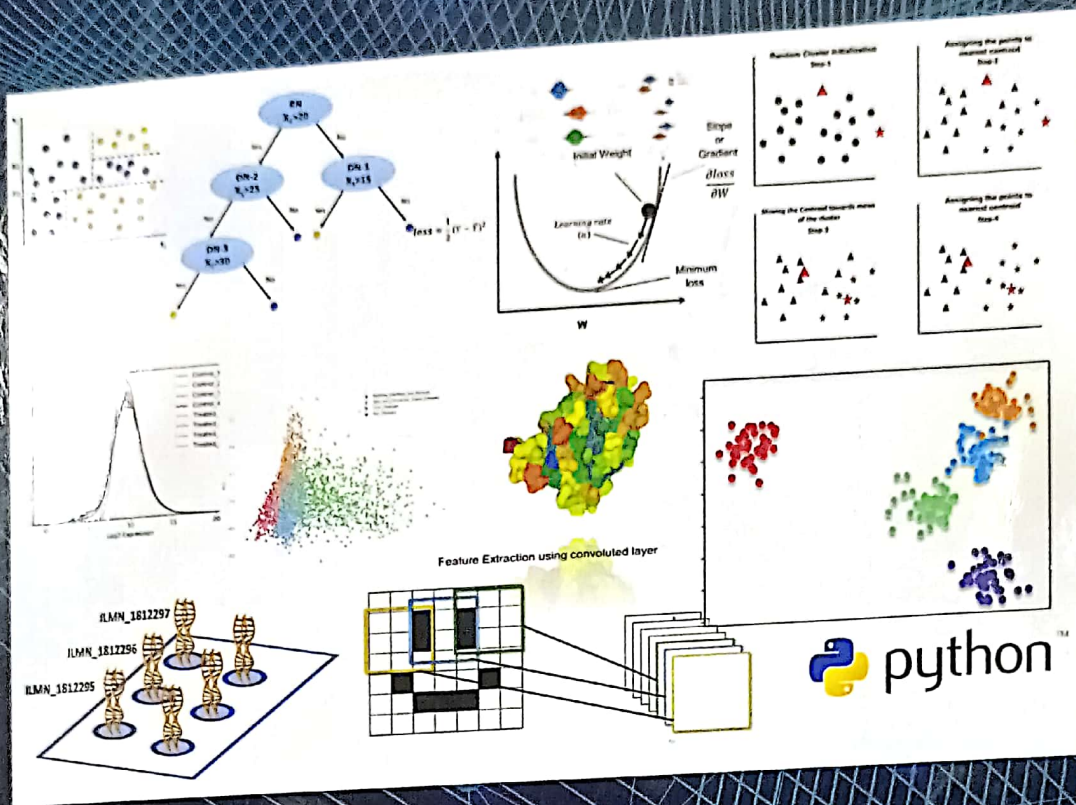


# Hands-on Data Science for Biologists using Python



**Yasha Hasija**  
**Rajkumar Chakraborty**

570.285

H2733 Ha



265879



**CRC Press**  
Taylor & Francis Group





# Hands-On Data Science for Biologists Using Python

Yasha Hasija and Rajkumar Chakraborty



**CRC Press**  
Taylor & Francis Group  
Boca Raton London New York

CRC Press is an imprint of the  
Taylor & Francis Group, an informa business



Scanned with OKEN Scanner

Sc  
570.285  
H2733 Ha

First edition published 2021  
by CRC Press  
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

© 2021 Taylor & Francis Group, LLC  
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

ISBN 13: 978-0-367-54679-3 (hbk)  
ISBN 13: 978-0-367-54678-6 (pbk)

570.285  
H2733 Ha  
  
265879

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyrighted material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under US Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, access [www.copyright.com](http://www.copyright.com) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC please contact [mpkbookspermissions@tandf.co.uk](mailto:mpkbookspermissions@tandf.co.uk)

*Trademark notice:* Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

*Library of Congress Cataloging-in-Publication Data*

Names: Hasija, Yasha, author. | Chakraborty, Rajkumar, author.

Title: Hands on data science for biologists using Python / Yasha Hasija and Rajkumar Chakraborty.

Description: First edition. | Boca Raton : CRC Press, 2021. | Includes bibliographical references and index.

Identifiers: LCCN 2020044939 | ISBN 9780367546793 (hardback) | ISBN 9780367546786 (paperback) | ISBN 9781003090113 (ebook)

Subjects: LCSH: Biology--Data processing. | Python (Computer program language)

Classification: LCC QH324.2 .H373 2021 | DDC 570.285--dc23

LC record available at <https://lccn.loc.gov/2020044939>

Typeset in Times  
by MPS Limited, Dehradun

Visit the Taylor & Francis Web site at  
<http://www.taylorandfrancis.com>

and the CRC Press Web site at  
<http://www.crcpress.com>

Jawaharlal Nehru University  
Accession No. 265879  
Source Varadhiran Books  
Bill No. & Date BL131-27-1-2024  
Price ₹ 69.99  
Centre/School SCIS  
Accessioned by SIM  
Catalogued by



---

# Contents

---

Preface.....	xi
Author Bio .....	xii
<b>1. Python: Introduction and Environment Setup .....</b>	<b>1</b>
Why Learn Python.....	1
Installing Python.....	2
Installing Anaconda Distribution .....	3
Running the Jupyter Notebook .....	3
The Building Blocks of Programs .....	5
Errors in Python.....	5
Exercise .....	6
<b>2. Basic Python Programming.....</b>	<b>7</b>
Datatypes and Operators .....	7
Variables .....	9
Strings .....	11
Lists and Tuples.....	16
Dictionary in Python .....	22
Conditional Statements .....	26
Loops in Python.....	29
Functions.....	33
Classes and Objects.....	37
File Handling in Python .....	40
Exercise.....	43
<b>3. Biopython .....</b>	<b>45</b>
Introduction.....	45
Installing Biopython .....	45
Biopython Seq Class .....	45
Parsing Sequence Files.....	47
Writing Files .....	51
Pairwise Sequence Alignment.....	53
BLAST with Biopython .....	57
Multiple Sequence Alignment.....	59
Construction of a Phylogenetic Tree .....	62
Handling PDB Files.....	64
Exercise.....	70
<b>4. Python for Data Analysis.....</b>	<b>71</b>
Introduction.....	71
NumPy .....	71
NumPy Arrays versus Lists.....	71
Two-Dimensional Matrices .....	73



Matrix Operations .....	74
Comparing Matrices .....	77
Generating Data Using NumPy .....	78
Speed Test .....	79
“Pandas” Dataframe .....	80
Selecting Rows and Columns .....	81
Conditional Filtering in Dataframe .....	84
Writing CSV Files from Pandas Dataframe .....	85
Apply() Function .....	85
Concatenating and Merging .....	87
Exercise .....	89
<b>5. Python for Data Visualization .....</b>	<b>91</b>
Introduction .....	91
Matplotlib .....	91
Matplotlib Functional Method .....	92
Matplotlib Object-Oriented Method .....	93
Resolution and Saving Figures .....	97
Legend .....	98
Customization of the Plot Appearance .....	99
Scatterplot .....	102
Histogram .....	102
Boxplot .....	104
Seaborn .....	104
Distribution Plots .....	105
Joint Plots .....	106
Pairplot .....	108
Barplot .....	111
Boxplot .....	113
Violin Plot .....	114
Heatmaps .....	114
Cluster Maps .....	116
Regression Plot .....	117
Plotly – Interactive Data Visualization .....	118
Geographical Plotting .....	120
Exercise .....	122
<b>6. Principal Component Analysis .....</b>	<b>123</b>
Introduction .....	123
Variance as Information .....	123
Data Transformation .....	124
Case Study .....	125
PCA: Step-by-Step .....	127
Standardization of the Features .....	127
Obtain the Eigenvectors and Eigenvalues .....	128
Choosing Axes with Maximum Variance .....	130
Programing Drive .....	133
Exercise .....	135

<b>7. Hands-On Projects</b> .....	137
Differential Gene Expression Analysis.....	137
Quality Control.....	138
Normalization.....	141
Differential Expression Analysis.....	146
Cluster Map.....	151
Gene Enrichment Analysis.....	152
SNP Analysis.....	153
Exercise.....	160
<b>8. Machine Learning and Linear Regression</b> .....	161
Introduction to Machine Learning and Its Applications in Biology.....	161
Types of Machine Learning Systems.....	161
Optimization of Models.....	165
Challenges in Machine Learning Projects.....	167
Linear Regression.....	169
General Workflow of a Machine Learning Project.....	171
Implementation of Linear Regression Using Scikit-Learn.....	172
Loading Dataset.....	172
Train-Test Split.....	173
Training Model.....	173
Model Evaluation.....	174
Predicting Child Height Based on Parents Height.....	176
Predicting the Height of Sons.....	178
Predicting the Height of Daughters.....	180
Exercise.....	181
References.....	181
<b>9. Logistic Regression</b> .....	183
Introduction.....	183
Implementation of Logistic Regression Using Sklearn.....	184
Train-Test Split.....	187
Training the Logistic Regression Model.....	187
Evaluation of Model.....	187
Retrieving Intercept and Coefficient.....	188
Data Scaling.....	189
Predicting a New Result.....	192
Breast Cancer Prediction Using Logistic Regression.....	193
Model Evaluation.....	194
Exercise.....	196
References.....	196
<b>10. K-Nearest Neighbors (K-NN)</b> .....	197
Introduction.....	197
Implementation of K-NN Using Sklearn.....	198
Loading the Dataset.....	198
Splitting the Dataset into the Training Set and the Test Set.....	198
Training the K-NN Model on the Training Set.....	199
Evaluation with K 1.....	199
Choosing a K-Value.....	199



Data Scaling .....	201
Predicting New Values .....	202
Diagnosing the Liver Disease Using $K$ -NN .....	203
Missing Value Imputation .....	204
Data Scaling .....	205
Splitting the Dataset into the Training Set and the Test Set .....	205
Choosing a $K$ -Value .....	206
Evaluation of the Model .....	208
Exercise .....	208
References .....	209
<b>11. Decision Trees and Random Forests .....</b>	<b>209</b>
Introduction .....	211
Random Forests .....	212
Implementation of Decision Tree and Random Forest Using Sklearn .....	212
Train Test Split .....	212
Decision Trees .....	213
Prediction and Evaluation .....	213
Predicting New Values .....	214
Random Forests .....	214
Prediction and Evaluation of Random Forest Model .....	214
Predicting Prognosis of Diabetes Using Random Forest .....	214
Loading Dataset .....	215
Train-Test Split .....	215
Training Classifier .....	216
Cross-Validation .....	217
Exercise .....	217
Reference .....	219
<b>12. Support Vector Machines .....</b>	<b>219</b>
Introduction .....	219
Kernel Trick .....	221
Implementation of Support Vector Machines Using Sklearn .....	221
Train Test Split .....	221
Train the Support Vector Classifier .....	221
Predictions and Evaluations .....	222
Grid Search .....	223
Prediction of Wheat Species Based on Wheat Seed Data .....	224
Train Test Split .....	224
Training Support Vector Classifier and Tuning Its Parameters Using a Grid Search .....	225
Exercise .....	225
References .....	225
<b>13. Neural Nets and Deep Learning .....</b>	<b>227</b>
Introduction .....	227
Neural Networks Architecture .....	228
The Working Principle of Neural Networks .....	228
Activation Functions .....	229
Steps of Forward Propagation .....	229
Gradient Descent .....	229

Backpropagation .....	230
Implementing Neural Networks Using TensorFlow.....	231
Data Scaling.....	232
TensorFlow 2.0.....	232
Creating a Model.....	232
Model – As a List of Layers.....	232
Model – Adding in Layers One by One.....	233
Building Model.....	233
Training Model.....	234
Overfitting .....	235
Dropout and Early Stopping .....	235
Model Evaluation.....	238
Predicting New Instance.....	239
Predicting Breast Cancer Using Neural Networks.....	239
Separating the Dependent and Independent Dataset.....	239
Data Scaling.....	239
Splitting the Dataset into the Training Set and Test Set.....	239
Creating the Model.....	240
Model Evaluation.....	241
Convolutional Neural Network.....	241
Implementation of CNN Using TensorFlow.....	243
Import Libraries.....	243
Importing the Dataset.....	244
Splitting the Dataset into the Training Set and Test Set.....	245
Building Model.....	246
Training Model.....	247
Model Evaluation.....	248
Exercise.....	249
Reference .....	249
<b>14. The Machine Learning Project.....</b>	<b>251</b>
Introduction.....	251
Importing the Libraries.....	251
Importing the Dataset.....	251
PCA.....	252
Splitting the Dataset into the Training Set and the Test Set.....	253
Training the Logistic Regression Model and Evaluation.....	254
Training the K-NN Model and Evaluation.....	254
Choosing K-Value .....	254
Training the Random Forest Model and Evaluation.....	255
Training the SVM Model and Evaluation.....	256
Training the ANN Model and Evaluation.....	257
Exercise.....	259
Reference .....	259
<b>15. Natural Language Processing.....</b>	<b>261</b>
Introduction.....	261
Vectorizing the Text.....	261
Bag of Words.....	261
TF-IDF .....	262



Classification of Abstracts into Various Categories Using NLP .....	263
Importing the Dataset .....	263
Text Processing .....	264
Label Encoding .....	265
Text Tokenization Bag-of-Words .....	265
Splitting the Dataset into the Training Set and the Test Set .....	267
Building Model .....	267
Model Evaluation .....	267
TF-IDF Implementation .....	267
Splitting the Dataset into the Training Set and the Test Set .....	268
Building Model .....	268
Model Evaluation .....	268
Artificial Neural Networks in NLP .....	270
Splitting the Dataset into the Training Set and the Test Set .....	271
Building Model .....	271
Training Model .....	271
Model Evaluation .....	272
New Prediction .....	272
Exercise .....	273
References .....	273
<b>16. K-Means Clustering .....</b>	<b>275</b>
Introduction .....	275
Implementation of K-Means Clustering Using Sklearn .....	275
Choosing the Number of Clusters .....	277
K-Means Clustering of Genes Based on the Co-Expression .....	279
Exercise .....	283
<b>Index .....</b>	<b>285</b>