Shiv Kumar

+91 8696147894 | shivkumarkr21@gmail.com | /shivkumar31 | /shivkumar31

Education

Vellore Institute of Technology **2022 – 2027**
Integrated M.Tech in Computer Science and Engineering (Computational and Data Science) *Madhya Pradesh*

Experience

Software Development Intern **Oct 2025 – Dec 2025**
Imperative Code *Jaipur, India*

- Designed and normalized relational MySQL schemas for large-scale business datasets, enabling structured data storage and improving downstream analytics reliability.
- Performed SQL-based exploratory data analysis (EDA) and optimized multi-table JOIN operations, reducing query runtime by 25% and improving dashboard refresh performance.
- Developed parameterized SQL views and stored procedures to automate KPI monitoring and trend analysis, reducing manual reporting workload by 30%.
- Collaborated with cross-functional teams to translate business requirements into data-driven solutions and analytics workflows.

Projects

Distributed Big Data ETL Pipeline | *PySpark, AWS S3, SQL* **2025**

- Built an end-to-end distributed ETL pipeline using PySpark to ingest, clean, transform, and load 5GB+ semi-structured log datasets into analytics-ready SQL tables.
- Applied data preprocessing, feature transformation, and schema validation techniques to improve data quality and pipeline robustness.
- Optimized Spark transformations using partitioning, lazy evaluation, and caching strategies, reducing processing latency by 35%.
- Implemented automated data validation rules and anomaly detection checks to ensure 99.9% data consistency within the pipeline.

Automated Document Extraction System | *Python, OCR, NLTK, Scikit-Learn* **2025**

- Developed an NLP-based document intelligence system using OCR and text mining techniques to extract structured information from unstructured PDF documents.
- Performed feature engineering using TF-IDF vectorization and trained Naive Bayes classification models achieving 89% accuracy.
- Designed text preprocessing workflows including tokenization, lemmatization, and noise filtering to improve model performance on OCR-generated data.
- Automated document indexing and categorization, reducing manual document handling time by 60%.

Customer Churn Analysis & Prediction Model | *Python, XGBoost, AWS EC2* **2025**

- Conducted exploratory data analysis and statistical feature selection to identify churn indicators from customer lifecycle datasets.
- Developed supervised machine learning models using Random Forest and XGBoost achieving 85% predictive accuracy.
- Evaluated models using classification metrics including precision, recall, F1-score, and ROC-AUC curves to ensure robust model performance.
- Deployed scalable training workflows on AWS EC2 for handling large datasets and improving model training efficiency.
- Generated business insights through data visualization using Matplotlib to support customer retention strategies.

Technical Skills

Programming Languages: Python, SQL, Java, PySpark
Data Science & Machine Learning: Supervised Learning, Unsupervised Learning, Feature Engineering, Model Evaluation, Statistical Analysis, Predictive Modeling
Algorithms: XGBoost, Random Forest, Naive Bayes, K-Means, Regression Models
Big Data & Cloud: Apache Spark, AWS (S3, EC2), Distributed Data Processing, Git, GitHub
Libraries: Pandas, NumPy, Scikit-learn, NLTK, SpaCy, Matplotlib
Core CS: Data Structures and Algorithms, DBMS, Operating Systems, Object-Oriented Programming

Certifications

Introduction to Machine Learning | *NPTEL* **Nov 2024**