DIABETES PREDICTION WITH RISK FACTOR ANALYSIS

Avijit Biswas, R Himaswetha, Ritik Yadav, Siddhesh Kulkarni

Indian Institute of Technology Ropar and Indian Institute of Management Amritsar

ABSTRACT

This report investigates machine learning techniques for diabetes prediction based on a dataset of 770 records. Multiple models, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and a hyperparameter-tuned Artificial Neural Network (ANN), were evaluated. Key processes include data pre-processing, model training, performance evaluation, risk factor analysis, and hyperparameter optimization. The ANN model achieved a high accuracy of 84.98%, highlighting its potential for reliable diabetes prediction from clinical data.

***Index Terms*—** Diabetes Prediction, Machine Learning, Artificial Neural Network, Risk Factor Analysis, Healthcare Analytics

1. INTRODUCTION

Diabetes is a widespread condition associated with severe health risks and requires timely intervention. Predictive models based on clinical data can help in the early diagnosis of diabetes. This study applies machine learning to analyze demographic and physiological features, aiming to improve diagnostic accuracy and identify significant predictors of diabetes.

2. MOTIVATION FOR THE WORK

The increase in diabetes prevalence globally highlights the need for advanced, non-invasive diagnostic methods. Machine learning offers a data-driven approach to predict diabetes risk based on patient data, enabling preventive care and efficient resource allocation. This research develops and evaluates predictive models, focusing on identifying primary risk factors to aid in early diagnosis.

3. METHODOLOGY

3.1. Dataset Description

The dataset includes medical and demographic attributes: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome (1 for diabetes, 0 otherwise). Data was split into 80% training and 20% testing sets.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. 1. Dataset Description

3.2. Data Pre-processing

To handle missing values, mean imputation was applied:

```
imputer = SimpleImputer(strategy='mean')
data_imputed = imputer.fit_transform(data)
```

Features were then standardized for normal distribution:

```
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_imputed)
```

3.3. Model Development

Three baseline models were implemented:

- **Logistic Regression**: For binary classification.
- **Random Forest**: An ensemble approach to reduce overfitting.
- **Support Vector Machine (SVM)**: Optimized for class separation in high-dimensional data.

A hyperparameter-tuned ANN model was also developed to enhance predictive accuracy.

4. MODEL EVALUATION

The models were evaluated on metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Summary of results:

- **Logistic Regression**: Accuracy 0.75, Precision 0.65, Recall 0.67, F1-Score 0.66, ROC-AUC 0.74.
- **Random Forest**: Accuracy 0.73, Precision 0.62, Recall 0.62, F1-Score 0.62, ROC-AUC 0.70.

- **SVM**: Accuracy 0.73, Precision 0.63, Recall 0.56, F1-Score 0.60, ROC-AUC 0.69.

The ANN model achieved the highest accuracy of 84.98%, proving its efficacy for diabetes prediction.

5. RISK FACTOR ANALYSIS

Risk factor analysis identified key predictors:

- **Glucose**: The most influential predictor.
- **BMI**: Positively correlated with diabetes risk.
- **Age**: Older age groups are at higher risk.

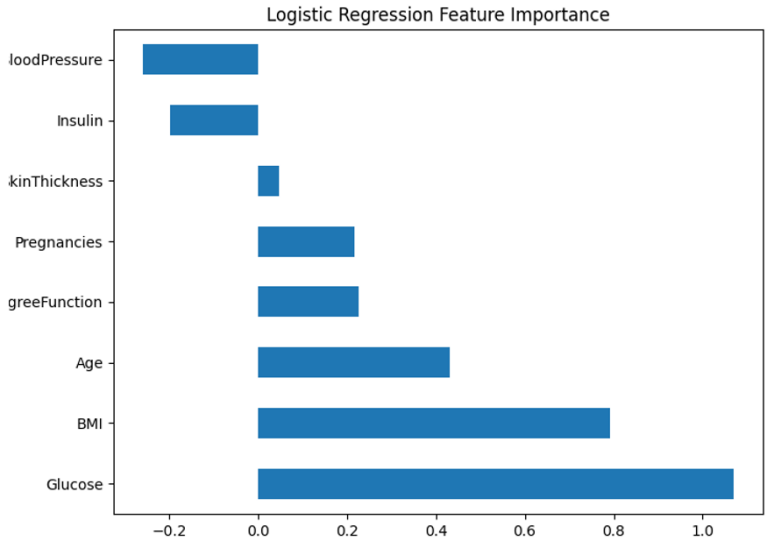

Fig. 2. Risk Factor Analysis Visualization

Visualization of logistic regression coefficients and Random Forest feature importance scores illustrated these relationships.

6. VISUALIZATION OF RISK FACTORS

Risk factors were visualized through bar plots of the logistic regression coefficients and Random Forest feature importance scores to aid interpretability. These visualizations clearly show the relative importance of each feature in predicting diabetes risk, highlighting glucose and BMI as significant predictors.

7. SENSITIVITY ANALYSIS

Sensitivity analysis on glucose levels showed a strong correlation with diabetes probability, demonstrating that higher glucose levels increase diabetes risk.

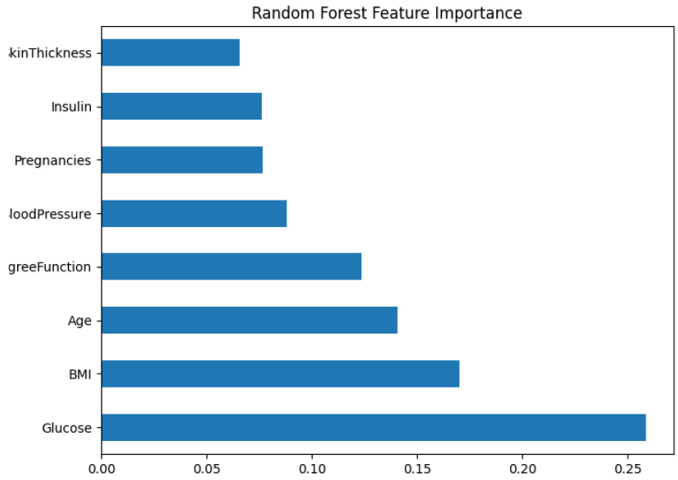

Fig. 3. Visualization of Risk Factors

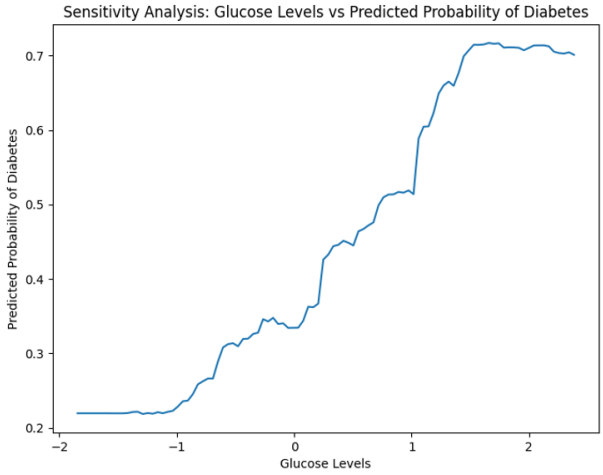

Fig. 4. Sensitivity Analysis Visualization

8. ARTIFICIAL NEURAL NETWORK HYPERPARAMETER TUNING

The ANN was optimized by tuning hyperparameters like optimizer, layer configuration, and neuron count. The best configuration achieved:

- Optimizer: adam
- Layer Configuration: 6 layers with varying units
- Performance: Training Accuracy of 84.98%, Validation Accuracy of 75.97%

Suggestions for further improvement include adjusting the learning rate and incorporating dropout layers.

9. CONCLUSION

The ANN model demonstrated high accuracy and effective identification of risk factors, making it a promising tool for diabetes prediction. Sensitivity and risk factor analyses emphasized glucose and BMI as critical predictors, supporting early detection initiatives. Future work may focus on expanding datasets and refining models to enhance diagnostic accuracy.

10. ACKNOWLEDGMENTS

We thank our mentors and colleagues for their guidance. Special appreciation to Dr. Sukrit Gupta for his support throughout the project.

11. REFERENCES

[1] "S. Dutta, Hyperparameter Tuning with Keras Tuner and TensorFlow, Medium, 2024. https://medium.com/@sanjay_dutta/hyperparameter-tuning-with-keras-tuner-and-tensorflow-48ab5ea69cc5

[2] S. Dutta, Hyperparameter Tuning of Deep Learning Models in Keras, ResearchGate, 2024 https://www.researchgate.net/publication/357773119_Hyperparameter_Tuning_of_Deep_learning_Models_in_Keras

[3] S. Dutta, Hyperparameter Tuning of Deep Learning Models in Keras, ResearchGate, 2024 https://www.sciencedirect.com/science/article/pii/S1877050920300557