جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Engineering & Technology
Electrical & Computer Engineering Department

ENCS5341

Assignment #3

Prepared by: Nidal Zabade-1200153 & Diaeddin Tahboub-1200136

Instructor: Dr. Yazan Abu Farha

Section: 1

Date: 26/1/2024

Introduction

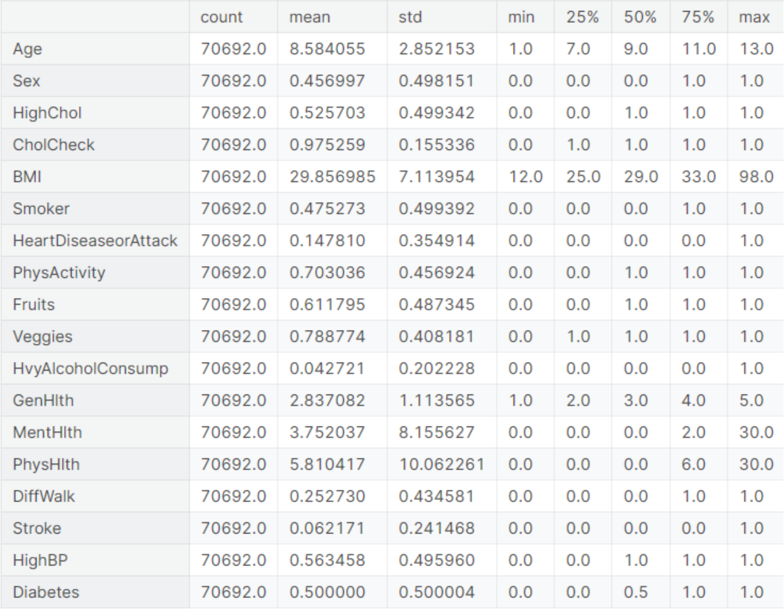
The goal of this assignment is to explore and evaluate various machine-learning models for a predictive task on a diabetes classification problem. The task involves predicting if the input features. The diabetes's dataset provided for this assignment contains instances of both input features and corresponding target variable values. In this assignment, we used three kind of models the first one is K nearest neighbor's classifier with two different values of K, the second one is XGBoost classifier and apply a Grid Search to find the best hyper-parameters for the XGBoost model, and the third one was using Artificial Neural Network model (ANN) we make three different ANN models with different number of layers. We applied z-score to remove the outliers after checking and handling the missing data (we didn't have any missing values). In addition, for the evaluation metrics, because the dataset was a balanced dataset we calculate the accuracy, precision, recall and the F1-score and chose the model with the highest values.

Dataset

The Diabetes Prediction dataset from Kaggle, This dataset has 70000 record and designed for the task of predicting the presence or absence of diabetes in individuals based on several key body and health parameters. The dataset encompasses a diverse set of features, including physiological and lifestyle-related metrics like (Age, Sex, HighChol, CholCheck …) in total of 17 features and the 1 target value. The features described as follow

Feature	Description
Age	13-level age category arranged in 5 years steps: 1 = 18-24; 2 = 25-29; ...; 9 = 60-64; ...; 13 = 80 or older.
Sex	Patient's gender 1 = male; 0 = female.
HighChol	0 = no high cholesterol; 1 = high cholesterol.
CholCheck	0 = no cholesterol check in 5 years; 1 = yes cholesterol check in 5 years.
BMI	Body Mass Index.
Smoker	Have you smoked at least 100 cigarettes in your entire life? (5 packs = 100 cigarettes) 0 = no; 1 = yes.
HeartDiseaseorAttack	Coronary heart disease (CHD) or myocardial infarction (MI) 0 = no; 1 = yes.
PhysActivity	Physical activity in past 30 days - not including job 0 = no; 1 = yes.
Fruits	Consume Fruit 1 or more times per day 0 = no; 1 = yes.
Veggies	Consume Vegetables 1 or more times per day 0 = no; 1 = yes.
HvyAlcoholConsump	(Adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no; 1 = yes.
GenHlth	Would you say that in general your health: scale 1-5; 1=excellent; 2=very good; 3=good; 4=fair; 5= poor.
GenHlth	Days of poor mental health scale 1-30 days.
PhysHlth	Physical illness or injury days in past 30 days scale 1-30.
DiffWalk	Do you have serious difficulty walking or climbing stairs? 0 = no; 1 = yes.
Stroke	You ever had a stroke 0 = no; 1 = yes.
HighBP	0 = no high; BP 1 = high BP.

Table 1: Features description



	count	mean	std	min	25%	50%	75%	max
Age	70692.0	8.584055	2.852153	1.0	7.0	9.0	11.0	13.0
Sex	70692.0	0.456997	0.498151	0.0	0.0	0.0	1.0	1.0
HighChol	70692.0	0.525703	0.499342	0.0	0.0	1.0	1.0	1.0
CholCheck	70692.0	0.975259	0.155336	0.0	1.0	1.0	1.0	1.0
BMI	70692.0	29.856985	7.113954	12.0	25.0	29.0	33.0	98.0
Smoker	70692.0	0.475273	0.499392	0.0	0.0	0.0	1.0	1.0
HeartDiseaseorAttack	70692.0	0.147810	0.354914	0.0	0.0	0.0	0.0	1.0
PhysActivity	70692.0	0.703036	0.456924	0.0	0.0	1.0	1.0	1.0
Fruits	70692.0	0.611795	0.487345	0.0	0.0	1.0	1.0	1.0
Veggies	70692.0	0.788774	0.408181	0.0	1.0	1.0	1.0	1.0
HvyAlcoholConsump	70692.0	0.042721	0.202228	0.0	0.0	0.0	0.0	1.0
GenHlth	70692.0	2.837082	1.113565	1.0	2.0	3.0	4.0	5.0
MentHlth	70692.0	3.752037	8.155627	0.0	0.0	0.0	2.0	30.0
PhysHlth	70692.0	5.810417	10.062261	0.0	0.0	0.0	6.0	30.0
DiffWalk	70692.0	0.252730	0.434581	0.0	0.0	0.0	1.0	1.0
Stroke	70692.0	0.062171	0.241468	0.0	0.0	0.0	0.0	1.0
HighBP	70692.0	0.563458	0.495960	0.0	0.0	1.0	1.0	1.0
Diabetes	70692.0	0.500000	0.500004	0.0	0.0	0.5	1.0	1.0

Table 2: Descriptive statistics

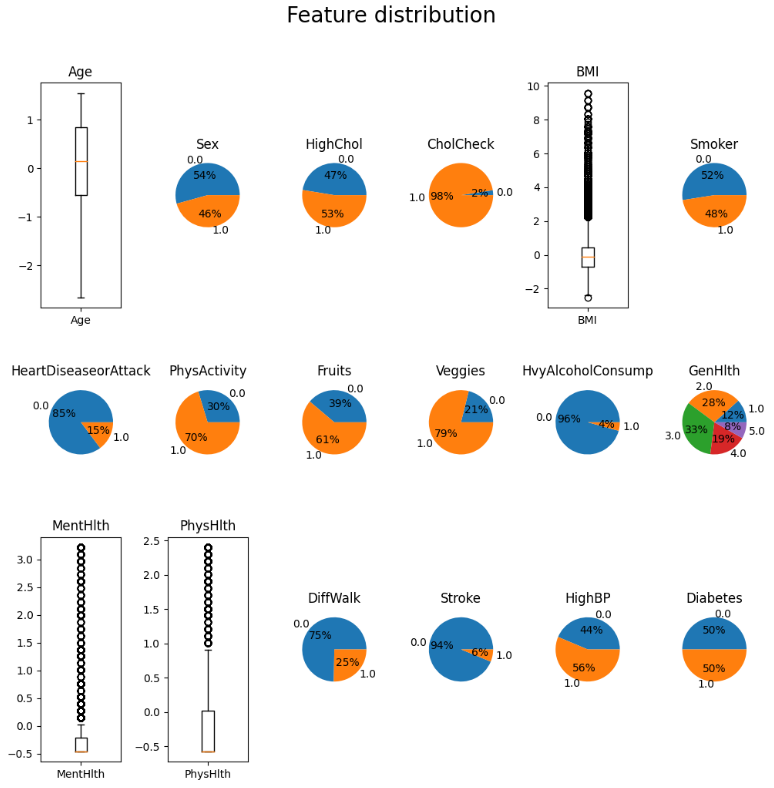


Figure 1: Feature Distribution

Experiments and Results

KNN model

The first one is K nearest neighbor's classifier with two different values of K the first value was k=1 and the k=3 with using Euclidean distance for both and the results were as follow:

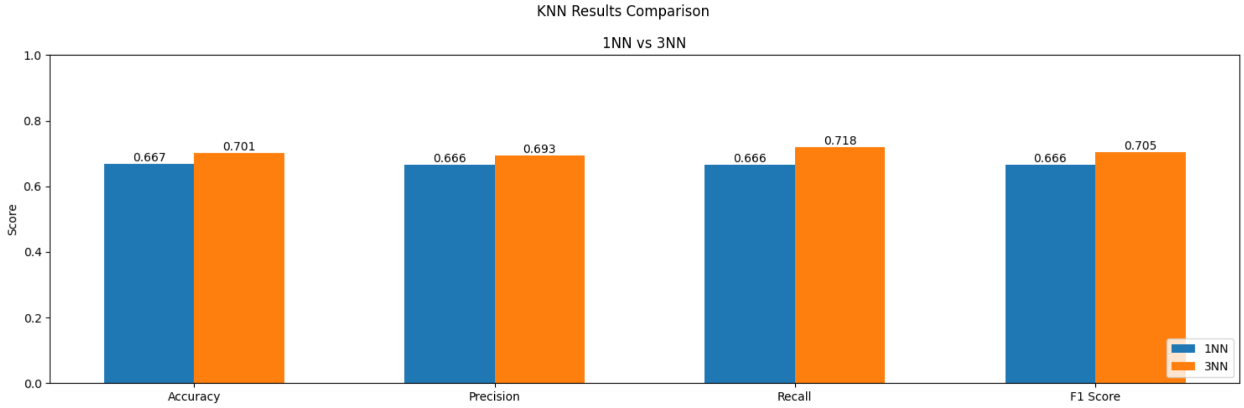


Figure 2: KNN Results Comparison

As shown in the above figure the knn model with k=3 has performed and give higher results on the validation set than the knn model where k=1.

XGBoost Model

For the XGBoost model, we applied three hyper-parameters with at least three values for each hyper-parameters and the hyper-parameters ware (n_estimators, max_depth, and learning_rate), we applied a grid search to find the best combination of the parameters based on the recall. The best model for XGBoost was (n_estimators=50, max_depth=3, and learning_rate=0.001), with the following results.

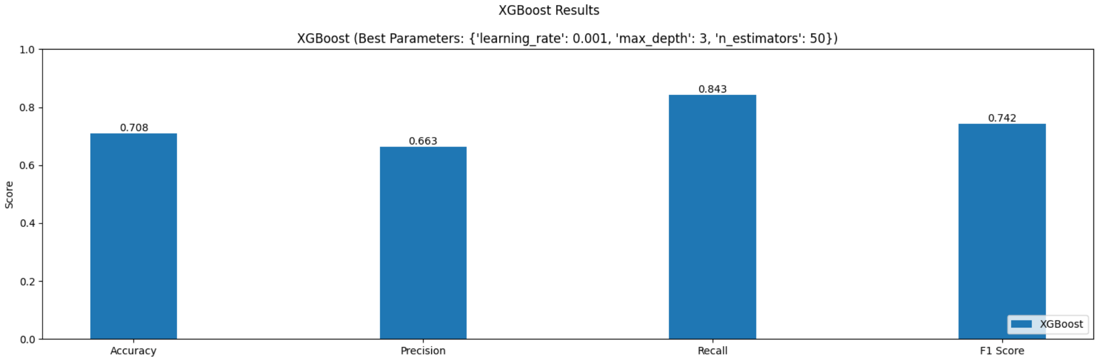


Figure 3: Best XGBoost performance

From the above figure, the XGBoost has better results than the KNN for all metrics except precision it was low.

ANN Model

For ANN Model we created 3 ANN models each has more layers than the previous one so for the first model only has 3 layers (1 input, 1output), the second one has 4 layers (1 input, 1output) and the last one has 5 layers (1 input, 1output). Each model has factor of hidden units which 30 and training epochs which 200 and the results were as follow:

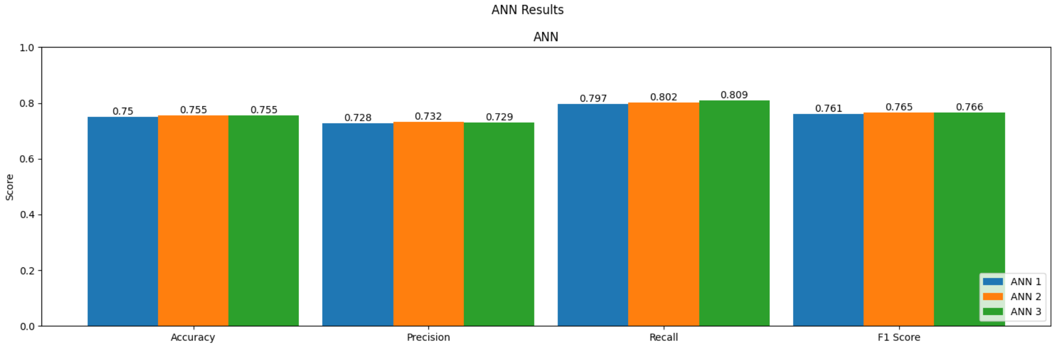


Figure 4: ANN Comparison

As shown in the figure the third model (ANN3) performed the best on the validation set and learned very well, moreover it overall performance was better the KNN and XGBoost models so we chose it as our classification model.

Analysis

After we chose the ANN3, model we wanted to see of the model has learned as much as possible and the loss reduced with increasing the epochs.

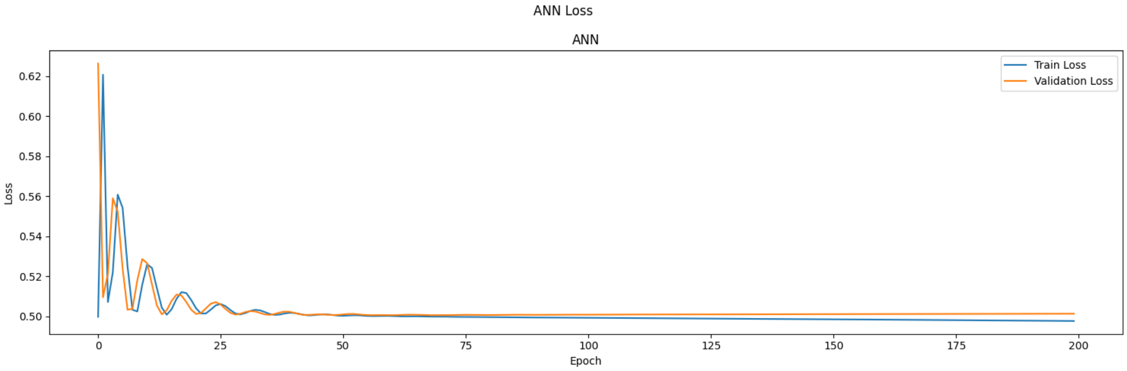


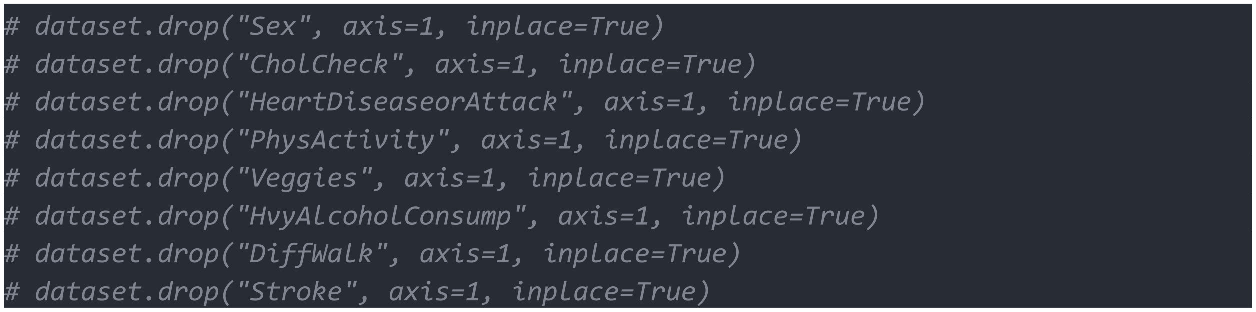
Figure 5: Training loss VS Validation loss

Form the above figure we saw that our model loss decreased with the increasing of the epochs so we tried to see how it would perform in the testing set.

Accuracy	0.7595133781433105
Precision	0.7380585670471191
Recall	0.8086662888526917
F1 Score	0.771750807762146

Table 3: ANN performance on the testing set

The results was good but not good enough so we tried to see any common patterns on the incorrect prediction (1700 incorrect example found using the code), so stated to drop feature or more and make the model relearn the data to see any improvement.



```
# dataset.drop("Sex", axis=1, inplace=True)
# dataset.drop("CholCheck", axis=1, inplace=True)
# dataset.drop("HeartDiseaseorAttack", axis=1, inplace=True)
# dataset.drop("PhysActivity", axis=1, inplace=True)
# dataset.drop("Veggies", axis=1, inplace=True)
# dataset.drop("HvyAlcoholConsump", axis=1, inplace=True)
# dataset.drop("DiffWalk", axis=1, inplace=True)
# dataset.drop("Stroke", axis=1, inplace=True)
```

Figure 6: Some attempts of dropping

The results did not change a lot for example after we dropped the Stroke and the DiffWalk features the results was as follow:

Accuracy	0.7521573305130005
Precision	0.7422327399253845
Recall	0.7847995758056641
F1 Score	0.7629228830337524

Table 4: ANN performance on the testing set after the dropping

Therefore, we leave the dataset as it is. Moreover, we did not increase the number of the epochs as seen in the Training loss VS Validation loss figure the loss remain the same after 50 epochs.

Conclusion and Discussion

In conclusion, our exploration and evaluation of machine-learning models for the diabetes classification problem revealed valuable insights. The K nearest neighbors (KNN) classifier with varying values of K, the XGBoost classifier optimized through Grid Search for hyper-parameter tuning, and the Artificial Neural Network (ANN) models with different layers were all employed in this study. The models were assessed using accuracy, precision, recall, and F1-score metrics on the balanced dataset. The results indicated that the XGBoost classifier with optimized hyper-parameters outperformed the other models, demonstrating the importance of parameter tuning for enhancing model performance. For the limitation, the features were not enough for the models to perform very well and there were not any direct correlation between the features and the target but overall with a performance like Accuracy = 0.759, Precision= 0.738, Recall= 0.808 and F1 Score= 0.772, I think our model and our experimentations goes well.