Electrical and Computer Engineering Department

Machine Learning and Data Science || ENCS5341

Assignment #3

Early-stage diabetes risk prediction

Prepared By: Lana Batnij __ 1200308

Instructor: Dr.Yazan Abu Farha

Section: 2

Table of Contents:

1) Introduction:

The fundamental goal of this study is to solve a binary classification problem in early-stage diabetes risk prediction. Various machine learning models have been checked, each using a different method and technique. The models used were k-nearest Neighbors (KNN) with different values of k, Support Vector Machine (SVM) with a radial basis function kernel, and Random Forest. Each model has unique strengths and behaviors, and its performance on the given diabetes dataset was evaluated.

The models were evaluated using basic classification measures such as precision, recall, F1-score, and accuracy. These evaluations gave an accurate representation of the model's efficiency in identifying both positive and negative outcomes, and knowledge about their overall performance.

2) Dataset:

The dataset that was utilized in this experiment was relocated from the Sylhet Diabetes Hospital in Bangladesh, and it featured sign and symptom information about patients at risk of early-stage diabetes. The dataset, which was gathered using direct questions and authorized by medical professionals, turned out to be a great resource for modeling predictions in the field of diabetes risk identification.

The dataset consisted of 520 examples, each with 17 features. These attributes captured multiple symptoms and signs related to diabetes risk, resulting in a holistic perspective of patient profiles. Investigative data analysis, which included both quantitative indicators and visuals, provided insights into the dataset's properties. Descriptive statistics, such as mean, standard deviation, and quartiles, were calculated for each feature to provide an overview of the data's central tendency and distribution.

A histogram was utilized to display the age distribution among people in the dataset, providing insights into how the ages were distributed. This image helps to comprehend the age group breakdown of the investigation. To examine class imbalance, a bar chart was used to show the distribution of categories (positive/negative). This graphic, known as the class distribution chart, helps with identifying the number of instances of each class within the dataset, which is essential in evaluating any possible biases in modeling.

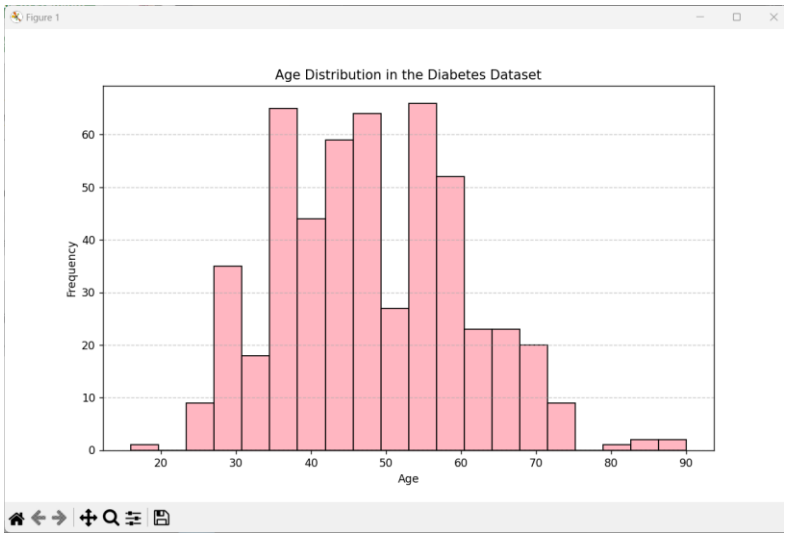

Figure 1: Dataset histogram

In addition, a correlation heatmap was established to show the correlations between various features. The heatmap presented a color-coded picture of the correlation coefficients, which helped highlight probable linkages and dependencies between the dataset's variables. This visualization was very useful for feature selection as well as for determining how many elements may have been linked in the setting of risk for diabetes prediction.

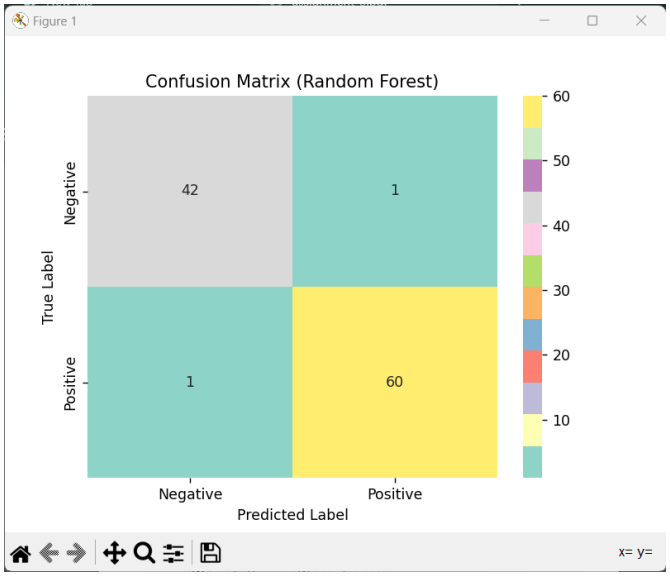

Figure 2: Confusion matrix of Random Forest

3) Experiments and Results:

Many types of systematic tests, involving baseline model launch, evaluation of the model, and hyperparameter modifications, were carried out to thoroughly evaluate the performance of models of prediction for early-stage diabetes risk classification.

3.1: Baseline Models:

The implementation of two baseline models established the foundation for additional investigations. The k-nearest neighbors (KNN) models with k=1 and k=3, which use the Euclidean distance metric, served as starting references. The classification reports revealed some remarkable results. The k=1 model obtained 97% accuracy, with precision, recall, and F1-score metrics approaching 95%. The k=3 model maintained a high accuracy of 95%, demonstrating its effectiveness in biasing between positive and negative examples.

3.2: Support Vector Machine (SVM):

An effective Support Vector Machine (SVM) model was constructed, complete with an attribute scaling process. The SVM model, trained on the prepared dataset, obtained 93% accuracy. It showed great precision (95%), recall (94%), and F1-score (95%), demonstrating its benefit in diabetes risk prediction.

3.3: Random Forest Model:

The Random Forest model, which uses a group of decision trees, performed exceptionally well, achieving 98% accuracy. Precision, recall, and F1-score metrics all reached 98%, demonstrating the model's ability to capture complicated patterns in the dataset. The model's impressive selective power was also represented in the area under the ROC curve.

3.4: Results:

All models maintained the capacity to identify early-stage diabetes risk, while the Random Forest model outperformed the others. It achieved precision and recall scores, with an overall accuracy of 98%. These findings provide important insights into machine learning models' prediction powers in the framework of diabetes risk classification.

4) Analysis:

4.1: K-Nearest Neighbors Model:

Classification Report (k=1):

	precision	recall	f1-score	support
0	0.93	1.00	0.97	42
1	1.00	0.95	0.98	62
accuracy			0.97	104
macro avg	0.97	0.98	0.97	104
weighted avg	0.97	0.97	0.97	104

Figure 3: KNN at K=1

k=1: Obtained 97% accuracy with great precision, recall, and F1-score in both classes. The model excelled at capturing individual cases. k=3: Showed somewhat lower accuracy (96%), with outstanding recall for class 0 while maintaining precision and recall for class 1.

Classification Report (k=3):

	precision	recall	f1-score	support
0	0.89	1.00	0.94	42
1	1.00	0.92	0.96	62
accuracy			0.95	104
macro avg	0.95	0.96	0.95	104
weighted avg	0.96	0.95	0.95	104

Figure 4: KNN at K=3

The K-Nearest Neighbors (KNN) model demonstrated several advantages, including high accuracy with k=1, good recording of individual cases, and appropriate precision-recall metrics. It demonstrated efficiency in capturing local patterns. However, drawbacks included somewhat decreased accuracy at k=3. showing potential for local variations, and a limited capacity to identify global patterns

4.2: Support Vector Machine Model:

```
Accuracy (SVM): 0.93
Precision (SVM): 0.95
Recall (SVM): 0.94
F1-Score (SVM): 0.94
```

Figure 5: SVM performance

The Support Vector Machine (SVM) model achieved high accuracy with incredible precision, recall, and F1-score metrics, demonstrating its ability to capture complex patterns and handle nonlinear relations. Generalize to the test set performed well across a variety of measures. However, considering its high level of accuracy, there was some potential for improvement. SVM was selected as an effective choice due to its general high performance and effectiveness in detecting complicated connections, making it appropriate for real-world implementation. Imperfections analysis discovered challenges in precision, recall, and F1-score, reflecting the difficulty in identifying specific patterns or instances, which could be due to fundamental complexities in the dataset that require more advanced models or more features.

4.3: Random Forest Model:

The Random Forest model shows strengths such as high accuracy and balanced precision-recall metrics, allowing it to handle complicated patterns and relationships between features successfully. Generalization to the test set performed well across several measures. However, although it has excellent accuracy, there is some potential for improvement. The Random Forest model was selected as the best alternative due to its excellent results across all evaluation measures, making it an acceptable choice for real-world implementation. Despite its overall good performance, minor differences in precision, recall, and F1 score suggested difficulties in recording specific patterns or instances.

```
Accuracy (Random Forest): 0.98
Precision: 0.98
Recall: 0.98
F1-Score: 0.98
AUC-ROC: 0.98
```

Figure 6: Random Forest Model

4.4: Misclassifications at each Model:

Possible improvements in mistaken classification for the K-Nearest Neighbors (KNN) model can be explored by looking into cases where broader trends fail to be detected or local context is strongly depended on. This could include considering the addition of new features or altering the value of k. Similarly, errors in classification in the Support Vector Machine (SVM) model could happen due to unclear patterns or a lack of selective characteristics. Analyzing specific misclassified examples can provide alternatives, such as collecting more relevant information or investigating complex model designs. The Random Forest approach provides potential improvements for dealing with misclassifications caused by confusing patterns or insufficient filtering features. Analyzing individual misclassified instances might provide insights into potential improvements, such as gathering extra relevant features or investigating more advanced model structures

5) Conclusions and Discussion:

In summary, experiments using K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest models on the diabetes dataset showed impressive results, each with its own set of strengths and limitations. KNN excelled in recognizing local patterns, whereas SVM showed great accuracy and strong generalization. The Random Forest model outperformed the other models because of its balanced precision-recall metrics. Despite their results, all models had small weaknesses that might be improved through parameter tuning and the use of advanced methodologies. The most appropriate model is chosen based on specific job needs, and continuous tracking, improving, and evaluation of dataset features is essential for the best real-world applicability.

6) Database Reference:

[1] B. Doctors at Sylhet Diabetes Hospital in Sylhet, "UCI Machine Learning Repository," Sylhet Diabetes Hospital in Sylhet, Bangladesh, 7 11 2020. [Online]. Available: https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset. [Accessed 10 1 2024].