Part-1.1.1

Domain: Healthcare

Part-1.1.2

Data science is revolutionizing the healthcare sector by enabling early disease detection, personalizing treatments, and improving the quality of patient outcomes. For example, AI and machine learning models may identify illnesses such as lung cancer, allowing for earlier intervention and more effective treatment. Likewise, looking at real-world data contributes to the development of precision medicine, which customizes treatments for specific genetic profiles and lifestyles.

Part-1.1.3

In my ideal data science position, I'd like to work on collecting insights from data to create decisions that truly improve healthcare outcomes. I'm passionate about in predictive analytics, where I can assist estimate patient outcomes and follow illness development. This type of job reflects my enthusiasm for problem solving and ambition to make a significant impact in people's lives through data-driven healthcare solutions.

Part-1.1.4

Diversity and inclusion are critical in healthcare to guarantee that data science initiatives accurately reflect all communities. Using different datasets improves the accuracy of research, addresses health inequities, and improves therapies across several populations. Furthermore, including multiple viewpoints in healthcare teams promotes creativity and leads to more culturally and contextually suitable remedies, hence improving patient care for all people.

Part-1.2.1

Data Set:Breast Cancer Wisconsin (Diagnostic)
URL:https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Description:This dataset consists of 569 instances with 31 attributes, including features like mean radius, mean texture, and worst area, related to breast cancer diagnosis. It’s primarily used for binary classification tasks to distinguish between malignant and benign tumors.

Data Set:Pima Indians Diabetes Database
URL:https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Description:This dataset contains 768 records with 8 attributes, including variables like pregnancies, glucose, blood pressure, and BMI. It is used to predict the likelihood of diabetes onset in patients.

Part-1.2.2

I picked the Breast Cancer Wisconsin (Diagnostic) and Pima Indians Diabetes datasets because they present promising potential for predictive analytics in healthcare. Both datasets match with my goal of improving patient outcomes through early diagnosis and risk assessment. The breast cancer dataset is good for creating models to distinguish between malignant and benign tumors, whilst the diabetes dataset is useful for measuring diabetes risk. Using these datasets will allow me to create models that will help improve health management and patient care, both of which are important goals for my data science career.

Part-1.3.1

To assess the efficacy of the machine learning models used on the Breast Cancer Wisconsin and Pima Indians Diabetes datasets, we first carried out multiple preprocessing processes to verify the data was appropriate for modeling.

Data Preparation:

1. Load Datasets
Breast Cancer Dataset: Loaded straight from the machine
Diabetes Dataset: Loaded from the specified URL.

2. Column Changes:
Diabetes Dataset: Added column names for better clarity and uniformity.
Cancer Dataset: Removed an empty column that was not useful for analysis.

3. Basic Preprocessing:
Verified data integrity by checking for "missing values", "duplicates", and "data types". Even though the datasets were pre-cleaned, this step was critical for accuracy.

4. Data Splitting:
Divide each dataset into 80% for training and 20% for testing to assess models on previously unknown data.

Machine Learning Algorithms Applied

1)Support vector machines (SVMs):

Purpose:SVM was chosen due to its efficacy in binary classification tasks and capacity to handle high-dimensional data, such as the cancer dataset's characteristics.

Application: Used independently on each dataset to determine if a tumor is malignant or benign in the cancer dataset and to forecast the start of diabetes in the diabetes dataset.

2) Random Forest:

Purpose:Random Forest was chosen for its resilience and ability to handle complicated feature interactions, making it ideal for datasets like as cancer and diabetes where feature significance and interpretability are crucial.

Application:The model was applied separately to both datasets. In the cancer dataset, Random Forest was used to classify tumors as malignant or benign. In the diabetes dataset, it was employed to predict the onset of diabetes. Its ensemble approach, which aggregates the predictions of multiple decision trees, provided reliable and accurate predictions.

Result for Random Forest

```
Random Forest - Diabetes:
Accuracy: 0.7208
Precision: 0.6071
Recall: 0.6182
F1 Score: 0.6126
ROC AUC: 0.8110

Random Forest - cancer:
Accuracy: 0.9649
Precision: 0.9756
Recall: 0.9302
F1 Score: 0.9524
ROC AUC: 0.9972
```

Result for SVM

```
SVM - Diabetes:
Accuracy: 0.7662
Precision: 0.7209
Recall: 0.5636
F1 Score: 0.6327
ROC AUC: 0.8066

SVM - cancer:
Accuracy: 0.6228
Precision: 0.0000
Recall: 0.0000
F1 Score: 0.0000
ROC AUC: 0.3816
```

Random Forest - Diabetes:
The model demonstrated moderate effectiveness with an accuracy of 72.08%. The precision (60.71%) and recall (61.82%) showed a balanced ability to correctly classify diabetic cases, with an F1 score of 61.26% indicating overall solid performance. The ROC AUC of 81.10% suggests that the model has good discriminative power.

Random Forest - Cancer:
The model excelled, achieving a high accuracy of 96.49%. With precision at 97.56% and recall at 93.02%, the model was highly effective in distinguishing between malignant and benign tumors. The F1 score of 95.24% and near-perfect ROC AUC of 99.72% underscore its reliability and accuracy.

SVM - Diabetes:
The SVM model provided a slightly better accuracy of 76.62% compared to Random Forest but showed a lower recall (56.36%), indicating it might miss more true positives. The F1 score of 63.27% and ROC AUC of 80.66% suggest that SVM was effective, though with trade-offs in precision and recall.

SVM - Cancer:
SVM's performance was significantly poorer in the cancer dataset, with an accuracy of 62.28% and a complete failure in precision, recall, and F1 score (all 0.00%). The ROC AUC of 38.16% indicates that the model was ineffective, possibly due to a poor fit for the dataset's characteristics.

The insights from Random Forest and SVM were complimentary for the diabetes dataset, as both delivered moderate to high performance, yet with some trade-offs in accuracy and recall. However, with the cancer dataset, the models produced inconsistent results. While Random Forest was very

effective, SVM failed totally, possibly due to SVM's sensitivity to certain data features, such as feature scaling or kernel selection.

Part-1.3.2

Effectiveness Analysis using Evaluation Metrics

Accuracy: Used to determine overall accuracy. It was more instructive for the diabetes dataset, where both models performed rather well. However, for the cancer dataset, accuracy alone was deceptive for SVM since it did not account for the model's failures in precision and recall.

Precision and recall are critical in healthcare, as false positives (precision) and false negatives (recall) have serious consequences. The Random Forest model's strong accuracy and recall in the cancer dataset indicate its dependability, whilst SVM's zero scores emphasize its inefficiency.

F1 Score: This metric was very beneficial for balancing accuracy and recall on the diabetes dataset, because both models performed differently.

The ROC AUC provided a more detailed picture of each model's ability to differentiate across classes. Random Forest's high ROC AUC in both datasets demonstrated great performance, however SVM's low ROC AUC in the cancer dataset revealed poor performance.

Conclusion:

Reflecting on the results, it is evident that model selection is quite important depending on the dataset. Random Forest's outstanding performance on both datasets, particularly in cancer detection, demonstrates that it is a dependable model for healthcare applications. On the other hand, SVM's poor performance on the cancer dataset emphasizes the need of testing and understanding why a model may fail in certain conditions.

This analysis went beyond just using machine learning algorithms; it also required critical thinking and a thorough examination of what the results indicate in a real-world healthcare setting. By carefully picking evaluation indicators and pondering on their relevance, I was able to reach deeper conclusions.

In conclusion, the process of selecting, testing, and assessing these models was well-planned and fully fit with the case study's objectives. The insights acquired were not only about statistics, but also about comprehending their consequences in a hospital environment. This technique teaches how to go beyond simply employing AI tools and how to critically assess and expand the findings to make them genuinely valuable.

Declaration:

I hereby declare that I have used AI tool (like ChatGPT) to get insights for this assignment

Appendix:

Documenting the chatgpt promt

Prompt:what are the best evaluvation metrics to compare the two ml algorithms for health care domain

AI response:[The best evaluation metrics for comparing machine learning algorithms depend on the nature of the problem, the data, and the specific goals of the analysis. Given that you are dealing with a healthcare domain, where predicting heart disease and cancer is critical, the following metrics are particularly useful……]

Prompt:Benefits of using svm machine learning Algorithm

AI response:[Support Vector Machines (SVM) are a powerful and versatile machine learning algorithm, particularly well-suited for classification tasks. Here are the key benefits of using SVM……..]

Prompt:Benefits of using random forest Algorithm:

AI response:[Random Forest is a popular and powerful machine learning algorithm that offers several benefits, especially in classification and regression tasks. Here are some of the key advantages……..]

Reference:

https://www.jnj.com/innovation/how-data-science-ushers-in-new-era-of-modern-medicine
https://www.connectplus.health/post/how-does-data-diversity-improve-healthcare-outcomes
https://pubmed.ncbi.nlm.nih.gov/32336480/
https://www.connectplus.health/
https://www.jnj.com/
https://scikit-learn.org/stable/modules/model_evaluation.html