Report on Assignment_2_Log_Reg

This report presents a comprehensive analysis of the predictive models built to determine the likelihood of a patient having diabetes based on certain features. Three different modeling approaches were evaluated: Logistic Regression without oversampling, with oversampling, and with cross-validation using Stochastic Gradient Descent (SGD).

Dataset Overview:

The "diabetes" dataset contains medical history features such as Glucose level, Blood Pressure, BMI, etc., along with a target variable indicating whether the patient has diabetes (1) or not (0).

Modeling Approaches:

1. **Without Oversampling:** This approach involved building a logistic regression model without oversampling the data.

   Accuracy: 0.766

   Precision: 0.686

   Recall: 0.636

   F1 Score: 0.660

   Insights: While achieving decent accuracy, the model lacked in recall, indicating a lower ability to correctly identify true positive cases of diabetes.

2. **With Oversampling:** This approach involved oversampling the minority class (patients with diabetes) to balance the dataset before training the logistic regression model.

   Accuracy: 0.765

   Precision: 0.755

   Recall: 0.792

   F1 Score: 0.773

   Insights: Oversampling significantly improved precision, recall, and F1 score, indicating better performance in identifying patients with diabetes while minimizing false predictions.

3. **With Cross-Validation using SGD:** This approach utilized cross-validation with Stochastic Gradient Descent (SGD) to train the model and evaluate its performance.

   Accuracy of 3-folds: [0.72284644, 0.61797753, 0.7443609]

   Cross-Validated Precision: 0.688

   Cross-Validated Recall: 0.712

   Cross-Validated F1 Score: 0.700

   Insights: Cross-validation with SGD provided competitive performance metrics, indicating the model's robustness across different folds of data.

Interpretation of the model coefficients :

Outcome (Target Variable) has

- Moderate positive correlations with Glucose (0.49) and Age (0.24).
- Weak positive correlations with Pregnancies (0.22), BMI (0.31), and Skin Thickness (0.18).
- Weak positive correlation with Insulin (0.18).
- Weak negative correlation with Diabetes Pedigree Function (-0.17).

These correlation values provide insights into the relationships between different features and the likelihood of diabetes. Features with higher correlation coefficients may have a stronger influence on the outcome.

Among the features considered, **Glucose** has the highest impact on the likelihood of diabetes, with a correlation coefficient of 0.492908, which means that as the Glucose level increases, there is a stronger tendency for an individual to have diabetes. **Blood Pressure** has the lowest impact on the likelihood of diabetes, with a correlation coefficient of 0.162986, which means that there is a weaker association between Blood Pressure and the likelihood of diabetes compared to other features in the dataset.

Model Comparison:

Precision: The model with oversampling achieved the highest precision (0.755), indicating its ability to correctly identify patients with diabetes while minimizing false positives.

Recall: Similarly, the oversampled model exhibited the highest recall (0.792), indicating its ability to capture a high proportion of actual positive cases.

F1 Score: The oversampled model also achieved the highest F1 score (0.773), indicating a good balance between precision and recall.

Accuracy: While the accuracy was similar across all approaches, the oversampled model showcased superior performance in terms of precision, recall, and F1 score.

Conclusion:

- Oversampling significantly improves precision, recall, and F1 score, indicating better model performance in identifying patients with diabetes.
- While cross-validation with oversampling using SGD provides insights into model performance across multiple folds, the average accuracy and precision are slightly lower compared to the scenarios without oversampling and with oversampling.
- Overall, oversampling appears to be an effective technique for improving the logistic regression model's performance in predicting the likelihood of patients having diabetes.
- **The logistic regression model with oversampling stands out as the best-performing model**, offering superior precision, recall, and F1 score compared to other approaches. Further fine-tuning and validation on unseen data may be necessary to confirm the model's effectiveness in real-world scenarios.