PG3102-1 Examination

1st 999995
dept. Oslo, Norway
Artificial Intelligence
Oslo, Norway

2nd 999301
dept. Oslo, Norway
Artificial Intelligence
Oslo, Norway

line 1: 3rd 112015
dept. Oslo, Norway
Artificial Intelligence
Oslo, Norway

***Abstract* – Diabetes stands as a significant public health challenge globally, the number of people with diabetes is increasing due to factors like increasing obesity, lifestyle changes, and aging populations. This concerning trend require innovative approaches for better understanding and managing the disease. Our study aims to address this need by applying advanced machine learning techniques to analyze the Diabetes Health Indicators Dataset. Our approach involves using both unsupervised and semi-supervised learning to search into the complexities of diabetes, uncovering hidden patterns and predictive factors using different health indicators. The objective of this research is to enhance the understanding of diabetes, offering new insights into this area. These insights are intended to help healthcare professionals and policymakers in developing more targeted and effective strategies for diabetes prevention and management, thereby contributing to improved health outcomes in the face of this global health challenge.**

I. Introduction

This paper explores the application of machine learning techniques to the Diabetes Health Indicators Dataset, to uncover insights into diabetes prevalence and associated health factors. We implement both unsupervised and semi-supervised learning approaches to analyze this dataset. The unsupervised learning component utilizes K-Means Clustering to identify inherent groupings within the data, revealing hidden patterns among health indicators. Principal Component Analysis (PCA) is then applied to reduce dimensionality, enhancing the dataset's interpretability and computational efficiency. For the semi-supervised learning aspect, Logistic Regression is implemented to model the probability of diabetes occurrence based on various health indicators. Additionally, tree-based methods, including Decision Trees, Random Forests, and Gradient Boosted Trees, are used to further classify patients and uncover the most significant predictors of diabetes. This study aims to provide a comprehensive understanding of diabetes indicators, offering valuable insights for healthcare professionals and policymakers in designing targeted interventions and preventive measures.

II. Dataset

The Diabetes Health Indicators Dataset, by Alex Teboul, is a comprehensive collection of data aimed at understanding the factors influencing diabetes among individuals. It comprises a wide range of health indicators, including demographic details, behavioral information, and various health-related metrics. Key features in the dataset encompass aspects such as age, gender, race, income, physical activity levels, alcohol consumption, smoking status, body weight, and blood pressure, among others. The dataset consists of 253,680 entries, each with 22 columns.

This dataset is particularly valuable for researchers and healthcare professionals seeking to analyze the correlations between lifestyle choices, socio-economic factors, and the risk of diabetes. It offers a rich source of information for conducting detailed statistical analyses, developing predictive models for diabetes, and formulating targeted healthcare strategies. The dataset's structure supports various forms of data analysis, including supervised and semi-supervised machine learning tasks, making it a versatile tool for academic research and practical healthcare applications.

In the Kaggle report it was three different datasets, we chose to use the dataset with two classes; 0 for non-diabetes and 1 for diabetes. The data is collected through a telephone survey that is conducted annually by the Centers for Disease Control and Prevention (CDC). This has been conducted every year since 1984, with an average of responses from over 400.000 Americans.

Here is some information about the dataset. All columns has dtype float64 (A. Teboul, 2015).

Name	**Explanation**	**Value**
Diabetes_binary	Do you have diabetes?	0 = no, 1 = yes
HighBP	Highblood pressure	0 = no, 1 = yes
HighChol	High cholesterol	0 = no, 1 = yes
CholCheck	Cholesterol check in the last 5 years	0 = no, 1 = yes
BMI	Body mass index	0-100
Smoker	Have smoked at least 100 cigarettes during your lifetime	0 = no, 1 = yes
Stroke	Ever had a stroke	0 = no, 1 = yes
HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI)	0 = no, 1 = yes
PhysActivity	Physical activity in the last 30 days	0 = no, 1 = yes
Fruits	Eat one ore more fruits per day	0 = no, 1 = yes
Veggies	Eat one or more vegtables per day	0 = no, 1 = yes
HvyAlcoholConsump	14 or more drinks per week (men), 7 or more drinks per week (women)	0 = no, 1 = yes

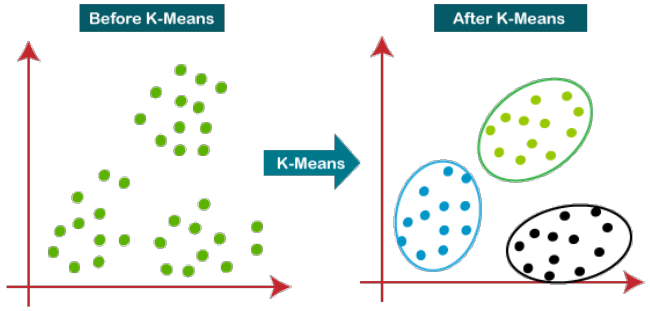
AnyHealthcare	Any form for healthcare coverage/insurance	0 = no, 1 = yes
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?	0 = no, 1 = yes
GenHlth	How Is your general health?	From 1-5 (1 being the best, and 5 being the worst)
MentHlth	Days of bad mental health in the last 30 days	1-30
PhysHlth	Days of bad physical health or injury the last 30 days.	1-30
DiffWalk	Do you have trouble walking or climbing stairs?	0 = no, 1 = yes
Sex	Gender	0 = female. 1 = male
Age	13 different groups of age	1 = 18-24, 9 = 60-64, 13 = 80 or older
Education	Level of education	From 1-6, 1 = Never attended school or only kindergarten 2 = elementary etc.
Income	Income/money earned	From 1-8, 1 = less than 10,000, 5=*less than* 35,000, 8 = $75,000 or more

III. Models

In the next paragraphs there is a short description of the models we are going to use in our project. Including the description of the models, there is information on why the model is useful for our project.

IV. K-Means clustering

This is an unsupervised algorithm, which divides your data into a fixed number (K) of clusters. It clusters all the different data points into clusters which have similarities. By doing this the algorithm might discover underlying patterns. The amount of clusters (K) you define, is referring to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. So after the K-Means clustering algorithm is applied, your data is divided into K clusters (Education Ecosystem, 2018).

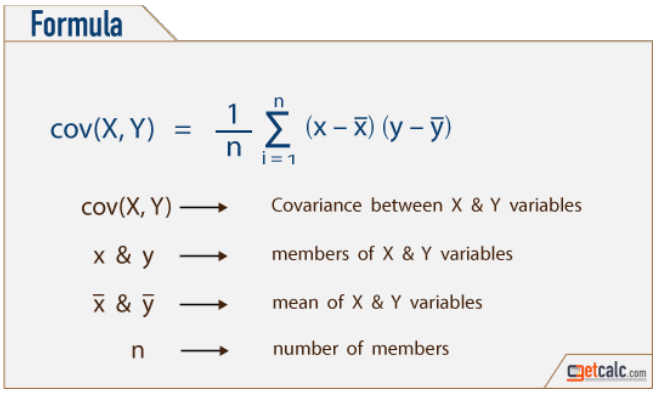


Figur 1 – Ilustation of K-Means Clustering (Sharma, 2023)

In the context of our project, K-Means can be used to identify inherent groupings among patients based on health indicators. This could reveal patterns or subgroups within the data that are not immediately obvious. Which might give us valuable information.

V. PCA

Principal Component Analysis (PCA) is in the unsupervised algorithm category. PCA is a dimensionality reduction algorithm used to reduce the number of variables in your data by extracting the most important information. It transforms the data into a new set of variables, the principal components. In the end the dataset variables is reduced, while preserving as much information as possible (Jaadi & Whitfield, 2023).



Figur 2 - Formula for PCA (Dubey, 2018)

In our project, PCA can be utilized to reduce the number of health indicators while retaining the most significant information. This can simplify the dataset and potentially improve the performance of other algorithms by eliminating noise and redundancy.

VI. Logistic regression

Logistic regression is a model that is in the supervised learning category. This type of model is often used for prediction or classication. The outcome of this model is either

1 or 0 (True or False), in our example diabetes or non-diabetes (Kanade, 2022).

$$f(x) = \frac{1}{1 + e^{-x}}$$

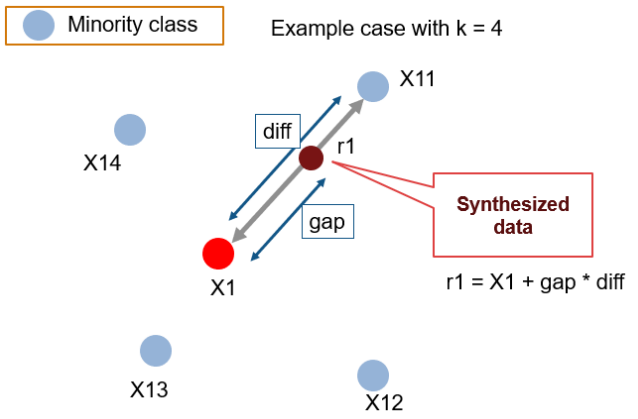
Equation of Logistic Regression

Figur 3 - Formula for logistic regression (Kanade, 2022)

In the context of our project, logistic regression can be used to predict the likelihood of a patient having diabetes based on their health indicators. This model is beneficial for understanding the relationship between different health indicators and the binary outcome (diabetes or non-diabetes).

VII. SMOTE

Smote or Synthetic Minority Over-sampling Technique is an over-sampling technique used to address class imbalance in a dataset, which is a common issue in machine learning, especially in medical diagnostics like diabetes prediction. It works by creating synthetic samples from the minority class instead of creating copies. This is done by selecting examples that are close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line. This approach helps to overcome the overfitting problem posed by random oversampling (Satpathy, 2023).

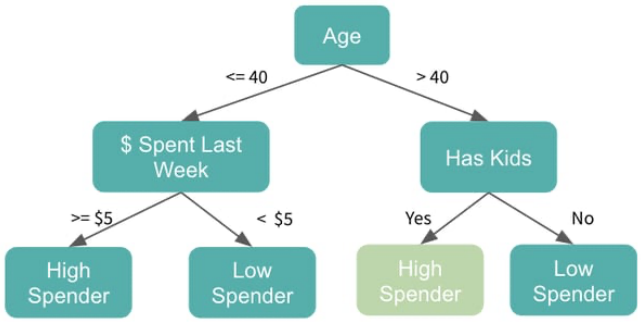


Figur 4 - Illustration of Smote (Satpathy, 2023)

The usage of Smote in our project is applied to improve the performance of the logistic regression model. By balancing the dataset, Smote Logistic Regression can improve the model's performance, especially in terms of sensitivity (the ability of the model to correctly identify diabetic cases). This is crucial in medical diagnostics, where missing a positive case (a person with diabetes) could have serious implications.

VIII. Tree Based

This includes algorithms like Decision Trees, Random Forests, and Gradient Boosted Trees (XGBoost). These are versatile methods used for both classification and regression tasks. The input of the model is being splitted into subsets to build a decision tree, and then each branch is tested regarding accuracy, efficiency and effectiveness. The output of the method, generates the most important variables at the top of the hierarchy and the ones that are irrelevant is being dropped (C3.ai, 2023).



Figur 5 - Example of Tree-Based Model (Gross, 2020)

Regarding our project, Tree-Based Methods can be used to classify patients based on their health indicators. They are particularly useful for interpreting the data, as they provide insights into which features are most important in predicting diabetes.

IX. Results

Logistic Regression without hyperparameters

The results shows high accuracy for class 0, with a accuracy of 0.86. As for the class 1, it does not achieve as great accuracy, reaching 0.54. The F1-score indicates that class 0 have a good performance and class 1 does not.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.98      0.92     38813
         1.0       0.54      0.15      0.23      7082

    accuracy                           0.85     45895
   macro avg       0.70      0.56      0.57     45895
weighted avg       0.81      0.85      0.81     45895

Confusion Matrix:
 [[37913   900]
 [ 6027  1055]]
```

Figure 1 - Results Logistic Regression without hyperparameters

Logistic regression with hyperparameters

The results did not change after we applied the hyperparameters, telling us that hyperparameter tuning did not influence the result on this model.

```
Best hyperparameters: {'C': 0.01, 'penalty': 'l2', 'solver': 'sag'}
Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.98      0.92     38813
         1.0       0.54      0.15      0.23      7082

    accuracy                           0.85     45895
   macro avg       0.70      0.56      0.57     45895
weighted avg       0.81      0.85      0.81     45895

Confusion Matrix:
 [[37934   879]
 [ 6048  1034]]
```

Logistic Regression with Smote

Applying Smote to the logistic regression improves the accuracy for class 0, but the accuracy of class 1 is dropping. Now the accuracy for class 1 is at 0.32.

```
Classification Report (with SMOTE):
              precision    recall  f1-score   support

         0.0       0.94      0.71      0.81     38813
         1.0       0.32      0.75      0.45      7082

    accuracy                           0.72     45895
   macro avg       0.63      0.73      0.63     45895
weighted avg       0.84      0.72      0.76     45895

Confusion Matrix (with SMOTE):
 [[27710 11103]
 [ 1799  5283]]
```

Decision Tree

As for the Decision Tree method, it did offer moderate accuracy with a balanced precision between classes. However, recall and F1-scores are relatively low, suggesting room for improvement in class prediction.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.87      0.85      0.86     38813
         1.0       0.29      0.32      0.30      7082

    accuracy                           0.77     45895
   macro avg       0.58      0.59      0.58     45895
weighted avg       0.78      0.77      0.78     45895

Confusion Matrix:
 [[33181  5632]
 [ 4826  2256]]
```

Random Forest

Demonstrates good precision for class 0 and improved precision for class 1 compared to Logistic Regression. The class 1 is having an accuracy of 0.49 with random forest.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.87      0.96      0.91     38813
         1.0       0.49      0.19      0.28      7082

    accuracy                           0.84     45895
   macro avg       0.68      0.58      0.60     45895
weighted avg       0.81      0.84      0.81     45895

Confusion Matrix:
 [[37376  1437]
 [ 5705  1377]]
```

XGBoost

Shows a comparable performance to Random Forest, with slight variations in precision and accuracy, suggesting similar strengths and weaknesses in classification tasks.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.87      0.96      0.91     38813
         1.0       0.48      0.20      0.28      7082

    accuracy                           0.84     45895
   macro avg       0.68      0.58      0.60     45895
weighted avg       0.81      0.84      0.82     45895

Confusion Matrix:
 [[37309  1504]
 [ 5670  1412]]
```

50/50 Dataset

We decided to test the trained model on a dataset which is split into 50/50 of non-diabetes and diabetes. This achieves the highest accuracy with a more balanced performance across classes.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.76      0.73      0.74      7090
         1.0       0.74      0.77      0.75      7049

    accuracy                           0.75     14139
   macro avg       0.75      0.75      0.75     14139
weighted avg       0.75      0.75      0.75     14139

Confusion Matrix:
 [[5157 1933]
 [1625 5424]]
```

Each model shows distinct strengths and weaknesses in handling class imbalances and predictive accuracy. The model with the highest accuracy of class 0 was Logistic Regression with Smote, achieving a accuracy of 0.94. For class 1 the model who achieved the best accuracy was Logistic Regression with and without hyperparameters.

X. Conclusion

In conclusion, this paper demonstrates the effectiveness of various machine learning models in predicting diabetes. Our analysis compared Logistic Regression, Decision Trees, Random Forest, and XGBoost, finding that each model offers unique strengths in terms of accuracy, precision, and recall. The results underscore the potential of machine learning in healthcare, particularly for early diagnosis of diabetes. However, there are limitations related to data diversity and model generalizability that warrant further investigation. Future research should focus on refining these models, incorporating more diverse datasets, and exploring real-world applications to enhance predictive accuracy and healthcare outcomes.

References

[1] C3.ai. (2023). Tree-Based Models https://c3.ai/glossary/data-science/tree-based-models/

[2] Dubey, A. (2018). The Mathematics Behind Principal Component Analysis. https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643

[3] Education Ecosystem. (2018). Understanding K-means Clustering in Machine Learning.

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[4] Gross, K. (2020). Tree-Based Models: How They Work (In Plain English!). https://blog.dataiku.com/tree-based-models-how-they-work-in-plain-english

[5] Jaadi, Z. & Whitfield, B. (2023). A Step-by-Step Explanation of Principal Component Analysis (PCA). https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[6] Kanade, V. (2022). What is logistic regression? Equation, assumptions, types, and best practices. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

[7] Satpathy, S. (2023). SMOTE for Imbalanced Classification with Python. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

[8] Sharma, P. (2023). K Means Clustering – Step-by-Step Tutorials for Clustering in Data Analysis. https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/

[9] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[10] Alex Teboul, Diabetes Health Indicators Dataset, 2015. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset