Real–Time Car Crash Hotspot Prediction and Analysis in B.C

Gloria Mo, Swaifa Haque, Rojin Lohrasbinejad

GitHub Repo

Live BC Hotspot Map

MOTIVATION

- More than **200,000 vehicle crashes** occur annually in British Columbia
- Drivers and policymakers often lack real-time insights into crash risks
- Our project predicts and identifies crash hotspots by combining historical crash data with real-time traffic and weather information
- **Goal:** Enable proactive, data-driven interventions to prevent collisions *before* they happen

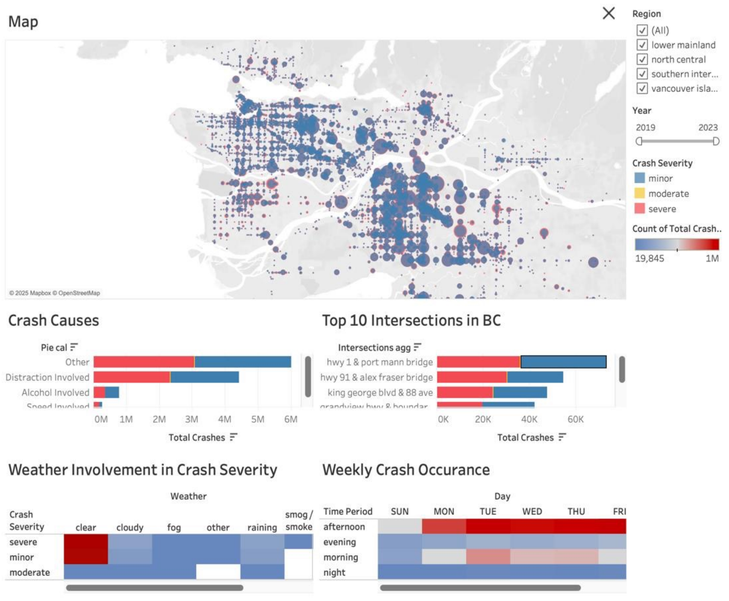


Tableau Dashboard

TOOLS & TECHNOLOGIES

- **Kafka** – Streams live traffic and weather data from APIs
- **Apache Spark (Structured Streaming)** – Real-time data processing and enrichment
- **Docker** – Containerized deployment for reproducibility
- **Spark MLlib** – Builds scalable ML pipelines
- **S3** - Stores ML models and processed data
- **Redis** – Fast, in-memory database to store and access live predictions
- **Tableau** - Interactive dashboard with historical crash trends

SFU SCHOOL OF COMPUTING SCIENCE

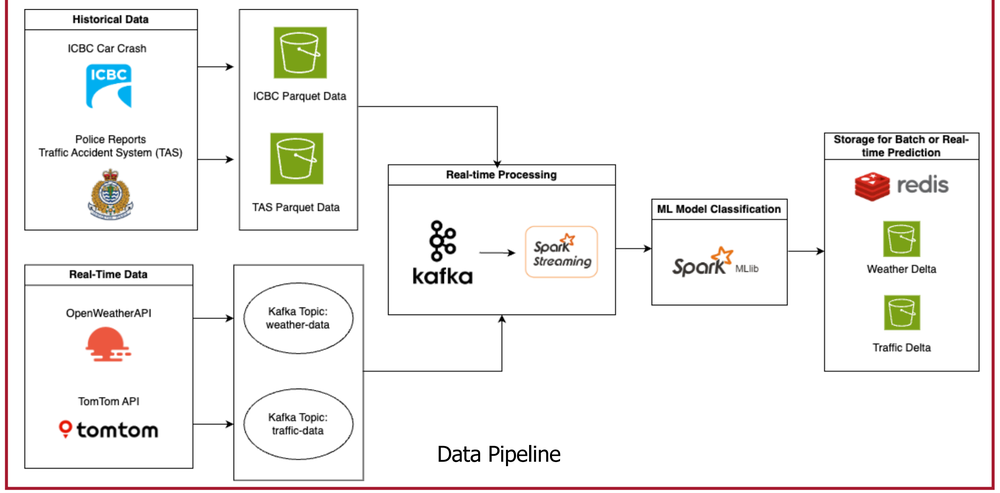
METHODOLOGY

1. Data Collection & Ingestion
- Historical Data: ICBC car crash data & Police-reported data
- Real-time Data: Kafka producers stream weather (OpenWeather API) & traffic data (TomTom API)

2. Batch Model Training (Offline Model Training)
- Combined historical crash data with 7 days of hourly-batched real-time data
- Trained on **Random Forest Classifier**

3. Real-Time Processing (Online Inference)
- Real-time data processed using Spark Structured Streaming
- Merged streams by time and location in 10-minute windows

4. Live Prediction & Storage
- Inputs: Real-time and historical crash features
- Output: Hotspot Risk Level – Low, Moderate, High
- Redis: Stores live predictions with location and metadata
- S3: Archives batch outputs and trained model



Data Pipeline

INNOVATION

- **Real-time** data ingestion and processing with **Kafka + Spark**
- **Low-latency** risk prediction with instant access via **Redis**
- **Modular ML pipeline** supports both batch training and real-time inference
- **Data-driven recommendation system** offering actionable safety interventions

Heatmap of risk zones in Downtown, Vancouver

Latitude	Longitude	Municipality	Hotspot Risk Level	Intervention
48.411	-123.357	Victoria	Low	More Patrols
49.238	-122.868	Coquitlam	Moderate	Install Warning Signs
49.878	-119.432	Kelowna	High	Improve Road Surface

Example ML Outputs with Risk Level & Suggested Interventions

IMPACT

- **Public Safety**: Real-time crash risk warnings for drivers and emergency responders
- **City Planning**: Data-driven insights for infrastructure and urban design
- **Policy-making**: Identifying high-risk zones for proactive regulation
- **Social Benefit**: Enables smarter mobility, reduced collisions, and informed civic decision-making
- **Cost Efficiency**: Helps allocate city resources effectively through targeted interventions
- **Scalability**: Easily adaptable to other regions and real-time data feeds

RESULTS

- Trained a **Random Forest Classifier** for multi-class risk prediction
- Achieved **86% accuracy** with cross-validation and hyperparameter tuning
- Fully **dockerized, modular pipeline** with clear documentation for retraining and reproducibility
- **Recommendation system** suggesting interventions at high-risk zones
- **Interactive Tableau dashboard** to explore historical crash trends
- **Live web-based heatmap** visualizing real-time hotspot predictions

WHAT'S NEXT?

- Develop **real-time alert dashboards** for major incidents and events
- Integrate **external data sources**: road closures, construction, and traffic incidents
- Expand **geospatial risk analysis** across evolving urban areas and nationwide

MASTER'S IN PROFESSIONAL COMPUTER SCIENCE