**YAHAN HU**

Boston, MA | 857-269-4673 | debbiehu@bu.edu

## SUMMARY

Actively seeking research opportunity in the data analysis, data mining relative field. Strengths in programming, SQL, machine learning algorithms, and big data analytical tools.

## EDUCATION

**Boston University, Boston MA** **Jan 2021 - Jan 2022**

*Master of Applied Data Analytics*

- Specialization: Data Mining, Data Structure, Algorithm, Data Science, Web Analytic, Data Analysis, Machine Learning, Big Data Analysis

**Yuan Ze University, Zhongli Taiwan** **Sep 2016 - Jun 2020**

*Bachelor of Industrial Engineering and Management*

- Specialization: Work Study, Human Factor, International Logistics Management, Operation Management, Facilities Planning, Service Management, Decision Analysis, Material Management, Quality Control, Statistic, Experimental of Design, Statistical Analysis

## PROJECTS

**Satellite Image Classification - Tensorflow, Keras, Machine learning, Convolutional Neural Networks** **Dec 2021**

- Created ten convolutional layers neural networks in Python (Tensorflow, Keras). Each two layers followed by max pooling, Batch normalization and dropout layer.

**Mushroom Classification - PySpark, Google Cloud Platform, Big Data, Logistic Regression, Linear SVC** **Oct 2021**

- Generated and preprocessed Spark DataFrame into ready format for analysis with PySpark.
- Implemented algorithm for Multiple Classification based on preprocessed dataset by data cleaning, converted categorical feature transformation.
- Instructed the model with Machine Learning method contain Logistic Regression and Support Vector Machine and optimized with PCA.

**Covid-19 Testing Result in University - Python, Prediction, Model Evaluation** **Aug 2021**

- Gathered the testing result from each school website in US.
- Compiled the Machine Learning model in Python with K-Nearest Neighbors, Logistic Regression, Random Forest, Support Vector Machine to predict the testing result.
- Visualized the evaluation with MAE, MSE, RMSE.

**Football Team Evaluation - R, Statistics method, Correlation analysis, Logistic Regression** **Aug 2021**

- Applied variables in Statistic method (t-test, f-test, ANOVA) with R. Tested the correlation between the features.
- Produced evaluation model with Logistic Regression to discovery the important factors.

**Heart Attack Analysis - R, Feature Analysis, Logistic Regression** **Jun 2021**

- Established a Binary Classification with Logistic Regression to predict if patient has heart attack.
- Analysis each variable with statistics method.

**The Impact of Video Characteristics on Millions Subscribe YouTubers - Weka, Feature Selection, Supervised Classification**

- Collected data with Python and Google YouTube Data API by web crawler. Selected feature with Information Gain, One Rule, Symmetrical Uncertain and self-chosen.
- Constructed Supervised Classification with Naïve Bayes, One Rule, Decision Tree, K-Nearest Neighbors and Logistic Regression for each feature selected method to find important video characteristics for those millions subscribe YouTubers.

**Predicting Primary Lung Cancer by Machine Learning and Statistical Methods – R, LR, DT, SVM, RF, PCA, KNN** **Dec 2019**

- Amassed data through the hospital patient information system by doctor and collate information based on labeled data, processed dataset by data cleaning categorical feature transformation. Attempted complete missing data with K-Nearest Neighbors.
- Trained supervised machine learning models including Logistic Regression, Decision Tree, Support Vector Machine, Random Forest with optimize parameters with PCA to predicted risk of lung cancer.
- Evaluated model performance of classification via K-fold Validation technique and analyzed feature importance to identify top factors that influenced the result.

**Traffic Collision Analysis - R, Statistics method, Correlation analysis** **Jun 2018**

- Processed traffic collision in Statistic method with R.
- Find the correlation between the variable.

**Taoyuan City Parking Management Information - HTML5, CSS, Pingendo, Node-red, JavaScript, MySQL** **Jun 2017**

- Built packing information database with MySQL and Node-red.
- Launched parking information system website with HTML5, CSS, Pingendo and JavaScript to shows the available plot to user.

## SKILLS

- **Programming Software:** Python (Pandas, Sklearn, Scipy, PySpark, Tensorflow, Keras), R, C#, Java, JavaScript
- **Data Science:** Data Mining (Pre-processing, Anomaly Detection), Feature selection, Hypothesis Testing
- **Machine Learning:** Supervised (KNN, SVM, RF, Naïve Bayes, Logistic Regression), Unsupervised (K-Means, PCA), NLP
- **Database/Warehouse:** MySQL, AWS (EC2, S3, EMR), Google Cloud
- **Visualization:** Matplotlib, ggplot2, Google Charts
- **Analysis Tools:** SAS, Microsoft Excel, Weka, JMP Pro, Minitab