**IBM CLOUD APPLICATION DEVELOPMENT PROJECT**

**PHASE 2: INNOVATION**

**Aditya Kishor**

**Dhanalakshmi Srinivasan college of Engineering and Technology**

**Innovation Phase: Incorporating Advanced Machine Learning Algorithms for Predictive Analysis and Anomaly Detection in Big Data**

**TABLE OF CONTENT**

**1. Executive Summary**

- A concise overview of the document's key findings and recommendations.

**2. Introduction**

- Introduction to the importance of predictive analysis and anomaly detection in big data.

- The role of advanced machine learning algorithms.

- Purpose and objectives of the documentation.

**3. Background**

- Understanding Big Data and Its Challenges

- Definition and characteristics of big data.

- Challenges in data volume, velocity, variety, and veracity.

- Machine Learning in Big Data Analytics

- The integration of machine learning into big data analytics.

- Benefits and limitations.

**Load dataset**

file\_names\_range **=** list(range(2009, 2016))

file\_paths **=** [f'/kaggle/input/airline-delay-and-cancellation-data-2009-2018/{file}.csv' **for** file **in** file\_names\_range]

**4. Problem Statement**

- Define the problem: Predictive Analysis and Anomaly Detection

- Real-world scenarios and industries where these techniques are critical.

**5. Data Acquisition and Preprocessing**

- Data Collection Strategies

- Sources, formats, and acquisition methods for big data.

- Data Cleaning and Transformation

- Techniques for handling missing data, outliers, and noise.

- Feature engineering for enhanced predictive modeling.

**Preprocessing**

carrier\_indexer **=** StringIndexer(inputCol**=**"OP\_CARRIER", outputCol**=**"OP\_CARRIER\_Index")

origin\_indexer **=** StringIndexer(inputCol**=**"ORIGIN", outputCol**=**"ORIGIN\_Index")

dest\_indexer **=** StringIndexer(inputCol**=**"DEST", outputCol**=**"DEST\_Index")

*# define onehotencoder for a index columns*

onehotencoder\_carrier\_vector **=** OneHotEncoder(inputCol**=**"OP\_CARRIER\_Index", outputCol**=**"OP\_CARRIER\_vec")

onehotencoder\_origin\_vector **=** OneHotEncoder(inputCol**=**"ORIGIN\_Index", outputCol**=**"ORIGIN\_vec")

onehotencoder\_dest\_vector **=** OneHotEncoder(inputCol**=**"DEST\_Index", outputCol**=**"DEST\_vec")

*# Pipelining the preprocessing stages defined above*

pipeline **=** Pipeline(stages**=**[carrier\_indexer, origin\_indexer, dest\_indexer,

onehotencoder\_carrier\_vector, onehotencoder\_origin\_vector,

onehotencoder\_dest\_vector])

transformed\_df **=** pipeline**.**fit(classify\_df)**.**transform(classify\_df)

**6. Advanced Machine Learning Algorithms**

- Supervised Learning for Predictive Analysis

- Random Forest, Gradient Boosting, Support Vector Machines.

- Unsupervised Learning for Anomaly Detection

- Isolation Forest, One-Class SVM, DBSCAN.

- Deep Learning for Complex Patterns

- Neural Networks and Deep Learning architectures.

- Hybrid Approaches and Ensemble Techniques

- Combining multiple algorithms for improved accuracy.

**Implementation of machine learning**

**import** altair **as** alt

**from** pyspark **import** SparkContext, SparkConf

**from** pyspark.sql **import** SparkSession

**import** pyspark.sql.functions **as** F

**import** pyspark.sql.types **as** T

**from** pyspark.ml.feature **import** OneHotEncoder, StringIndexer, VectorAssembler

**from** pyspark.ml **import** Pipeline

**from** pyspark.ml.classification **import** LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, GBTClassifier

**from** pyspark.ml.evaluation **import** BinaryClassificationEvaluator, MulticlassClassificationEvaluator

**import** warnings

warnings**.**filterwarnings('ignore')

**7. Model Training and Evaluation**

- Data Splitting and Cross-Validation

- Training Procedures and Hyperparameter Tuning

- Evaluation Metrics for Predictive Analysis and Anomaly Detection

- Accuracy, Precision, Recall, F1-score, ROC AUC, etc.

**Training Model**

*define the models*

log\_regress **=** LogisticRegression(labelCol **=** 'label', featuresCol **=** 'features')

decision\_tree **=** DecisionTreeClassifier(labelCol **=** 'label', featuresCol **=** 'features')

rand\_forest **=** RandomForestClassifier(labelCol **=** 'label', featuresCol **=** 'features')

gbt **=** GBTClassifier(labelCol **=** 'label', featuresCol **=** 'features')

log\_regress\_model **=** log\_regress**.**fit(train)

decision\_tree\_model **=** decision\_tree**.**fit(train)

rand\_forest\_model **=** rand\_forest**.**fit(train)

gbt\_model **=** gbt**.**fit(train)

**8. Case Studies**

- Real-world Applications

- Healthcare, Finance, Cybersecurity, and more.

- In-Depth Case Studies

- Detailed examples of applying advanced ML for predictive analysis and anomaly detection.

- Outcomes, challenges, and lessons learned.

**9. Implementation Considerations**

- Hardware and Software Requirements

- Scalability and Parallelization

- Model Deployment and Integration

- Cloud-based solutions, containerization, and real-time processing.

**10. Challenges and Solutions**

- Common Challenges in Big Data Analytics

- Strategies and Solutions for Handling Big Data

- Addressing Class Imbalance in Anomaly Detection

**11. Conclusion**

- Summary of key findings and takeaways.

- The significance of advanced machine learning in addressing big data challenges.

- The potential for innovation and impact.