CHUNLIN HE, PhD

Data Scientist, Microsoft Certified

Phone: 765-490-7331 • chhe09@gmail.com • https://datute.net

SUMMARY

- Proficient data scientist skilled in Python, utilizing powerful packages such as Pandas, NumPy, Matplotlib, SciPy, Seaborn, SciKit-Learn, Statsmodels, TensorFlow, and more.
- Experienced in developing advanced classification and regression models, employing machine learning algorithms such as Logistic Regression, k-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Regression (SVR), Linear Regression, and Gradient Boosting.
- Demonstrated leadership in managing databases and spearheading data analysis for the multi-million dollar alfalfa toolbox project, driving molecular breeding advancements in the 8-billion dollar alfalfa market.
- Successfully led and executed 100+ data projects in collaboration with 100+ academic and research institutions across 50+ developing countries, generating and analyzing genotypic data for diverse applications.
- Proficient in an extensive array of data science tools, encompassing SQL, SAS, MS SQL Server, MySQL, Jupyter Notebook, Google Colab, RStudio, AzureML, Power BI, Tableau, and Big Data analytics using MS R Client, HDInsight, and Spark.
- Additional expertise and keen interest in GBS (Genotyping-By-Sequencing), RNA-seq analysis, molecular genetics, genomics, and bioinformatics, enhancing data analysis capabilities for biological research.
- Accomplished author and co-author of 58 peer-reviewed papers, conference papers, and book chapters, focusing on statistical and quantitative genetics as well as molecular genetics.

EDUCATION

- Data Science Certificate: Microsoft Professional Program (MPP) for Data Science, Microsoft, USA
- Professional Certificate: Data Science Fundamentals, Microsoft, USA.
- Ph.D.: Plant Science with a focus on Biostatistics, University of Saskatchewan, Canada.
- M.Sc.: Biostatistics and Quantitative Genetics, Nanjing Agricultural University, Nanjing, China.
- B.Sc.: Crop Science, Hunan Agricultural University, Changsha, China.

SKILLS AND KNOWLEDGE

- Computing skills: Python, R, MS R Client, SAS, Minitab, SQL, Git Bash, JavaScript, AzureML, Power BI, Spark, Tableau, HDInsight, MS SQL Server, MySQL, HTML, CSS, Google Colab, Jupyter Notebook.

- Machine Learning (ML) and Deep Learning (DL): Linear Regression, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Gradient Boosting, AdaBoost, XGBoost, CatBoost, Neural networks, computer vision, image analysis, model assessment.
- Data wrangling and statistical analysis: Data pipeline, data wrangling, exploratory data analysis (EDA), hypothesis testing, correlation and regression analysis, Chi-square test, ANOVA, t-test, statistical modeling, principal component analysis, nonparametric statistics.
- Bioinformatics: Genome sequence assembly, annotation, DNA sequence analysis, gene identification, RNA-seq, gene expression analysis, sequence motif, SNP markers, DNASTAR, GeneMapper, MapQTL, QTL IciMapping, Biopython.

PROFESSIONAL EXPERIENCE

Modeller/Data Scientist, Tata Consultancy Services Ltd. | 2020 – present

- Conducted comprehensive assessments of predictive models, identifying gaps for improved model risk management and business performance.
- Specialized in models related to business operations, focusing on predicting customer satisfaction and membership retention.
- Evaluated model accuracy, weaknesses, and limitations through rigorous testing and benchmarking.
- Implemented effective model maintenance, monitoring, and documentation practices to drive continuous business performance improvements.

Microsoft/MPP Data Science Track & Freelance Data Scientist | 2019 - 2020

- Data mining and database management: Collaborated with a private company in the health industry under a non-disclosure agreement (NDA) to perform data cleansing, query and merging operations, and explore relationships within the dataset. Managed and maintained the database effectively to ensure data integrity and accuracy.
- Clinic data analysis: Conducted statistical analysis using R on a comprehensive dataset obtained from a private research organization (NDA). Analyzed data sets aggregated from tens of thousands of participants, providing valuable insights to support decision-making and enhance organizational outcomes.
- Relational database management project: Successfully completed a project for a private company involving the implementation of a relational database management system (RDBMS) using MS SQL Server. Designed and developed a robust database solution, optimizing data storage, retrieval, and management processes.
- Microsoft/MPP machine learning projects: Demonstrated expertise by successfully completing various machine learning projects, addressing real-world data science challenges commonly encountered by corporations and research organizations.

Alfalfa Toolbox/Data Curator, Noble Research Institute, Ardmore, OK | 2015 - 2018

- Curated and analyzed alfalfa data for the multi-million-dollar Alfalfa Toolbox project (ABT), aimed at accelerating molecular breeding for the 8-billion-dollar alfalfa market.

- Conducted data pipeline and exploratory data analysis (EDA) using Python and R on diverse datasets from multi-environmental trials, generating valuable insights.
- Designed and maintained databases using RDBMS (MS SQL Server), efficiently extracting populations of alfalfa with multiple traits of interest for the ABT, partners, and customers.
- Optimized data integration into the ABT web portal through the Toolbox API.

Breeding Services Manager/Senior Scientist, GCP, c/o CIMMYT, Mexico | 2010 - 2014

- Oversaw and managed 100+ data projects, generating genomic and genotypic data for more than 100 academic and research institutions across 50+ developing countries.
- Utilized data analysis techniques, including QTL mapping, to derive valuable insights bridging genotypes and phenotypes for international breeding programs.
- Mentored and guided research scientists, demonstrating data analytics tools and visualizations for effective analysis of genotypic and phenotypic data.