LinkedIn Job Posting and Profiles Insights

Chloe and Jacob

Abstract: This project analyzes job postings and LinkedIn profiles in the data field, focusing on insights into the job market and skill requirements. It employs web scraping to gather data, followed by thorough cleaning, exploratory data analysis, and visualization. Using Natural Language Processing, relevant skills are extracted from job descriptions to create a recommendation system for matching job postings with candidate profiles. Additionally, a Random Forest Classifier model predicts job prospects based on education, skills, and experience, evaluated through various metrics and feature importance analysis. The project aims to offer valuable insights for job seekers, employers, and educational institutions in data science and analytics.

1. Data Cleaning and EDA

Job Postings Dataset

	Industries	City	State	job_title_categorized	Job_title	Job_link	Company	Company_link	Post_time	Applicants_count	Job_description	Seniority_level	Employment_type	Job_function
1945	Government Administration	Boston	Massachusetts	data scientist	Life Scientist/E...	https://www.linkedin.com/jobs/view/life-scient...	US Environmental Protection Agency...	https://www.linkedin.com/company/us-epa?trk=pu...	2/6/24	Be among the f...	Help Help Requirements Conditions of Employmen...	Mid-Senior level	Full-time	Research, Analyst, and Information...
833	Technology, Information and Internet...	New York	United States	data engineer	Sr. DevOps Engin...	https://www.linkedin.com/jobs/view/sr-devops-e...	Experfy	https://www.linkedin.com/company/experfy?trk=p...	4/3/23	Over 200 appli...	We are looking for a Senior DevOps Engineer to...	Mid-Senior level	Contract	Information Technology
869	Renewable Energy Semiconductor Manuf...	Austin	Texas	data analyst	Data Analyst, Ne...	https://www.linkedin.com/jobs/view/data-analys...	Tesla	https://www.linkedin.com/company/tesla-motors?...	2/7/24	Over 200 appli...	What To ExpectThe Vehicle Operations team at T...	Entry level	Full-time	Information Technology
285	IT Services and IT Consulting, Softw...	Seattle	Washington	data analyst	Business Analyst...	https://www.linkedin.com/jobs/view/business-an...	Amazon	https://www.linkedin.com/company/amazon?trk=pu...	2/1/24	Be among the f...	DescriptionThe FBA Inventory and Capacity Mana...	Not Applicable	Full-time	Strategy/Planning, Analyst, and In...
3131	Appliances, Electrical, and Electron...	Baltimore	Maryland	data engineer	Quality Engineer...	https://www.linkedin.com/jobs/view/quality-eng...	Tbest Services Inc	https://www.linkedin.com/company/tbestservices...	2/6/24	Be among the f...	TBest Services Inc. is currently seeking a hig...	Mid-Senior level	Full-time	Quality Assurance

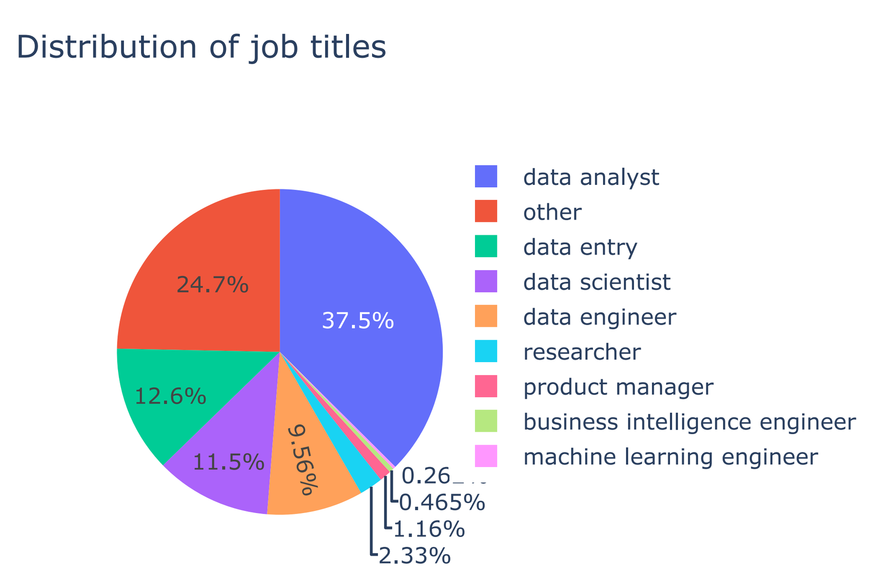
LinkedIn Profiles Dataset

	User Name	Headline	About	Job_title	Experience	Company	Company_size	University	Degree	Degree_type	...	R	Software_development	Git	HTML_CSS	AI	Has_certification	Follower_count	Connections	Uni_ranking	Has_job
11	Christophe H.	Data Scientist at Dropbox	.	data scientist	Data Scientist at Mount Sinai Health System	Dropbox	Big tech	Northern Arizona University	Bachelor of Science - BS, Business Administrat...	Bachelor	...	yes	no	no	no	yes	0	9632.0	9703.0	208.0	1
976	Nan Liu	Machine Learning Engineer	NaN	data engineer	Machine Learning Engineer at DoorDash	DoorDash	Big tech	Fordham University	Master's degree	Other	...	yes	no	no	no	no	0	181.0	181.0	NaN	1
1398	Yiting L.	Data Scientist at Amrock	· Good interpersonal skills, strong work ethic...	data scientist	Data Scientist at Amrock - ...data with NLP to...	Amrock	Other	Beijing International Studies University	Bachelor's Degree, Journalism & English	Bachelor	...	yes	no	no	no	yes	1	341.0	341.0	NaN	1
1013	Niyal Thakkar	Actively seeking Data Analyst Internship | Und...	I am a highly driven undergraduate student at ...	other	Operations Manager at Rutgers University–New B...	Rutgers University–New Brunswick	Grad school	Rutgers University–New Brunswick	Bachelor of Science - BS, Computer Science	Bachelor	...	yes	yes	no	no	no	0	798.0	800.0	NaN	0
707	Guru Prasad Kumar	Senior Data Analyst @ Capital One | Data Analy...	Enthusiastic data analyst with 5 years experie...	data analyst	Senior Data Analyst at Capital One	Capital One	Big tech	SBOA School & Junior College	High School, Computer Science	High School	...	yes	yes	no	no	no	1	1930.0	1929.0	NaN	1

5 rows × 27 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3440 entries, 0 to 3439
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   Industries             3389 non-null   object
 1   City                   3440 non-null   object
 2   State                  3440 non-null   object
 3   job_title_categorized  3440 non-null   object
 4   Job_title              3426 non-null   object
 5   Job_link               3440 non-null   object
 6   Company                3426 non-null   object
 7   Company_link           3426 non-null   object
 8   Post_time              3440 non-null   object
 9   Applicants_count       3439 non-null   object
 10  Job_description        3439 non-null   object
 11  Seniority_level        3439 non-null   object
 12  Employment_type        3389 non-null   object
 13  Job_function           3389 non-null   object
dtypes: object(14)
memory usage: 376.4+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1514 entries, 0 to 1513
Data columns (total 27 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   User Name            1514 non-null   object 
 1   Headline             1511 non-null   object 
 2   About                1093 non-null   object 
 3   Job_title            1514 non-null   object 
 4   Experience           1494 non-null   object 
 5   Company              1369 non-null   object 
 6   Company_size         1514 non-null   object 
 7   University           1383 non-null   object 
 8   Degree               1257 non-null   object 
 9   Degree_type          1514 non-null   object 
 10  Major                1514 non-null   object 
 11  Python               1514 non-null   object 
 12  Java                 1514 non-null   object 
 13  SQL                  1514 non-null   object 
 14  Machine_learning     1514 non-null   object 
 15  Statistical_analysis 1514 non-null   object 
 16  Visualization        1514 non-null   object 
 17  R                    1514 non-null   object 
 18  Software_development 1514 non-null   object 
 19  Git                  1514 non-null   object 
 20  HTML_CSS             1514 non-null   object 
 21  AI                   1514 non-null   object 
 22  Has_certification    1514 non-null   int64  
 23  Follower_count       1398 non-null   float64
 24  Connections          1398 non-null   float64
 25  Uni_ranking          498 non-null    float64
 26  Has_job              1514 non-null   int64  
dtypes: float64(3), int64(2), object(22)
memory usage: 319.5+ KB
```



3440

Job postings

8

Cities

1514

Profiles

Distribution of industries

2. Text analysis of job descriptions

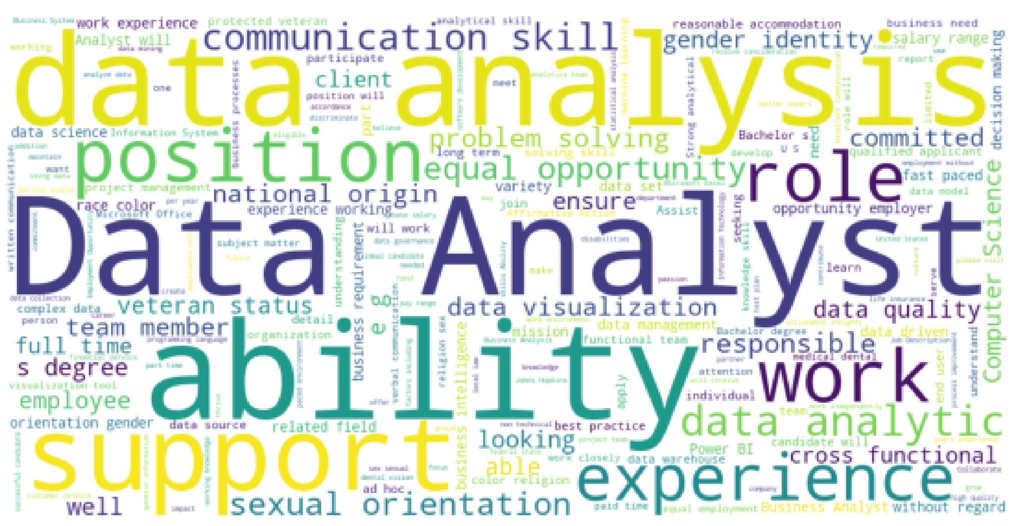
- **Text Preprocessing:** The job descriptions are preprocessed by converting the text to lowercase, removing punctuation, and other cleaning operations.
- **Tokenization:** The preprocessed job descriptions are tokenized into individual words or terms using NLTK's word_tokenize function.
- **Stop Word Removal:** Common stop words (e.g., 'the', 'and', 'is') that do not contribute significantly to the meaning of the text are removed from the tokenized job descriptions using NLTK's stopwords.
- **Skill Filtering:** A list of relevant skills is defined, and the tokenized job descriptions are filtered to keep only the relevant skills.

Choose your job title

data analyst

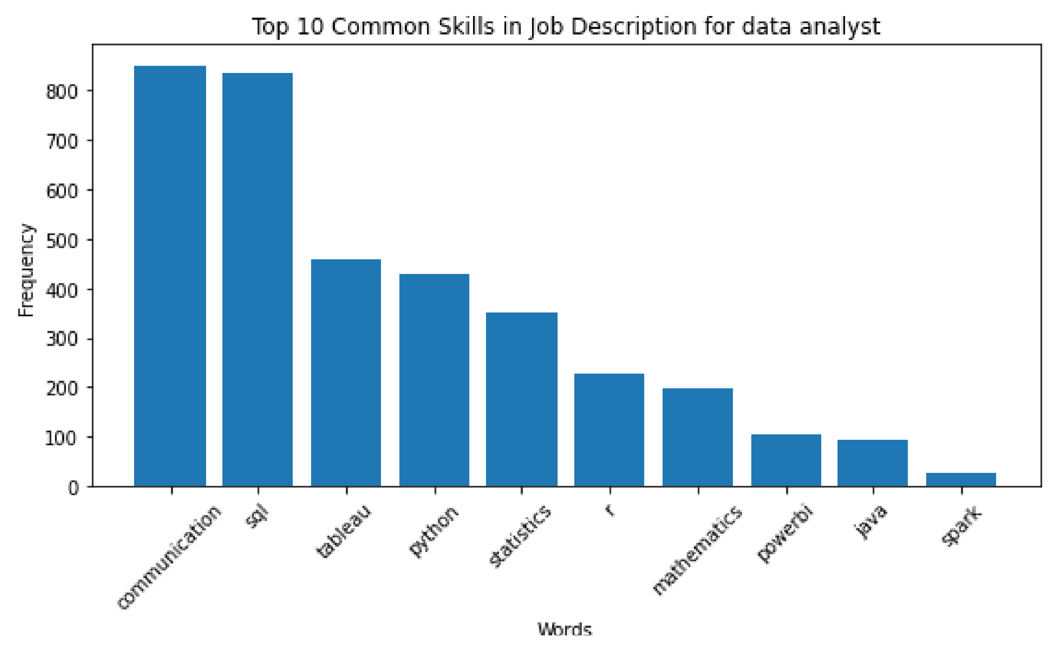
Job Description Word Cloud

Download



Skill sets required for each data job

Download



Recommendation system to match jobs with candidate's profile

- Create a TF-IDF (Term Frequency-Inverse Document Frequency) matrix from the job descriptions. This matrix represents how important each word or term is in each job description and gives higher scores to words that are more relevant and unique to a particular job description.
- Then, we also convert the user's profile text into a TF-IDF vector.
- Next, the cosine similarity between the user's TF-IDF vector and each job description. Cosine similarity is a measure of how similar two vectors are, in this case, the user's profile and each job description.
- The cosine similarity scores range from 0 to 1, where 1 means the vectors are identical, and 0 means they are completely different. So, job descriptions with higher cosine similarity scores to the user's profile are considered more relevant or similar to the user's interests and skills.

```
user_profile = """I am a data scientist with experience in machine learning, Python, and SQL.
I am interested in roles related to predictive modeling, and developing AI solutions."""
```

Recommended postings

	Job_title	Company	Job_link	Job_description
801	Senior Linux Sys...	Canonical	https://www.linkedin.com/jobs/view/senior-linu...	Job DescriptionPosition Description:...
800	Business Analyst...	Jobs for Humanity	https://www.linkedin.com/jobs/view/business-an...	Job DescriptionPosition Description:...
876	Machine Learning...	LeanDNA	https://www.linkedin.com/jobs/view/machine-lea...	Company OverviewLeanDNA is a dynamic software ...
868	Entry-Level AI/M...	Austin Fraser	https://www.linkedin.com/jobs/view/entry-level...	Austin Fraser is supporting a client in the AI...
2152	Junior Data Scie...	Flexon Technologies Inc.	https://www.linkedin.com/jobs/view/junior-data...	Job DescriptionJob Summary:We are se...
1839	Data Scientist I...	CodaMetrix	https://www.linkedin.com/jobs/view/data-scient...	CodaMetrix is revolutionizing Revenue Cycle Ma...
1907	Transmission Dat...	New Leaf Energy, Inc.	https://www.linkedin.com/jobs/view/transmissio...	Job Summary: As an AI Business Analy...
1906	AI Business Anal...	Futurism Technologies, INC.	https://www.linkedin.com/jobs/view/ai-business...	Job Summary:As an AI Business Analyst within o...
3085	Junior Software ...	GliaCell Technologies	https://www.linkedin.com/jobs/view/junior-soft...	We are looking for Machine Learning ...

3. Predictive Modeling using Random Forest

Research Question: What contributes to landing a data job?

Data Preprocessing: The LinkedIn profile data is loaded and preprocessed, including feature engineering, encoding of categorical variables, and imputation of missing values.

Data Splitting:

- The preprocessed data is split into features (X) and labels (y), where y represents the target variable ('Has_job' or 'No job').
- The data is further split into training and testing sets using techniques like train_test_split from scikit-learn.

Model Training:

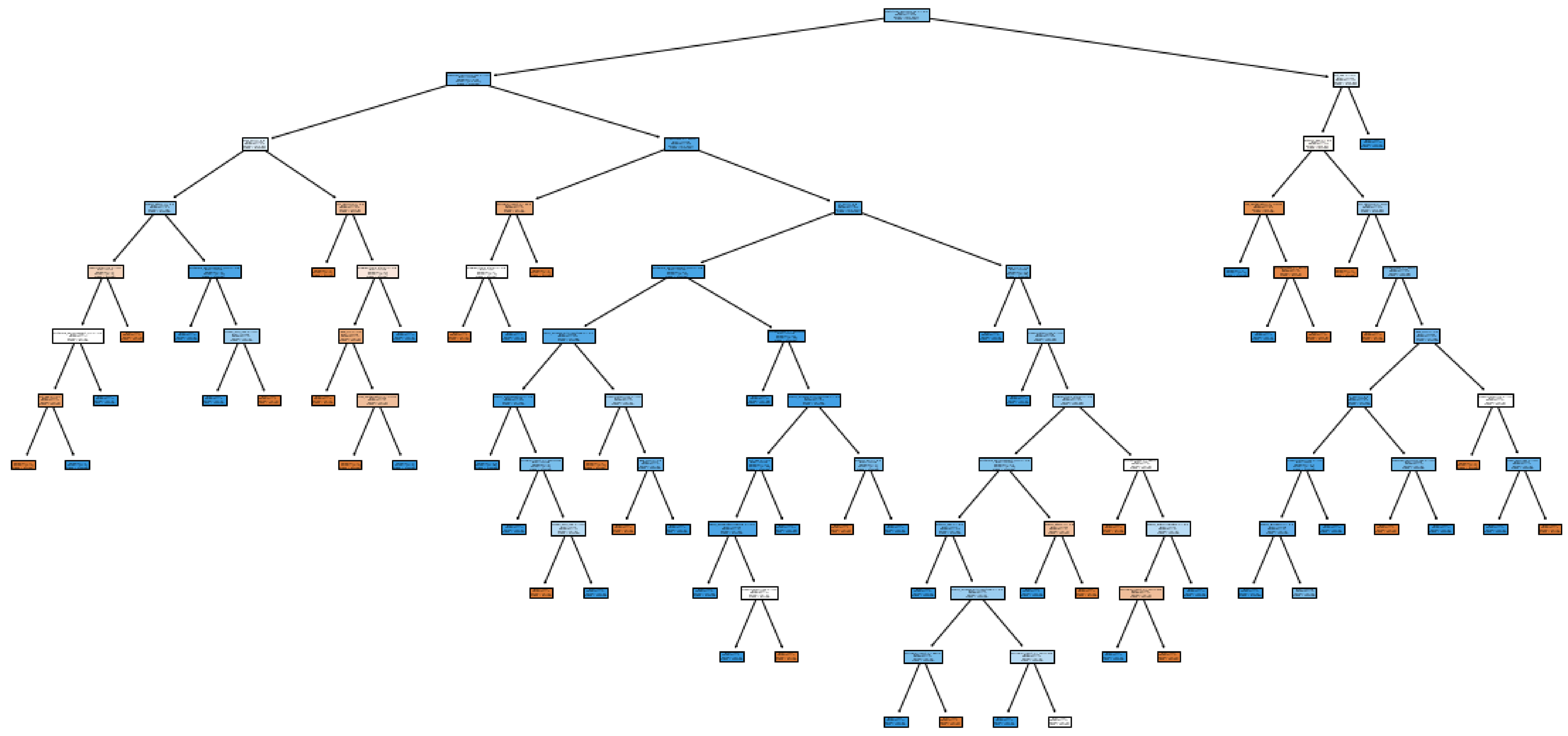
- A Random Forest Classifier is instantiated with appropriate hyperparameters, such as the number of trees (n_estimators) and random state for reproducibility.
- The model is trained on the training data (X_train, y_train) using the fit method.

Model Evaluation:

- The trained Random Forest model is evaluated on the test data (X_test, y_test) by making predictions using the predict method.
- Evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to assess the model's performance.
- The project calculates and visualizes the feature importance scores to identify the most relevant features for predicting the target variable.

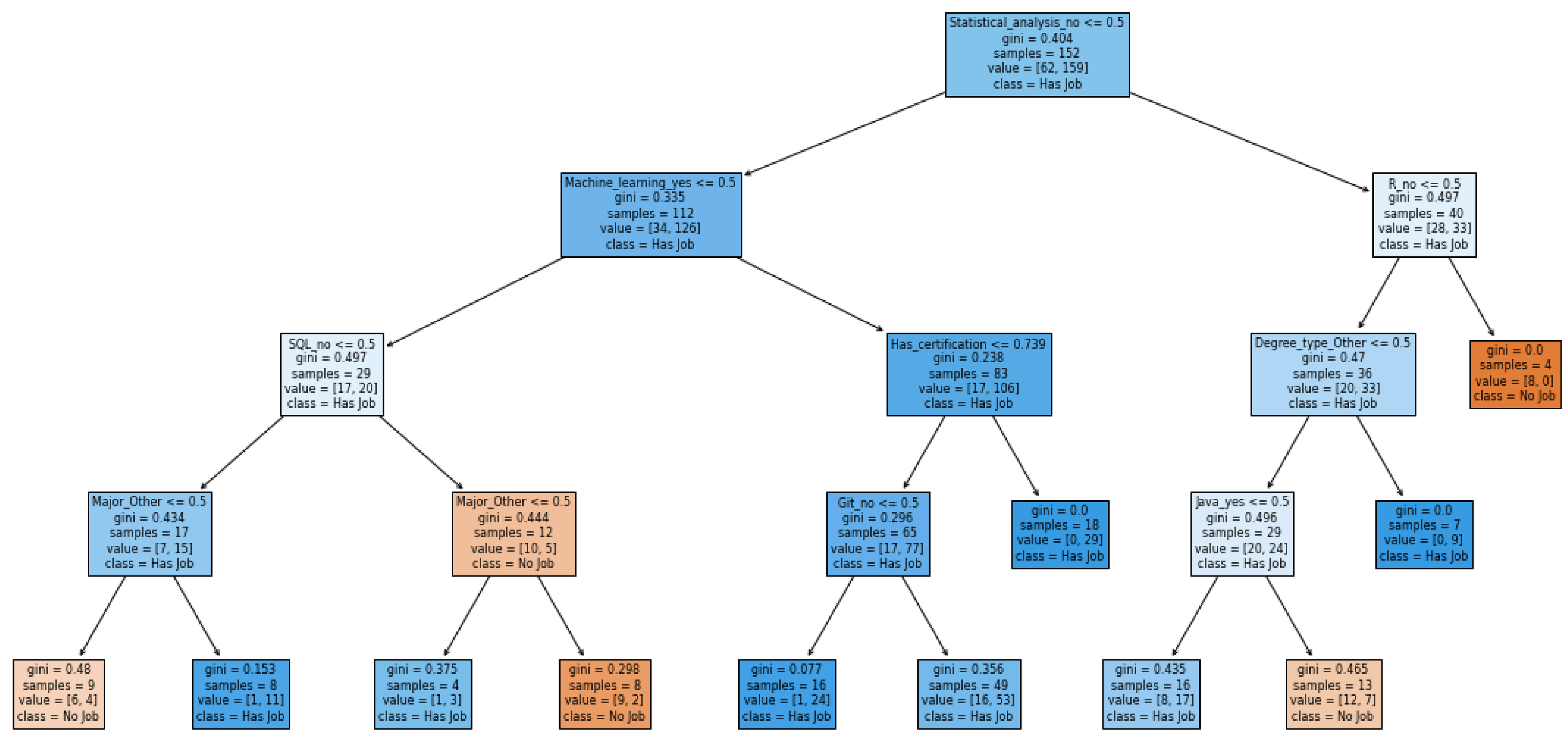
```
▾        RandomForestClassifier
RandomForestClassifier(random_state=0)
```

Visualizing single full decision tree



That looks like quite an expansive tree! Let's limit the depth of trees in the forest to produce an understandable image.

Download



The root node gives us several information :

- There are 152 profiles (`samples = 152`).
- `value = [62,159]` describes the repartition of these profiles among the tree possible classes (i.e. 62 for the 'No Job', 159 for 'Has Job').
- `class = job` . This is the job prospect predicted by the Decision Tree at the root node.
- Gini impurity is a metric that measures the probability from a randomly chosen element (here a profile) to be incorrectly classified.

Model Evaluation

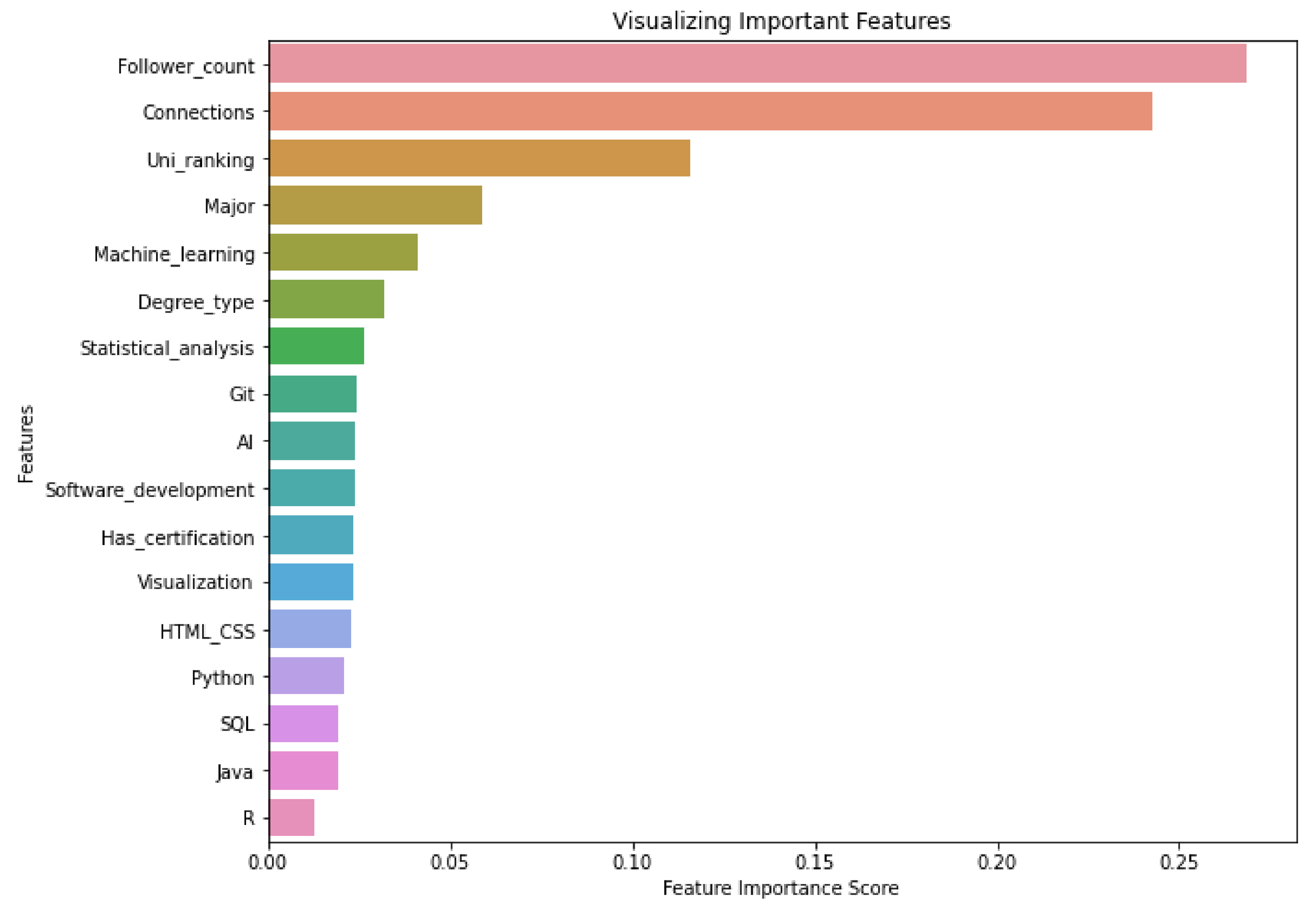
```
Accuracy: 0.92
Precision: 0.97
Recall: 0.91
F1-score: 0.94
```

Feature Importance

```
Follower_count          0.268526
Connections             0.242866
Uni_ranking             0.115728
Major                   0.059006
Machine_learning        0.041271
Degree_type             0.032234
Statistical_analysis    0.026456
Git                     0.024332
AI                      0.023935
Software_development    0.023730
Has_certification       0.023527
Visualization           0.023445
HTML_CSS                0.022668
Python                  0.020746
SQL                     0.019325
Java                    0.019177
R                       0.013026
dtype: float64
```

Visualize feature scores

Download



The feature importance scores emphasize that follower count and connections are the most influential factors, closely followed by university ranking. While skills like machine learning and statistical analysis contribute significantly, proficiency in programming languages such as Python and SQL, though important, holds slightly less weight in predicting job possibilities. Additionally, certifications and expertise in areas like artificial intelligence (AI) remain impactful in enhancing job probability.

4. Implications

4.1. Implications for Stakeholders:

- Job Seekers: The project provides valuable insights into job requirements, skill sets, and the probability of landing a data job based on a candidate's profile. This information can help job seekers tailor their resumes, acquire relevant skills, and apply to suitable job opportunities.
- Employers: The project's analysis of job descriptions and requirements can assist employers in crafting more effective job postings and aligning their expectations with industry standards.
- Educational Institutions: The identification of in-demand skills can help educational institutions update their curricula and prepare students for the job market.
- Career Counselors: The project's findings can aid career counselors in providing better guidance to individuals interested in data-related fields.

4.2. Ethical, Legal, and Societal Implications:

- Ethical Considerations: The project scrapes public LinkedIn profiles, the project does not violate any terms of service or privacy policies.
- Legal Implications: The project's findings can contribute to the development of unbiased hiring policies, promoting equal employment opportunities and preventing discrimination.
- Societal Impact: The project empowers job seekers by providing valuable insights into the skills and qualifications required in the data science and analytics fields, potentially leading to better career opportunities.