Project Report: Predicting Diabetes - Classification Using KNN

Predicting Diabetes - Classification Using KNN

Intern Name: Pyata Nandini

Internship Title: Tech Internship at Cloudcredits

Internship Duration: April 4, 2025 - June 4, 2025

Project Title: Predicting Diabetes - Classification Using KNN

Submission Date: [June 4, 2025]

Acknowledgment

I extend my sincere gratitude to Cloudcredits for providing this opportunity to learn and work on machine learning applications. I am also thankful to my mentors for their valuable guidance and encouragement throughout this internship period.

Table of Contents

Introduction

Project Report: Predicting Diabetes - Classification Using KNN

Classification problems are fundamental in machine learning, especially in domains like healthcare, finance, and fraud detection. This project involves building a supervised learning model using the K-Nearest Neighbors (KNN) algorithm to classify whether a person is likely to be diabetic based on health parameters. The project leverages the well-known PIMA Indian Diabetes Dataset to evaluate the performance of the model.

Objective

To design, train, and evaluate a KNN classification model that predicts diabetes based on patient data, aiming to support medical diagnostics and early prevention strategies.

Technology Stack

- Programming Language: Python
- Development Tools: Jupyter Notebook / Google Colab
- Libraries Used:
  - pandas
  - numpy
  - matplotlib
  - seaborn
  - scikit-learn

Methodology

1. Data Collection: Used publicly available PIMA Indian Diabetes dataset.
2. Data Preprocessing: Checked for missing or zero values. Normalized features using StandardScaler. Applied label encoding where needed.
3. Splitting Dataset: 80% training and 20% testing split.
4. Model Training: Used KNeighborsClassifier. Optimal k selected with cross-validation.

5. Evaluation: Used accuracy, precision, recall, F1-score, confusion matrix and heatmaps.

Implementation Details

Code snippet provided in notebook format using sklearn's KNeighborsClassifier, train_test_split, accuracy_score, and seaborn heatmap for confusion matrix visualization.

Results and Evaluation

- Achieved approx. 75-80% accuracy.

- Balanced precision and recall.

- Few false predictions in the confusion matrix.

- Scope for enhancement via hyperparameter tuning or advanced models.

Conclusion

The KNN model proved effective for diabetes classification. Simple yet accurate for medium-scale datasets. Future work can explore advanced ML models for better results.

References

- https://scikit-learn.org/

- https://pandas.pydata.org/

- https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

- https://matplotlib.org/

- https://seaborn.pydata.org/

Appendix

Code available in 'knn_diabetes_classification.ipynb' with full implementation and evaluation steps.