VISHAKHA MALHOTRA

drvishakhamalhotra@gmail.com | +1 (415) 910-4469 | San Francisco, CA
· linkedin/in/drvishakhamalhotra · github.com/dr-vish

EDUCATION

University of California San Francisco | Department of Biostatistics and Epidemiology — **San Francisco, CA**
Masters in Health Data Science **GPA:** 3.89 / 4.00 — *July 2022 – June 2024*

Mahatma Gandhi University of Medical Sciences & Technology — **Jaipur, India**
Bachelor of Dental Surgery **GPA:** 3.56 / 4.00 — *October 2015- December 2020*

SKILLS

- **Areas:** Machine Learning, Data Analysis, Data Mining, Data Wrangling
- **Programming Languages:** R**,** Python, Bash, SQL
- **Technologies:** Tableau, Power BI, Orange Data Mining, Microsoft Office Suite, Pandas, spaCy, Tidyverse, Jupyter, NumPy, scikit
- **Statistical Methods:** T-test, Z-test, Hypothesis testing, ANOVA, Correlation, Regression, Survival Analysis

WORK EXPERIENCE

Bakar Computational Health Sciences, Gastroenterology – *Data Science Intern* — *UC San Francisco · May 2023 – Present*
• Employed NLP techniques to clean, reformat and extract data from over 13,000 unstructured patient notes.
• Developed a novel information extraction pipeline to identify medications prescribed by gastroenterologists for Inflammatory Bowel Disease. The pipeline predicted medication recommendations with an accuracy of 90%.
• Experimented with Regex, and spaCy for classifying whether a doctor prescribed the medication in the past, future or present.
Technologies used: Python, Scikit-learn, Pandas, NumPy, spaCy, nltk

Research Facilities Services – *Data Analyst* — *UC San Francisco · July 2023 – Present*
• Queried, organized, analyzed and interpreted data related to seismic bracing from IBM Maximo. Corrected work order data in IBM Maximo and ensured that labs at UCSF were held in accordance with California earthquake codes.
• Provided remote freezer monitoring kits to the labs and programmed the kits with iMonnit with an escalation pathway if the temperature dropped below a designated range.
• Wrote financial contracts with vendors and drove continuous process improvements through participation in standing committees and partner initiatives.
Technologies used: Excel, SQL, iMonnit, IBM Maximo

Teaching Assistant – *BIOSTAT 202- Introduction to Science of Complex Biomedical Data* — *UC San Francisco · July 2023 – Present*
• Graded assignments and projects, held weekly office hours and assisted with computer labs.
• Assisted the course director in developing the course mechanics by updating assignments and adding to the course syllabus.

PROJECTS

Object Detection and Classification in Dental Radiographs — *UC San Francisco · March 2023*
• Developed a machine learning model using Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs) to detect and classify healthy vs diseased teeth in dental images.
• Preprocessed and augmented the data, trained the model, and optimized hyperparameters using grid search and cross-validation.
• Achieved an overall accuracy of 75% on a test set of 1,000 images.
Technologies used: Python, Scikit-learn, OpenCV, Pandas, NumPy, R, SciPy, Scikit-image, dplyr, Keras

HIV Testing Data Analysis — *UC San Francisco · November 2022*
• Coded, cleaned, and visualized the dataset using R programming.
• Conducted HIV testing analysis on 1385 male injecting drug users from 8 Afghan cities. Found living in Mazar-I-Sharif, Jalalabad, and Zaranj to be associated with HIV positive tests and heroin to be the most prevalent substance (99.4%)
• Identified frequent opium usage and old age to be significant predictors for lower adjusted odds ratio (AOR) of recent HIV testing.
Technologies used: R

KIRC Patients Undergoing Nephrectomy — *UC San Francisco · August 2022*
• The study included 243 KIRC patients with a mean age of 61.3 years; almost 37% of patients died during evaluation.
• Performed unsupervised learning using k-means clustering and identified genetic clusters in mRNA, CNV, and miRNA data subsets, and supervised learning using hyperparameter tuning was performed to predict vital status.
• The neural network model with 100 hidden layers performed best for CNV and clinical subset data, while k-nearest neighbors with 5 neighbors performed best for mRNA subset data. The CNV subset model had the highest AUC on test data at 87.5%.
Technologies used: Orange Data Mining

Case study on 'Endodontic Management Of Extra Oral Sinus' — *MGDCH · October 2020 - June 2021*
• Published a research paper in the International Journal of Scientific Research titled **'Endodontic Management Of Extra Oral Sinus in Lower left Canine with CBCT and 6 months follow-up'** that describes the clinical course of a case with extra-oral sinus tract formation, from clinical and radiographic diagnosis using periapical radiographs and CBCT scan and treatment.