UNIVERSITETI I PRISHTINËS “HASAN PRISHTINA”
FAKULTETI I SHKENCAVE MATEMATIKO-NATYRORE
DEPARTAMENT OF MATHEMATICS, COMPUTER SCIENCE PROGRAM

Machine Learning
Seminar Project
Data Analysis for Diabetes Prediction Using Health Indicators

Albana Rexhepi, Eljesa Kqiku, Kaltrina Kuka

May 2025

Përmbajtja

1 Introduction

1.1 Dataset description

This dataset contains detailed health indicators collected from a large population and is designed to support the analysis and prediction of diabetes risk. It includes a variety of columns describing lifestyle factors and demographic data, such as high blood pressure (HighBP), high cholesterol (HighChol), body mass index (BMI), smoking and alcohol use, physical activity, diet, mental and physical health status, as well as gender, age, education level, and income.

The main variable of interest is Diabetes_binary, which indicates whether an individual has been diagnosed with diabetes (1) or not (0). This dataset can be used to develop classification models aimed at predicting diabetes risk based on known behavioral and health-related factors.

index	feature	description
Diabetes_binary	0 = no diabetes, 1 = diabetes	Diabetes status
HighBP	0 = no high BP, 1 = high BP	High blood pressure
HighChol	0 = no high cholesterol, 1 = high cholesterol	High cholesterol
CholCheck	0 = no cholesterol check in 5 years, 1 = yes	Cholesterol check
BMI	Continuous	Body Mass Index
Smoker	0 = no, 1 = yes	Smoked at least 100 cigarettes in life
Stroke	0 = no, 1 = yes	Ever had a stroke
HeartDiseaseorAttack	0 = no, 1 = yes	Heart disease or heart attack history
PhysActivity	0 = no, 1 = yes	Physical activity in past 30 days (not job-related)
Fruits	0 = no, 1 = yes	Consumes fruit 1+ times per day
Veggies	0 = no, 1 = yes	Consumes vegetables 1+ times per day
HvyAlcoholConsump	0 = no, 1 = yes	Heavy alcohol consumption
AnyHealthcare	0 = no, 1 = yes	Healthcare coverage
NoDocbcCost	0 = no, 1 = yes	Couldn't see a doctor due to cost
GenHlth	1 = excellent, 5 = poor	General health status
MentHlth	1-30 days	Poor mental health days in the past 30 days
PhysHlth	1-30 days	Physical illness/injury days in past 30 days
DiffWalk	0 = no, 1 = yes	Difficulty walking or climbing stairs
Sex	0 = female, 1 = male	Gender
Age	1 = 18-24, 9 = 60-64, 13 = 80 or older	Age category
Education	1-6	Education level
Income	1-8	Income scale

Table 1: List of all attributes in the dataset

1.2 Motivation of the study

The motivation of this study is based on three main questions related to understanding the risk factors for diabetes and how these factors can be used to predict the likelihood of developing the disease:

1. **What are the main factors that predict the risk of diabetes?** One objective of this study is to identify the health and lifestyle factors most strongly associated with the risk of developing diabetes. These factors, such as `BMI`, `Age`, and `Physical_Health`, can help in early detection and prevention of diabetes.

2. **Can a subset of factors be used to accurately predict whether an individual has diabetes?** Another important question is whether only a few factors, such as `BMI` and `Age`, can be used to make an accurate prediction of diabetes, simplifying the prediction process without losing too much accuracy.

3. **How can machine learning models be used to predict diabetes risk more accurately?** This broader question relates to using advanced techniques to help create a reliable model for predicting diabetes. Machine learning can uncover relationships and patterns that may not be immediately apparent using traditional methods.

Through these questions, the study aims to improve our understanding of the factors influencing diabetes risk and to develop a simple yet accurate method for predicting who may be at risk.

1.3 Selection of algorithms for the dataset

For this dataset, where the target variable is `Diabetes_binary` (diabetes status), three well-known machine learning algorithms have been chosen: **MLP (Multilayer Perceptron)**, **Autoencoder**, and **I TRETI**. These algorithms are suitable for this type of problem for various reasons:

- **MLP (Multilayer Perceptron)**:
  MLP is a type of neural network that uses multiple processing layers and is excellent for handling classification and regression problems. This algorithm is powerful for capturing complex relationships between different features. For our dataset, it is well-suited to model the connections between health indicators and diabetes risk, as it can effectively analyze factors such as `BMI`, `Age`, `Physical_Health`, `Smoking`, and other related variables.

- **Autoencoder**:
  An Autoencoder is a type of neural network used primarily for unsupervised learning and dimensionality reduction. It can help identify hidden patterns in the data and is useful for reducing the number of features while maintaining the underlying structure. For our dataset, it can assist in detecting underlying factors contributing to diabetes risk by learning an efficient representation of the data. Autoencoders are also useful for feature extraction, especially in cases where the data is high-dimensional.

- **[Third Algorithm]**:
  flasum per tretin

2 Preprocessing

At this part of the project, we begin with the data preprocessing phase. This step is necessary to prepare the dataset for machine learning, ensuring the data is ready for analysis and modeling.

2.1 Data cleaning

2.1.1 Removing Duplicates

For this step, we first analyze if there are any duplicate rows in the dataset using the following Python code:

```python
import pandas as pd

# Load the dataset
df = pd.read_csv('datasets/diabetes_binary_5050split_health_indicators_BRFSS2015.csv')

# Check for duplicate rows
duplicates = df[df.duplicated()]

# Display the result
if duplicates.empty:
    print("No duplicates found!")
else:
    print(f"Found {duplicates.shape[0]} duplicate rows.")
    print(duplicates)
```

After running the code, we found that there are 1635 duplicate rows. However, since these duplicates could potentially represent individuals who share the same health indicators (such as age, BMI, or smoking status), and not necessarily represent data entry errors, we decided not to remove them. This is because, in medical datasets, it is common for multiple individuals to have identical data points across various features. Therefore, removing these rows could lead to loss of valid data.

```
Found 1635 duplicate rows.
```

As shown, there are 1635 duplicate rows, but they are not removed to ensure we preserve all relevant data for analysis.

2.1.2 Handling missing values

In this step, we aim to check whether the dataset contains any missing values. Missing data can impact the accuracy of machine learning models, so identifying and addressing it is an important part of the data preprocessing process. Below is the Python code used for this task:

2.1.3 Handling outliers

The first step we used to analyze the outlier or noise values was checking the min, max, average and mode of each attribute. From these data we can conclude that there are no noises since all the min-max fields are within the range declared on the metadata.

Next, we applied the interquartile range (IQR) method to identify outliers. However, this approach was not very effective, as it flagged 40,205 records as containing at least one outlier attribute—more than half of the dataset. Because this didn't seem like a logical result, we decided to ignore this method and try something else.
Finally, we used the z-score method, which was more suitable for our dataset. This method found 1,927 outlier records, which is only 2.7% of the total dataset. Since this number was small, we thought it was reasonable to remove these records without further analysis, as we still had enough data left to work with.

```python
import pandas as pd

# Load the dataset
df = pd.read_csv('datasets/diabetes_binary_5050split_health_indicators_BRFSS2015.csv')

# Check for missing values
missing_values = df.isnull().sum()

# Display missing values per column
print(missing_values)
```

The output after executing the script was:

```
Diabetes_binary         0
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
```

```
MentHlth              0
PhysHlth              0
DiffWalk              0
Sex                   0
Age                   0
Education             0
Income                0
```

As we can see, all columns in the dataset have a count of 0 missing values. This means the dataset is complete in terms of data presence, and no additional cleaning or imputation for missing values is necessary.

2.2 Data integration

In this project, all the required data is already combined into a single CSV file. Therefore, no additional integration from multiple sources is necessary. The dataset is self-contained and ready for the next steps of data preprocessing.

2.3 Data transformation

In this step, we apply feature scaling to the dataset in order to standardize the features. Scaling is necessary because some machine learning algorithms, such as the ones we have chosen for this analysis, are sensitive to the scale of the data. These algorithms work better when all features have a similar range.

```python
import pandas as pd
import numpy as np
from scipy.stats import zscore

df = pd.read_csv('datasets/diabetes_binary_5050split_health_indicators_BRFSS2015.csv')

####### Making a summary of the fields #######
min_vals = df.min()
max_vals = df.max()
avg_vals = df.mean()
mode_vals = df.mode().iloc[0]
summary_df = pd.DataFrame({
    'Min': min_vals,
    'Max': max_vals,
    'Mean': avg_vals,
    'Mode': mode_vals
})
print("####### Summary of the dataframe #######")
print(summary_df)


print("\n####### Checking for outliers with Interquartile Range (IQR) method #######")
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
outliers = (df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))
outlier_counts = outliers.sum(axis=1)
print("Rows with >=1 outliers:", (outlier_counts >= 1).sum())


print("\n####### Checking for outliers with Z-Score Method (Assumes Normal Distribution) #######")
z_scores = df.select_dtypes(include='number').apply(zscore)
outliers = (z_scores.abs() > 5)
print("Outlier rows (Z-score):", outliers.any(axis=1).sum())


####### Removing the outliers #######
rows_to_remove = outliers.any(axis=1)
```

```
cleaned_df = df[~rows_to_remove]
cleaned_df.to_csv('datasets/dataset_without_outliers.csv', index=False)
print("Cleaned dataset saved as 'dataset_without_outliers.csv'")
```

2.3.1 Data scaling

In this step, we apply feature scaling to the dataset in order to standardize the features. Scaling ensures that all features have a similar range. The StandardScaler from the `sklearn.preprocessing` library is used to perform scaling.

The following Python code demonstrates the process of scaling the features of the dataset:

```
from sklearn.preprocessing import StandardScaler
import pandas as pd

# Load the dataset
df = pd.read_csv('datasets/dataset_without_outliers.csv')

# Separate features and target
X = df.drop(columns=['Diabetes_binary'])  # Features
y = df['Diabetes_binary']                 # Target

# Initialize and apply the scaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Convert scaled features back to DataFrame
X_scaled_df = pd.DataFrame(X_scaled, columns=X.columns)

# Add the target column back
X_scaled_df['Diabetes_binary'] = y.values

# Save to a new CSV file
X_scaled_df.to_csv('datasets/diabetes_scaled.csv', index=False)

print("Scaled dataset saved as 'datasets/diabetes_scaled.csv'.")
```

This code scales all the features of the dataset using the `StandardScaler`, which standardizes the features by removing the mean and scaling to unit variance. The scaled features are then saved in a new CSV file called `diabetes_scaled.csv`.

3 Prepare for analysis

3.1 Correlation between features

Understanding how features interact within the dataset is a key step before building predictive models. One effective way to explore these interactions is through correlation analysis. This technique reveals the strength and direction of linear relationships between variables, helping us detect redundant features, highlight those most relevant to the target variable, and optimize the feature set for better model efficiency and accuracy. By refining the input space early on, we set a strong foundation for the performance of our machine learning models.

One important aspect of correlation analysis is examining how strongly each feature relates to the target variable, in this case, Diabetes_binary. Identifying variables that show a higher correlation with the target can guide us in selecting the most predictive features for our model. On the other hand, features that exhibit very weak relationships with the target may contribute little to the predictive power and can be excluded to simplify the model. Additionally, correlation analysis helps reveal whether certain features are highly related to each other. When two features show a strong correlation such as above 0.5 it may indicate redundancy. Retaining both could introduce multicollinearity, which not only complicates model interpretation but can also impair the algorithm's performance. In such cases, it's often beneficial to remove one of the correlated features to streamline the dataset and enhance model robustness.

To calculate the correlation between features, we used the `corr()` function from the `pandas` library. We visualized the results using a Heatmap. A heatmap is a graphical representation that uses color gradients to indicate the strength of relationships in a two-dimensional matrix. This is especially useful for identifying patterns and connections between features.

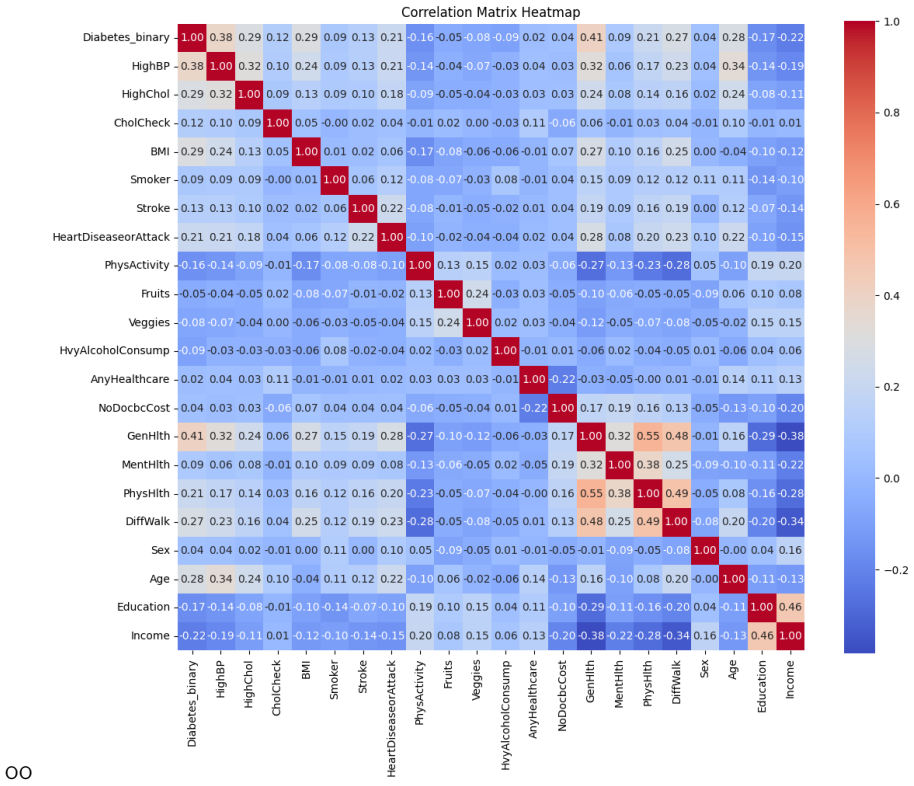
The implementation can be seen in the code below, and the visual results are presented in Figure 1:

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv('datasets/diabetes_binary_5050split_health_indicators_BRFSS2015.csv')

# Compute correlation matrix
corr_matrix = df.corr()

# Plot the heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap="coolwarm", square=True)
plt.title("Correlation Matrix Heatmap")
plt.tight_layout()
plt.savefig("report/images/diabetes_correlation_matrix.png")
plt.close()
```



oo

Figure 1: Correlation matrix between features.

From the heatmap visualization, we observe that some features are more strongly correlated with each other. We then filtered the feature pairs that have an absolute correlation higher than 0.45. The implementation of this filtering process in Python is shown below:

```python
import pandas as pd
import numpy as np

df = pd.read_csv('datasets/diabetes_binary_5050split_health_indicators_BRFSS2015.csv')
corr_matrix = df.corr()

# Unstack and filter correlations
high_corr = (
    corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(bool))
    .stack()
    .reset_index()
)
high_corr.columns = ['Feature 1', 'Feature 2', 'Correlation']
filtered_corr = high_corr[high_corr['Correlation'].abs() > 0.45]
print(filtered_corr.sort_values(by='Correlation', ascending=False))
```

The results are displayed in Table 2:

Feature 1	Feature 2	Correlation
GenHlth	PhysHlth	0.552757
PhysHlth	DiffWalk	0.487976
GenHlth	DiffWalk	0.476639
Education	Income	0.460565

Table 2: Feature pairs with correlation $c > |0.45|$

4 Algorithm implementation

Next, we will implement the three chosen algorithms: MLP (Multilayer Perceptron), Autoencoder, and [Third Algorithm]. These algorithms have been selected due to their suitability for classification problems and their ability to handle structured and complex datasets. First, we will provide a brief introduction to each algorithm, highlighting their key characteristics. After that, we will proceed with the implementation and performance analysis for our dataset.

4.1 Multilayer Perceptron (MLP)

The **Multilayer Perceptron (MLP)** is a core model in the field of machine learning, used for both classification and regression tasks. MLP consists of three primary layers: the **input layer**, one or more **hidden layers**, and the **output layer**. Each layer contains multiple neurons, and neurons from one layer are fully connected to neurons in the next layer. This dense connectivity allows MLP to capture complex relationships in data, particularly non-linear ones, which are common in many real-world applications [1].

In the context of our dataset, which includes health-related features such as `BMI`, `Age`, `Physical_Health`, and `Sleep_Time`, MLP is well-suited to identify hidden patterns and make predictions about whether an individual has diabetes or not. The network learns from the data through a process called *backpropagation*, where errors from the output are propagated back through the network to adjust the weights of the connections between neurons. This process is combined with *gradient descent*, an optimization technique that minimizes the prediction error by adjusting the weights during each iteration of training. Over time, the model learns the optimal weights, improving its ability to make accurate predictions [2].

The basic structure of an MLP is illustrated in the figure below. The input features, such as `BMI` and `Age`, are fed into the input layer. From there, they pass through one or more hidden layers, where neurons transform the data by applying learned weights and activation functions. The transformed data then flows to the output layer, which produces the final prediction, such as whether an individual has diabetes.

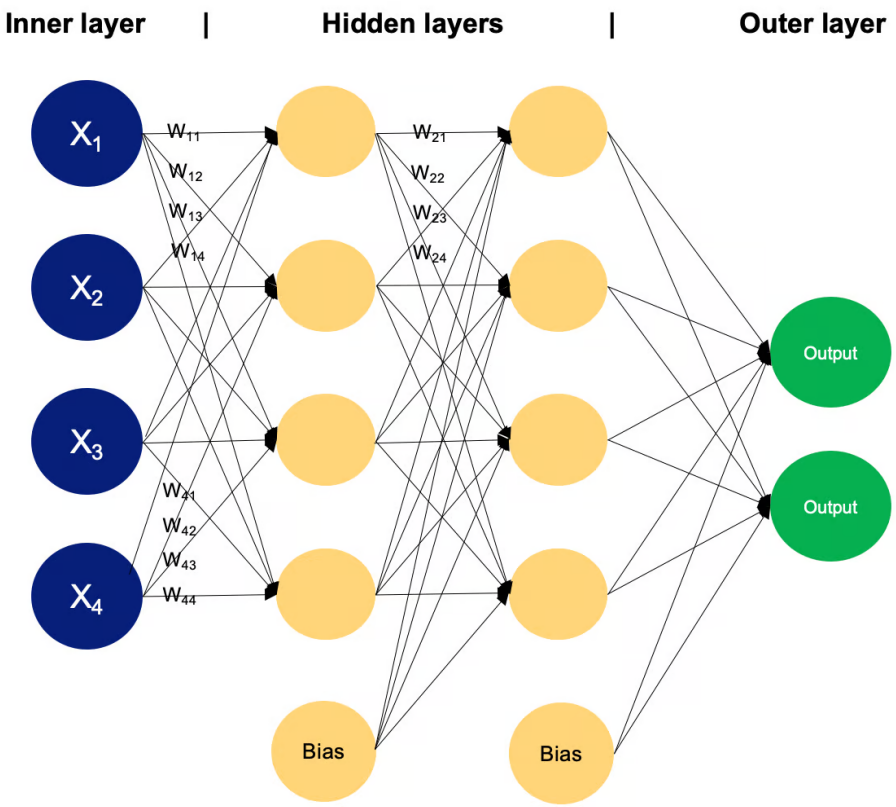


Figure 2: Basic structure of a Multilayer Perceptron (MLP) [3].

In the image, you can observe the flow of information from the input layer through the hidden layers to the output layer. Each layer plays a critical role in transforming the input data to make sense of the patterns and relationships. The neurons in the hidden layers learn features in the data, and as the network adjusts its weights through training, it becomes more accurate in predicting the target outcome.

MLPs are particularly effective in tasks where the relationships between input features and the target variable are non-linear, as is the case in health-related predictions like diabetes classification [1]. This model's ability to learn from complex and large datasets allows it to handle a wide variety of problems, from image recognition to medical diagnoses.

What makes MLPs particularly well-suited for our dataset is their capacity to model intricate relationships between multiple features simultaneously. By learning these relationships, MLPs can make accurate predictions, even when the data contains complex, interdependent factors. This makes MLP an ideal choice for predicting the likelihood of diabetes based on a variety of health indicators [3].

4.2 Autoencoder

4.3 Third algorithm

5 Result interpretation

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. Available: `https://www.deeplearningbook.org`

[2] Michael Nielsen. *Neural Networks and Deep Learning*. 2015. Available: `http://neuralnetworksanddeeplearning.com`

[3] *Multilayer Perceptrons in Machine Learning*. `https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning`

heey