Reflective Report on Portfolio 4

Name = "Pradyot Jain"

Student_id = "48479985"

Word Count = 984

1.01 Objective:

In Portfolio 4, I worked on two key objectives:

1. Predicting Diabetes Risk Using Regression and Neural Network Models with Cross Validation, Feature Selection, and KNN Optimization.
2. Identifying Key Risk Factors for Diabetes Using GaussianNB, CategoricalNB, and Bayesian Network Analysis.

These objectives were chosen to predict diabetes risk using multiple machine learning models and to identify the most influential risk factors. The analysis was aimed at both improving model prediction performance and interpreting key variables contributing to diabetes risk.

1.02 Process of Solving Problems and Learning to Use Notebooks:

At the start of this project, I was unfamiliar with Jupyter notebooks and machine learning techniques, but I gradually gained confidence in structuring data analysis, building models, and interpreting results.

The initial step was cleaning the dataset, where I addressed missing values, removed duplicates, and ensured that both categorical and continuous features were accurately identified. After this, I performed **Exploratory Data Analysis (EDA)** to understand the relationships between features like age, blood glucose level, HbA1c levels, and smoking history, and how they related to diabetes risk. EDA revealed key patterns and helped select features for further analysis.

To prepare the data for modeling, I applied **One-Hot Encoding** to categorical features such as smoking history and gender, converting them into numerical values suitable for machine learning models. I split the data into training and testing sets to ensure model performance could be evaluated on unseen data.

I used **Recursive Feature Elimination (RFE)** to select the most relevant features, enhancing model accuracy by focusing on influential variables. This process prevented overfitting by removing irrelevant features. I experimented with several models, including **Linear Regression**, **Polynomial Regression**, **K-Nearest Neighbors (KNN)**, and **Neural Networks**.

I optimized hyperparameters and evaluated model performance using metrics like accuracy, F1 score, and AUC.

To further explore diabetes risk factors, I used **Gaussian Naive Bayes**, **Categorical Naive Bayes**, and **Bayesian Network Analysis**. The Bayesian Network revealed the interactions between key features like age, HbA1c levels, blood glucose levels, and smoking history, offering a deeper understanding of how these factors contribute to diabetes risk. This helped in determining the probability of diabetes by analysing dependencies between variables like age and smoking history.

1.03 Why I Chose the Dataset:

I chose the diabetes dataset because it contained a balanced mix of continuous and categorical features, making it an excellent choice for applying a range of machine-learning models. The variables in the dataset, such as age, blood glucose levels, and smoking history, are well-known risk factors for diabetes, making it particularly relevant for predictive modeling. The dataset's structure allowed me to explore both regression-based models and probabilistic approaches to understand diabetes risk factors.

1.04 Progress from the Start of the Unit and Future Interests:

When I first started with machine learning models, understanding their differences and choosing the best one was challenging. Through hands-on practice, I developed a better grasp of how these models function and how to interpret their results. I also became more comfortable using Jupyter Notebooks for organizing and executing data analysis tasks. A key lesson I learned was addressing overfitting—initially, some models performed well on training data but poorly on testing data. By applying cross-validation and feature selection techniques like RFE, I improved model performance. This project has sparked my interest in applying machine learning in healthcare for disease prediction and prevention, and I'm eager to explore how these models can support healthcare professionals in making informed decisions.

1.05 Discussion Points:

1. Predicting Diabetes Risk Using Regression and Neural Network Models:

In this part of the project, I experimented with multiple models, including **Linear Regression**, **Polynomial Regression**, **KNN Regression**, **KNN Classification**, and **Neural Networks**.

- **Neural Networks** outperformed the other models due to their ability to capture complex non-linear relationships between features. This model achieved the highest accuracy and F1 scores in predicting diabetes risk.
- **KNN models** also performed reasonably well, but they required careful tuning of hyperparameters, especially the number of neighbors (k), to avoid overfitting.
- **Linear Regression** and **Polynomial Regression** provided valuable insights into the relationships between individual features and diabetes risk, but they were less effective than more complex models like Neural Networks.

2. Identifying Key Risk Factors Using GaussianNB, CategoricalNB, and Bayesian Network Analysis:

In addition to regression models, I applied **GaussianNB** for continuous features and **CategoricalNB** for categorical features. I also used **Bayesian Network Analysis** to explore feature interdependencies.

- **GaussianNB** performed well in modelling continuous features, providing strong predictive power with an AUC of 92%. This model offered probabilistic insights into how age, blood glucose levels, and HbA1c levels influenced diabetes risk.
- **CategoricalNB** was used to model categorical variables like smoking history and gender. While the accuracy was decent, the model struggled with the precision and F1 scores, particularly due to the distribution of the data.

Bayesian Network Analysis revealed important interactions between features, such as the relationship between age, smoking history, and blood glucose levels in determining diabetes risk. The network helped uncover how combinations of risk factors increase the likelihood of diabetes.

1.06 Conclusion:

This project provided valuable insights into the technical aspects of machine learning and how these techniques can be applied to healthcare data analysis. Through the use of Jupyter notebooks, I was able to iteratively explore various models, optimize their performance, and interpret the results in a structured manner.

Neural Networks emerged as the best model for predicting diabetes risk, due to their ability to capture complex, non-linear relationships between variables. Meanwhile, **GaussianNB** provided a solid probabilistic framework for understanding how continuous features influence the probability of developing diabetes.

Bayesian Network Analysis contributed significantly to understanding the relationships between different risk factors and how combinations of factors affect diabetes risk. This multi-faceted approach, combining regression models, classification models, and probabilistic analysis, deepened my understanding of diabetes prediction and the key factors involved.

In conclusion, the combination of these techniques allowed me to create a comprehensive analysis of diabetes risk factors and prediction, highlighting the importance of using multiple models to both predict and interpret healthcare-related outcomes.