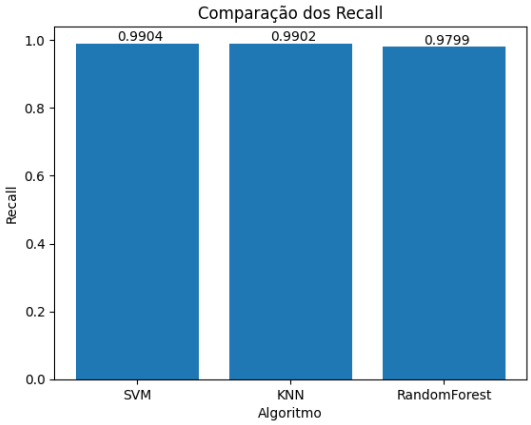

Figura 9 – Recall - Banco de dados Digits

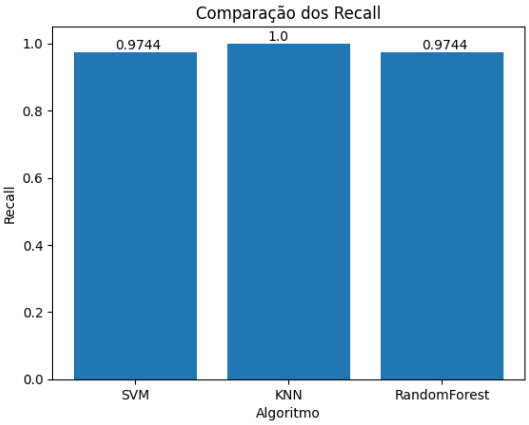

Figura 10 – Recall - Banco de dados Iris

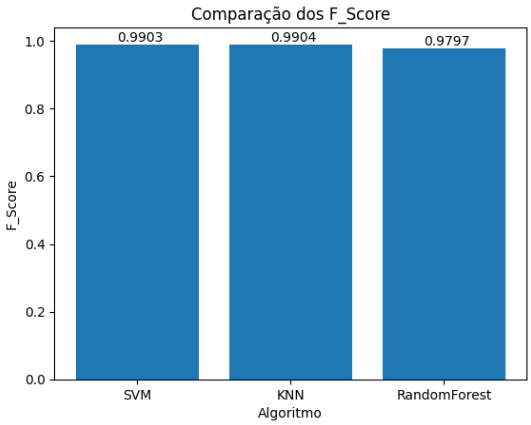

Figura 11 – F_Score - Banco de dados Digits

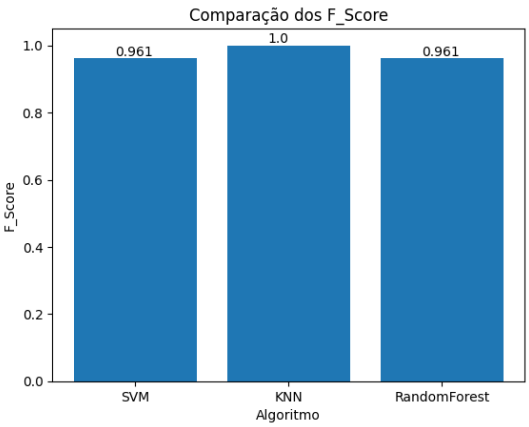

Figura 12 – F_Score - Banco de dados Iris

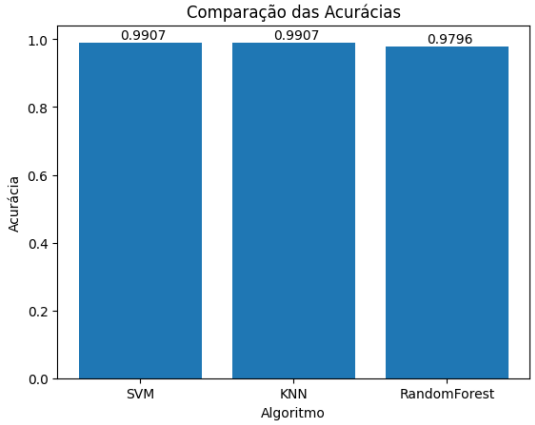

Figura 13 – Acurácia - Banco de dados Digits

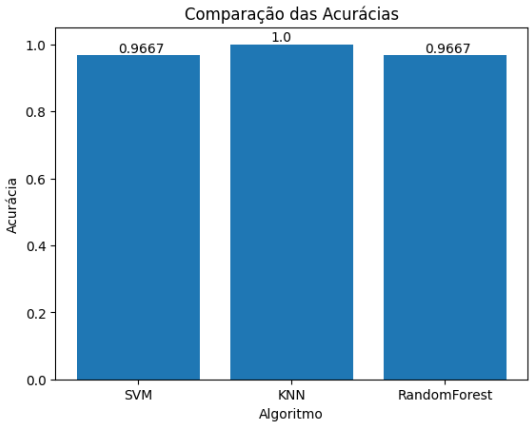

Figura 14 – Acurácia - Banco de dados Iris

1.6 Link para o código-fonte

O código de execução do tutorial e dos resultados obtidos pode ser encontrado **aqui**, em um notebook do Google Colaboratory, onde cada tópico foi divido semelhante ao tutorial.

2 PARTE 2

2.1 Introdução

O câncer de mama é uma das enfermidades mais prevalentes e devastadoras, impactando milhões de indivíduos globalmente. Uma detecção precoce e precisa é vital para um tratamento eficaz e para aumentar as taxas de sobrevivência. Dentro deste cenário, o banco de dados de câncer de mama de Wisconsin se apresenta como um recurso inestimável para pesquisadores e cientistas de dados. Na segunda parte deste estudo, realizaremos uma análise do banco de dados e uma comparação entre quatro algoritmos de aprendizado de máquina, utilizando-o como base.

2.2 Conjunto de dados utilizados

O conjunto de dados de câncer de mama de Wisconsin é um dos conjuntos de dados mais utilizados na área de aprendizado de máquina para tarefas de classificação. Com 569 amostras e 30 características descritivas, ele oferece uma base rica para o desenvolvimento e teste de algoritmos de classificação.

As características do banco de dados são computadas a partir de imagens digitalizadas de biópsia de mama e incluem medidas como a textura, o perímetro e a suavidade das células, fornecendo insights detalhados sobre a natureza do tecido mamário. Esses dados são categorizados em malignos e benignos, permitindo que os modelos de machine learning aprendam a diferenciar entre os dois tipos de condições.

2.3 Algoritmos utilizados

Para explorar e analisar este conjunto de dados foi aplicado não apenas os três algoritmos robustos de aprendizado de máquina apresentados em 1.3, mas também introduziremos um quarto algoritmo poderoso:

- **Logistic Regression**: Um método estatístico que estima a probabilidade de uma amostra pertencer a uma classe, fornecendo uma base probabilística para a classificação.

2.4 Metodologia

A metodologia aplicada segue os passos estipulados em 1.4.

2.5 Resultados obtidos

Para uma visualização preliminar dos dados, realizamos uma redução de dimensionalidade usando PCA (Análise de Componentes Principais), uma técnica estatística utilizada para reduzir a dimensionalidade de um conjunto de dados, reduzindo de 30 para 2 os principais componentes. Isso nos permitiu visualizar a distribuição das classes em um plano bidimensional, conforme ilustrado na Figura 15.

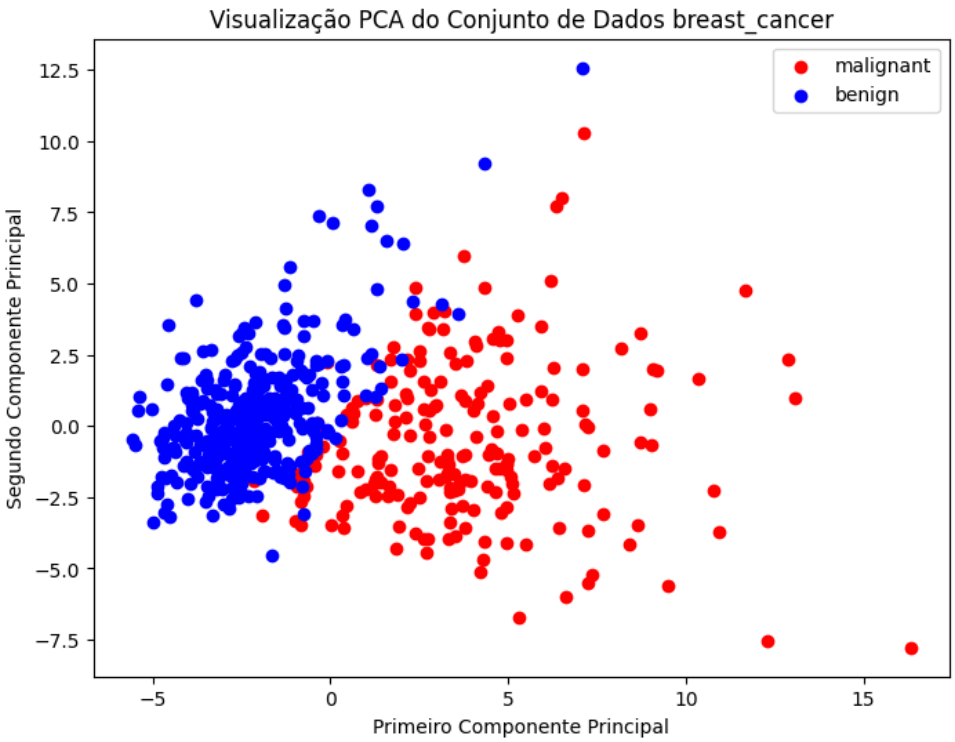

Figura 15 – Distribuição das classes

Temos na Figura 16 as matrizes de confusão dos 4 algoritmos com destaque para o SVM com apenas 2 erros.

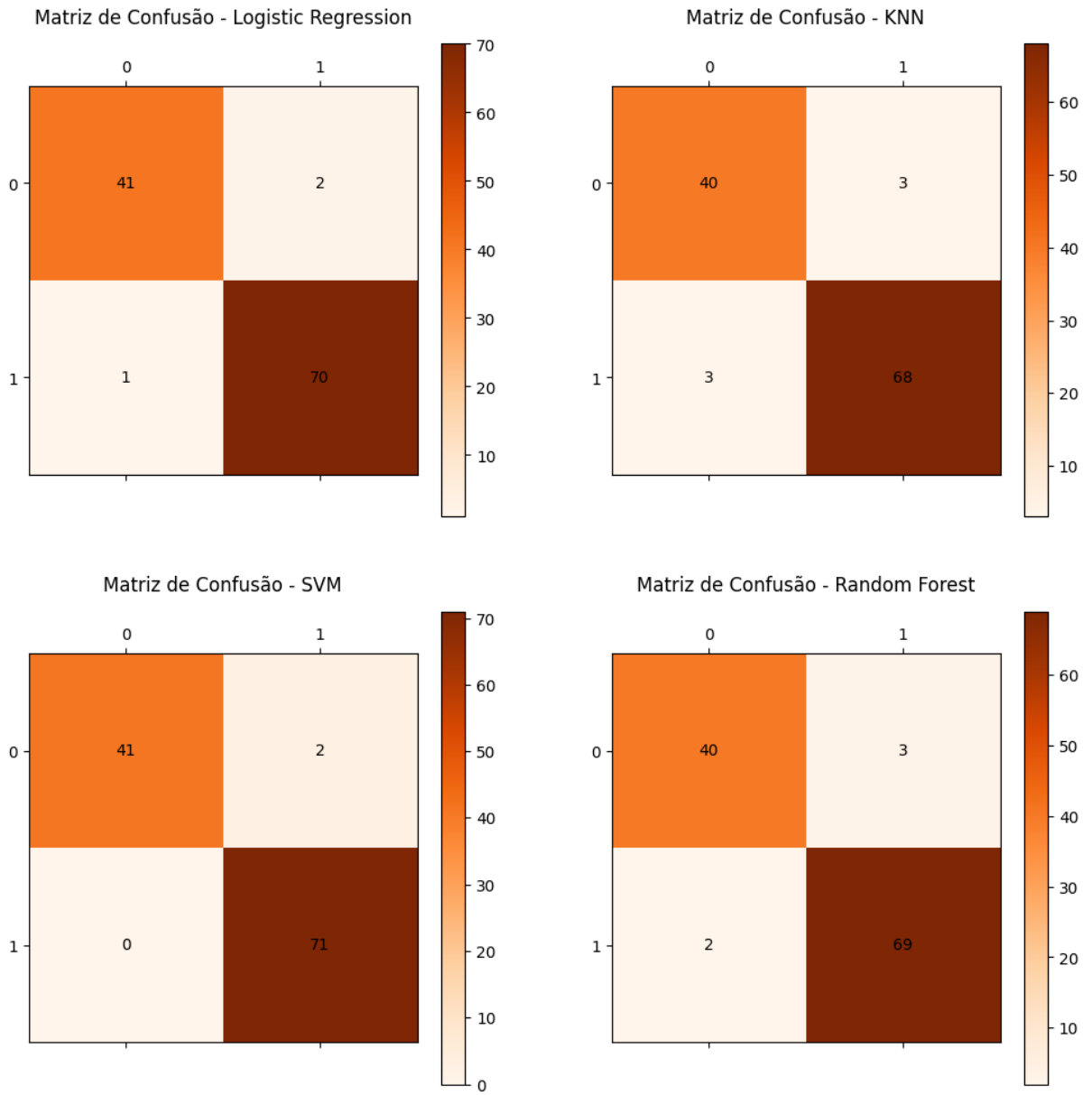

Figura 16 – Matrizes de Confusão

Na Figura 17, apresentamos os valores das quatro métricas para todos os algoritmos. A partir da análise, observamos que o SVM confirma o excelente desempenho já notado anteriormente, destacando-se nas quatro métricas com uma acurácia de 98.2%. Os demais algoritmos apresentaram resultados ligeiramente inferiores, com 97.3% para a Regressão Logística, 94.7% para o KNN e 95.6% para o Random Forest.

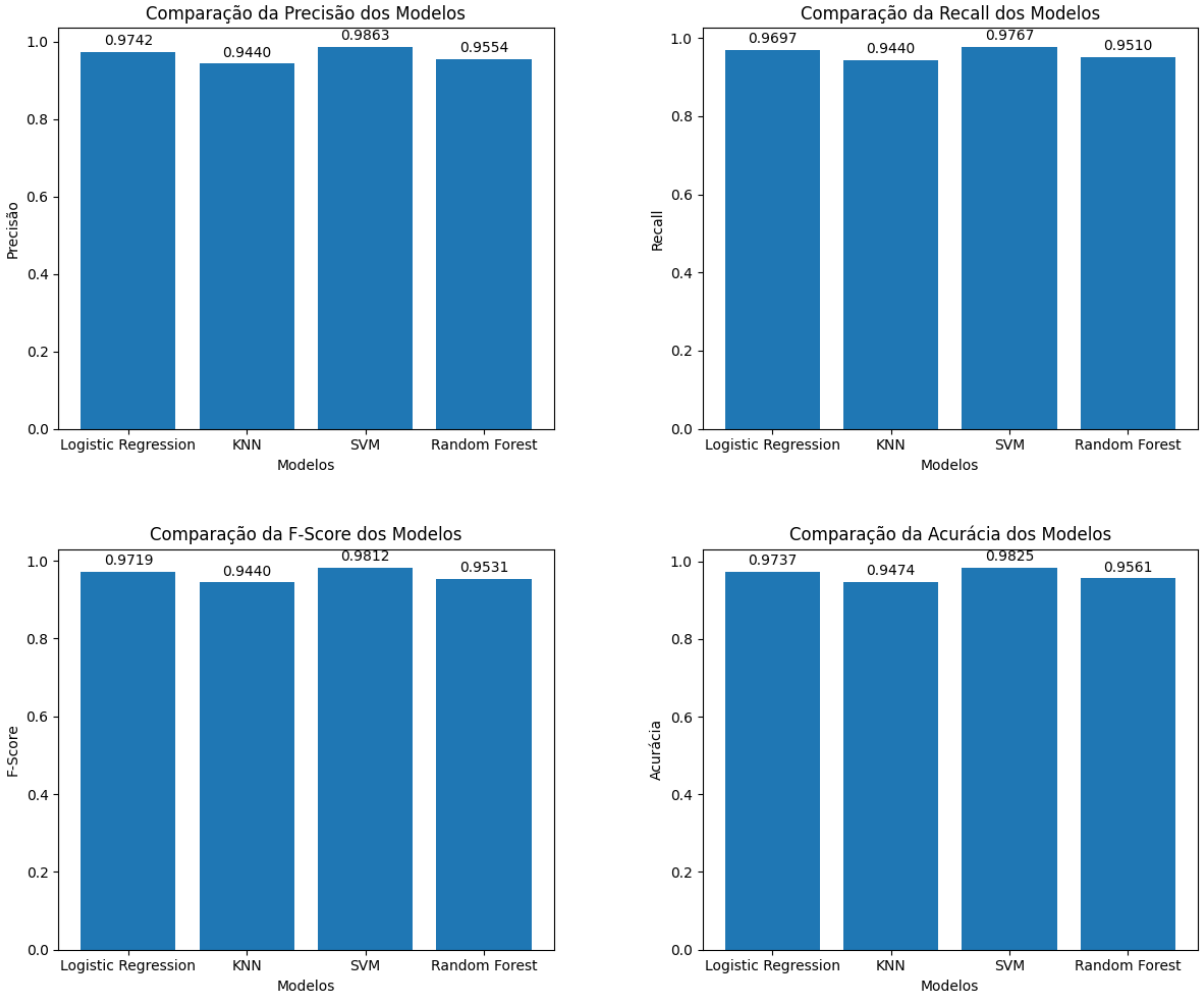

Figura 17 – Matrizes de Confusão

2.6 Link para o código-fonte

O código de execução dos resultados obtidos pode ser encontrado **aqui**, em um notebook do Google Colaboratory, onde a uma breve explicação do que está sendo feito.

REFERÊNCIAS

Scikit-learn developers. **Accuracy Score metrics documentation**. 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html. Acesso em: 16 abr. 2024.

Scikit-learn developers. **Breast Cancer dataset documentation**. 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer. Acesso em: 16 abr. 2024.

Scikit-learn developers. **Confusion Matrix metrics documentation**. 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html. Acesso em: 16 abr. 2024.

Scikit-learn developers. **Decomposition PCA documentation**. 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html. Acesso em: 16 abr. 2024.

Scikit-learn developers. **F_Score metrics documentation**. 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. Acesso em: 16 abr. 2024.

Scikit-learn Developers. **An introduction to machine learning with scikit-learn**. 2024. Disponível em: https://scikit-learn.org/stable/tutorial/basic/tutorial.html. Acesso em: 15 abr. 2024.

Scikit-learn developers. **Precision Score metrics documentation**. 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html. Acesso em: 16 abr. 2024.

Scikit-learn developers. **Recall Score metrics documentation**. 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html. Acesso em: 16 abr. 2024.

Scikit-learn developers. **Supervised learning algorithms documentation**. 2024. Disponível em: https://scikit-learn.org/stable/supervised_learning.html. Acesso em: 15 abr. 2024.