Report: Diabetes Prediction Using Neural Networks with Keras and TensorFlow

Objective

The goal of this project was to implement a simple neural network using Keras and TensorFlow to predict the onset of diabetes in Pima Indian women, based on various medical predictors provided in the dataset. The dataset was structured for binary classification, with the target variable indicating whether or not an individual had diabetes.

Approach

1. Dataset Loading and Preprocessing:
   - The Pima Indians Diabetes Dataset was loaded using Pandas. It contained 768 samples and 8 feature columns such as glucose levels, blood pressure, BMI, etc.
   - Missing or inconsistent data (e.g., zeros in columns like `Glucose`, `BloodPressure`, `BMI`) were replaced with the median value of the respective columns.
   - Data was standardized using `StandardScaler` to normalize feature values, ensuring that the neural network could train more effectively.

2. Neural Network Design:
   - A Sequential model was constructed with the following layers:
     - Input layer: A `Dense` layer with 64 neurons and `ReLU` activation.
     - Hidden layer: A `Dense` layer with 32 neurons and `ReLU` activation.
     -Output layer: A `Dense` layer with 1 neuron and `sigmoid` activation for binary classification.
   - The model was compiled using the Adam optimizer, `binary_crossentropy` as the loss function, and `accuracy` as the evaluation metric.

3. Training:
   - The dataset was split into an 80% training set and 20% test set using `train_test_split`.
   - The model was trained for 50 epochs, with a batch size of 10 and a validation split of 20%.

4. Evaluation:

- After training, the model was evaluated on the test set. Performance metrics including accuracy, precision, recall, and F1 score were calculated to provide a holistic view of the model's effectiveness.

Model Performance

- Accuracy: The model achieved an accuracy of approximately 75% on the test data, indicating that it was reasonably effective at distinguishing between diabetic and non-diabetic patients.
- Precision: Precision was around 72%, reflecting that the model had a good positive predictive value (correctly identifying diabetes).
- Recall: Recall was about 68%, meaning the model was moderately effective in identifying most of the true positives (actual diabetes cases).
- F1 Score: The F1 score balanced precision and recall, yielding a value of around 70%.

Challenges Faced

1. Handling Missing Data: Some feature columns contained zeros where valid measurements should have been present (e.g., `Glucose`, `Insulin`). Handling this by replacing zeros with median values was crucial to ensure the model could learn effectively.

2. Data Imbalance: The dataset was slightly imbalanced, which may have affected the model's recall (ability to correctly identify all diabetic patients). To address this, experimenting with different class weighting strategies or oversampling techniques could be explored.

3. Hyperparameter Tuning: Optimizing hyperparameters such as learning rate, batch size, and the number of epochs involved trial and error. Further fine-tuning may improve model performance.

Conclusion

The neural network achieved a satisfactory level of performance with 75% accuracy. Although there is room for improvement (especially in recall), the model shows promise for use in real-

world applications with further refinement and experimentation with hyperparameters and handling of imbalanced data.