Executive Summary

Afonso Nunes, Diogo Paixão, Marcos Fernandes

Universidade da Madeira

Executive Summary

This executive summary presents the key findings, insights, and recommendations from our comprehensive exploration of machine learning techniques applied to a given dataset. The project focused on implementing, evaluating, and comparing various machine learning models to extract actionable insights.

Our journey began with the implementation of the k-Nearest Neighbors (kNN) algorithm from scratch using numpy arrays. The implementation process was meticulously documented, followed by application and rigorous performance evaluation on the dataset. Despite its simplicity, the kNN model demonstrated reasonable performance but was limited by generalization issues and computational inefficiency.

Next, we explored supervised learning models using the sklearn library, including Logistic Regression, Decision Tree, and Multilayer Perceptron (MLP). Logistic Regression provided valuable insights into linear relationships within the data, while the Decision Tree model exhibited high accuracy but was prone to overfitting. The MLP model struck a balance between learning and generalization, making it a robust choice for predictive accuracy on unseen data.

We then delved into ensemble models, employing Bagging Classifier and AdaBoost techniques. The Bagging Classifier significantly improved performance metrics, proving more effective in generalization and robustness against overfitting. AdaBoost also enhanced overall performance but was outperformed by the Bagging Classifier in our experiments.

Our feature selection process involved utilizing the feature selector object, which we serialized to save its state, including the selected features. By employing the best model identified during the model building process, namely the Decision Tree model, we performed feature selection to enhance model performance and reduce complexity.

Our exploration extended to deep learning with the implementation of a Convolutional Neural Network (CNN) using TensorFlow. While the CNN demonstrated commendable performance, it was surpassed by the ensemble models, highlighting the importance of considering various model architectures and techniques.

Additionally, we applied clustering algorithms, including K-Means, Gaussian Mixture Model (GMM), Hierarchical Clustering and OPTICS Clustering, to uncover latent patterns within the dataset. These analyses provided valuable insights into the inherent data structures, aiding in the identification of cohesive clusters.

Key Findings and Insights

- **kNN model:** Reasonable performance but limited by generalization issues and computational inefficiency.
- **Logistic Regression:** Effective for linear relationships and highly interpretable but limited by the assumption of linearity.
- **Decision Tree:** High accuracy but prone to overfitting; visual and interpretable.
- **MLP:** Balanced performance for nonlinear data but requires careful tuning and is computationally intensive.
- **Bagging Classifier:** Best overall performance with robustness against overfitting but computationally expensive.
- **AdaBoost:** Enhanced weak learners but sensitive to noisy data.
- **CNN:** Powerful for complex patterns but data and resource-intensive; did not outperform ensemble models in this context.
- **Feature Selector:** Improved model performance by selecting relevant features, enhancing interpretability and reducing complexity.
- **Clustering:** Revealed latent patterns, with each method having unique strengths and limitations.

Recommendations for Further Improvements

- Implement advanced data preprocessing techniques, such as feature engineering and scaling, to improve model performance.
- Conduct extensive hyperparameter tuning for models like MLP and CNN to optimize their performance.
- Apply regularization methods to models like Decision Tree and MLP to reduce overfitting.
- Continue exploring ensemble methods and their variants to enhance predictive performance.
- Explore more advanced deep learning architectures and techniques
- Investigate clustering results to gain deeper insights into data structures and apply clustering algorithms to different data segments for more granular analysis.

In conclusion, this project demonstrates the transformative potential of machine learning in extracting actionable insights from data. By implementing these recommendations, future projects can achieve better model performance, improved generalization, and deeper insights.