IBDA2032 KECERDASAN BUATAN

FAKE JOB POSTING DETECTION

- MARTIN EMMANUEL CHANG
- MIKHA ALDYN YAUW
- PIXEL ARIEL CHRISTOPHER
- MOSES ANTHONY KWIK

IBDA2032 KECERDASAN BUATAN

OVERVIEW

JUDUL	RUMUSAN MASALAH	DATA YANG DIPILIH DAN EDA
DATA PRE-PROCESSING	MODEL MACHINE LEARNING	EVALUASI MODEL DAN VISUALISASINYA
PENYETELAN ALGORITMA	VISUALISASI HASIL	CONCLUSIONS

RUMUSAN MASALAH

Dalam era digital saat ini, platform digital untuk pencarian pekerjaan menjadi semakin populer. Namun, seiring bertumbuhnya platform digital ini, muncul pula masalah tentang adanya lowongan kerja palsu (*fake job posting*). Lowongan palsu dapat merugikan pencari pekerja dengan mengarahkan mereka ke penipuan atau pekerjaan yang tidak sesuai dengan deskripsi aslinya. Maka, kami mengusulkan *Fake Job Posting Detection* untuk memprediksi postingan lowongan kerja palsu.

DATA YANG DIGUNAKAN

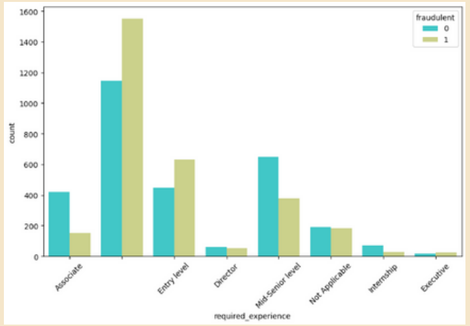
- Sumber data belum pernah dibahas di kelas.
- Dataset dari website KAGGLE.COM dengan judul “Real / Fake Job Posting Prediction”.
- Dataset memiliki 18 fitur.
- 17880 sampel data. Integer dan string.

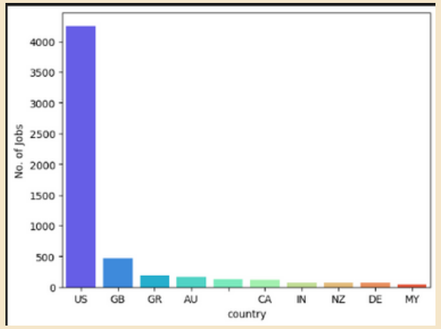
*(keunggulan bonus: data kita terdiri dari kombinasi angka dan huruf dan jumlahnya lebih dari 1000)

EDA

Exploratory Data Analysis

Bar Charts





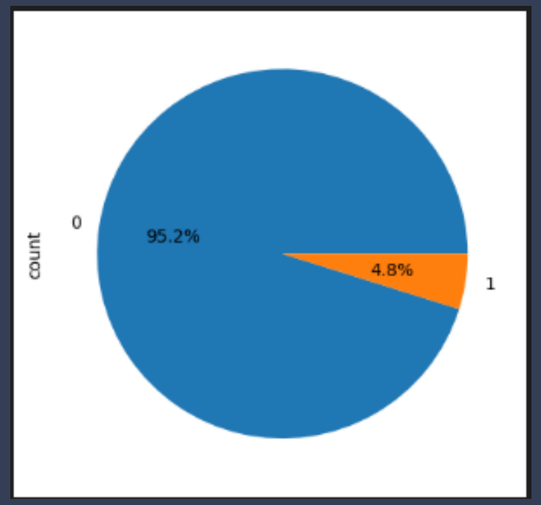
Fraudelent and Real Word Cloud

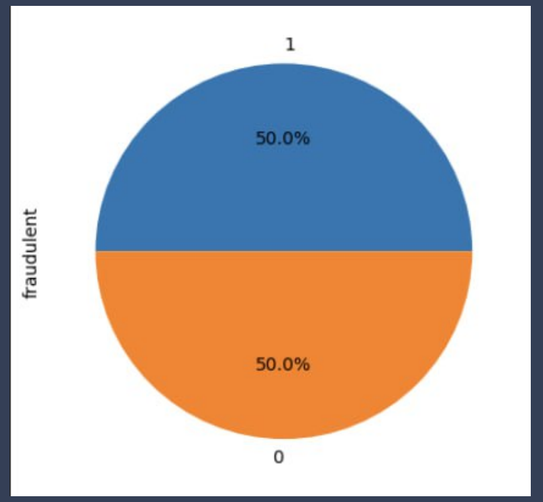
Memberikan 1 jenis visualisasi dengan penjelasan kesimpulan yang tepat mengenai data yang digunakan

Memberikan 1 jenis visualisasi dengan tambahan penjelasan kesimpulan yang tepat mengenai data yang digunakan

Bonus: kita juga punya bar chart untuk mengetahui number of fraud jobs berdasarkan negara ini adalah visualisasi extra yang tidak diminta tetapi dikerjakan

PIE CHART DATA SEPERATION



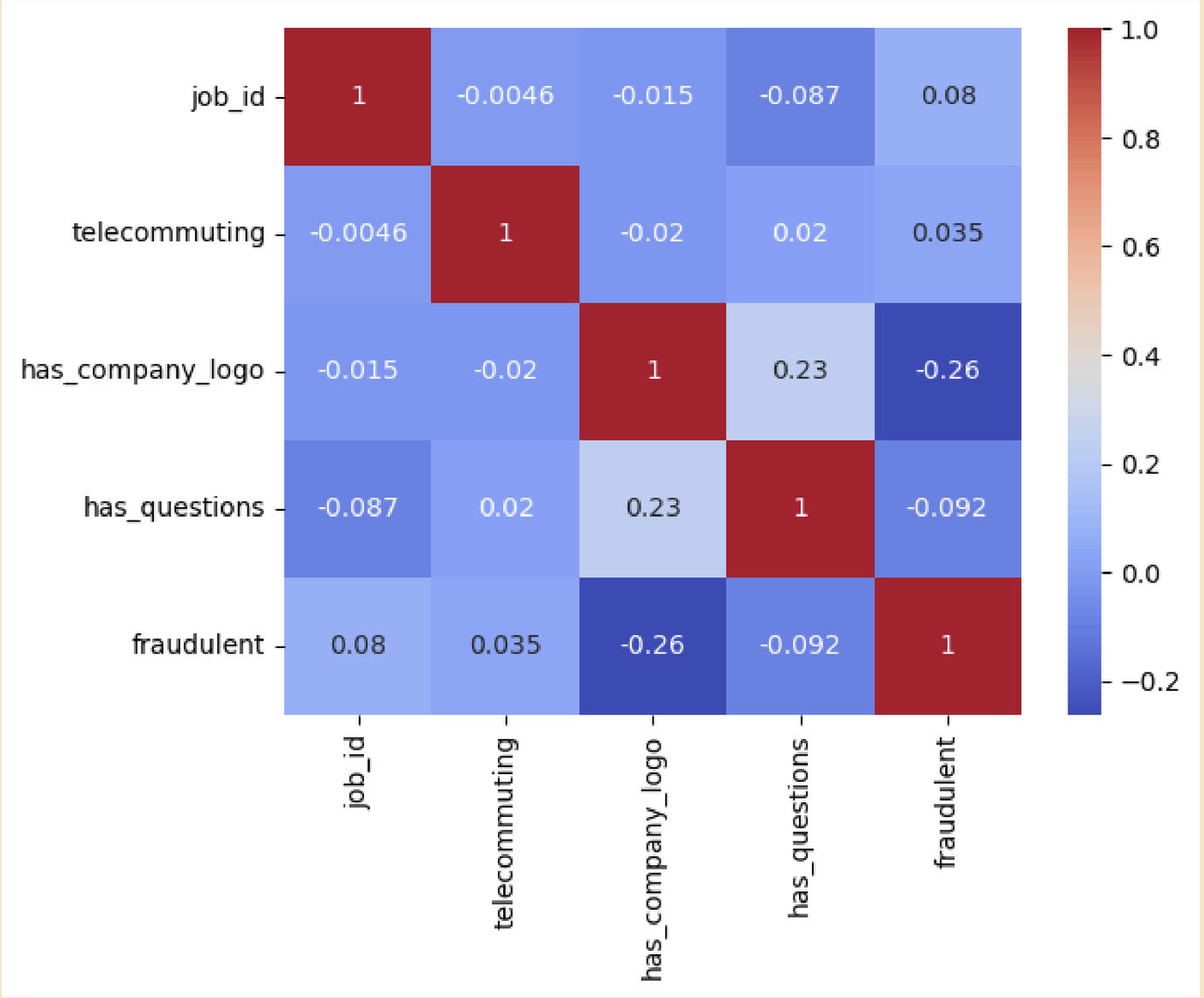


BEFORE

AFTER

DATA PREPROCESSING

PEMILIHAN FITUR



DATA PREPROCESSING

DATA MANIPULATION

01 FEATURE DROP

Menghapus fitur yang korelasi tidak terlalu signifikan terhadap fitur fraudulent melalui analisa heat map.

02 MISSING DATA IMPUTATION

Mengganti isi fitur yang berisi NaN menjadi Null agar dapat diproses oleh komputer.

DATA PREPROCESSING

DATA MANIPULATION

03 SPLIT DATA

Mengambil data nama country dari fitur location. Lalu, dimasukkan ke kolom baru bernama country.

04 BALANCING & SAMPLING DATA

Mengambil sampel data untuk menyeimbangkan jumlah data postingan pekerjaan yang asli dan palsu agar mudah diproses.

DATA PREPROCESSING

DATA MANIPULATION

05 FEATURE COMBINATION

fitur yang datanya berupa string digabung menjadi 1 fitur baru bernama combined_text.

DATA PREPROCESSING

DATA CLEANING

Data cleaning bekerja dengan lowercasing, menghapus html tags, mengahpus URLs, menghapus non-alphanumeric characters, tokenization, menghapus stopword, dan join token-token sebelumnya kembali menjadi string. Data cleaning membantu menyederhanakan text string yang ada.

Kita memilih menggunakan teknik ini karena kita hanya ingin mencari tahu kata-kata apa yang sering muncul pada poster lowongan kerja palsu maupun real. Oleh karena itu, kita tidak memerlukan tanda baca dan sebagainya.

DATA PREPROCESSING

DATA NORMALIZATION

Data normalization bekerja dengan tokenisasi, lemmatisasi, dan penggabungan token kembali. Tokenisasi adalah pemisahan teks menjadi kata-kata individu. Lemmatisasi adalah proses mengubah kata ke bentuk dasarnya, misal walked menjadi walk. Terakhir penggabungan token kembali menjadi satu string.

Kita memilih menggunakan teknik ini karena kita ingin mengganti semua kata menjadi bentuk bakunya. Sehingga, kata majemuk atau turunan yang sebenarnya memiliki akar kata yang sama, diklasifikasikan dalam satu klasifikasi yang sama.

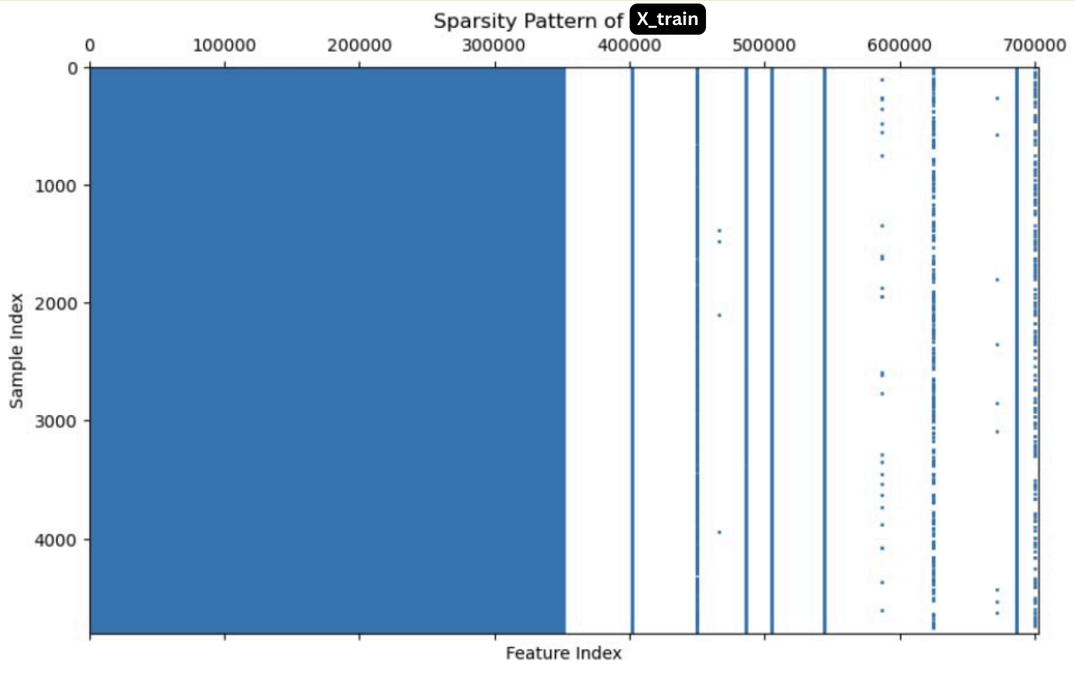
DATA PREPROCESSING

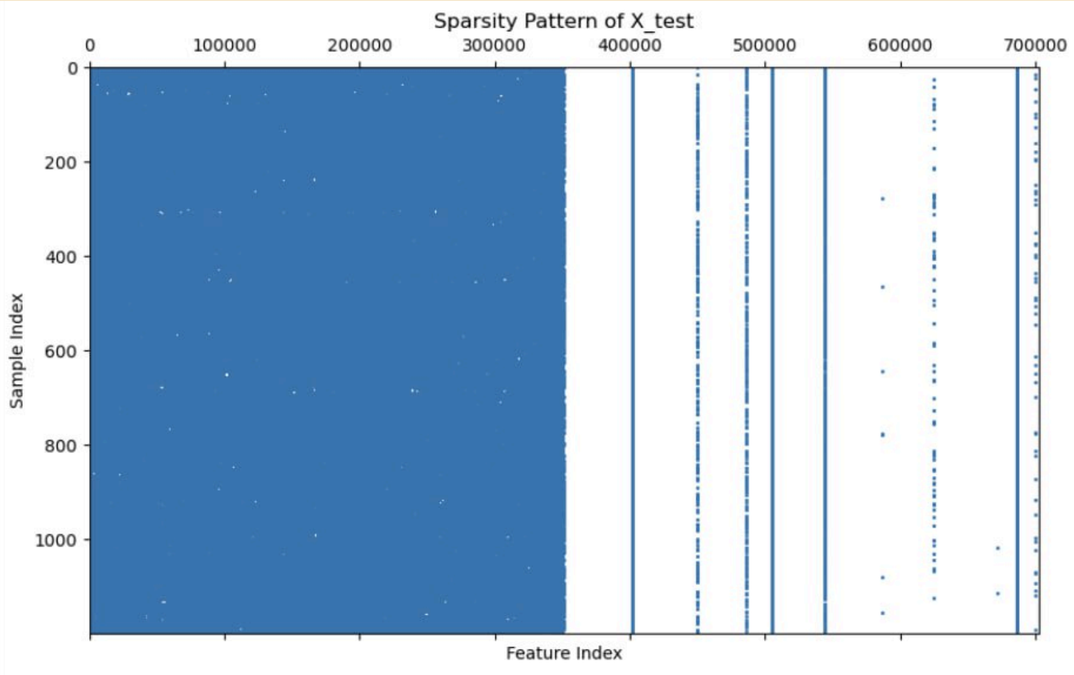
FEATURE EXTRACTION

Feature extraction bekerja dengan cara pembuatan fitur POS (Part of Speech), penggabungan fitur POS, pembuatan matriks fitur, transformasi teks dan fitur POS, penggabungan matriks fitur, penyimpanan vectorizer. Secara keseluruhan, kode ini mengubah teks mentah dan fitur POS menjadi representasi numerik yang dapat dipahami oleh algoritma machine learning.

Kita memilih menggunakan teknik ini karena kita ingin mengklasifikasi setiap kata sesuai part of speech-nya dan mengubahnya menjadi representasi numerik. Hal ini ditujukan untuk memudahkan komputer dalam membaca dan memproses data.

VISUALISASI TRAIN/TEST DATA





Sparsity pattern pada data string dan integer membantu kita melihat di mana terdapat kotak-kotak yang hanya berisi angka atau hanya kata-kata, membantu kita mengetahui di mana harus mencari angka atau kata-kata ketika menggunakan atau menganalisis data tersebut.

PEMILIHAN ALGORITMA

KESESUAIAN ALGORITMA ML DENGAN MASALAH

LOGISTIC REGRESSION

- Untuk klasifikasi binary yang bertujuan untuk prediksi apakah ‘real job posting’ atau ‘fake job posting’.
- Mudah diinterpretasikan karena memberikan estimasi probabilitas instansi pada sebuah kelas.
- Efisien secara komputasi; dapat menangani dataset besar.
- Efektif dalam menangani data teks.

NAIVE BAYES

- Dapat menangani dataset besar dengan baik tanpa mengorbankan sumber daya komputasi yang besar.
- Memprediksi probabilitas keaslian postingan berdasarkan kata-kata yang terkandung di dalamnya.
- Algoritma sederhana dan efisien, sehingga cepat untuk pelatihan dan dijalankan.
- Tidak terlalu dipengaruhi oleh perbandingan sampel antar kelas.

RBF SVM

- Radial Basis Function Support Vector Machine
- Kurang rentan terhadap overfitting. Ini bermanfaat dalam deteksi postingan pekerjaan palsu, di mana model perlu menggeneralisir dengan baik ke data yang belum pernah dilihat sebelumnya.

PEMILIHAN ALGORITMA

KESESUAIAN ALGORITMA ML DENGAN MASALAH

RANDOM FOREST

- Mudah melakukan perhitungan dengan jumlah fitur dan data input yang banyak.
- Anti overfitting karena ensemble learning.
- Mampu membuat decision tree walaupun data kompleks dan tidak linear

EXTRA TREES CLASSIFIER

- Lebih cepat dari Random Forest dalam menangani karena memilih nilai threshold dengan random dan bukannya threshold best-fit.
- Efisien secara komputasi dan dapat menangani dataset besar. Cocok untuk platform online.
- Lebih general dibanding RF.

NEURAL NETWORK 1 LAYER

- Selayaknya pemikiran manusia, Neural Network sangat baik untuk klasifikasi karena meniru pemikiran manusia yang dibilang mahluk pengklasifikasi semua hal karena mandat dari Tuhan.
- Walaupun model training sangat lama namun model akan menghasilkan hasil yang general

PENYETELAN ALGORITMA

PENYETELAN DAN PELATIHAN ALGORITMA

- Penyetelan hyperparameter setiap model machine learning ditentukan setelah banyak try and error.
- Tahap ini memerlukan pengujian dengan cara mempelajari dampak setiap parameter pada model, mengubah parameter, dan melakukan pengecekan ketepatan akurasi.
- Inilah alasan mengapa model kami memiliki ketepatan akurasi dan presisi tinggi.
- Pipeline digunakan untuk membuat data dapat diproses secara parallel sehingga mempercepat training
- Train dan Test split ratio adalah 0.8 untuk training dan 0.2 untuk testing

EVALUASI MODEL

CROSS VAL SCORE

Menggunakan teknik splitting dan folds yang sama seperti k-fold cross validation tetapi cross val score melakukan semua splitting training dan prediction dalam satu step. Method ini juga menghitung score seperti accuracy, precision habis setiap fold.

K-FOLD CROSS VALIDATION

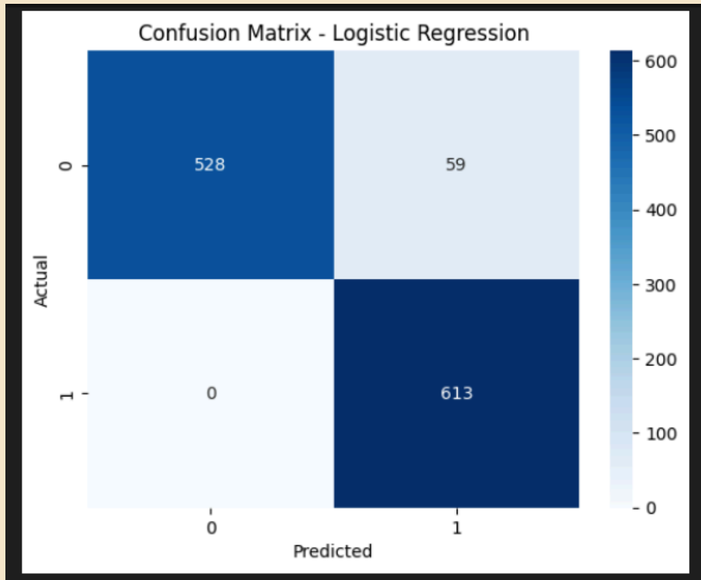
Untuk evaluasi model, original sample data dipisah menjadi subsampel ukuran sama sejumlah K. K-1 subsampel digunakan untuk training dan satu digunakan untuk testing. Ini akan diulang untuk K kali. K-fold memberikan kontrol lebih banyak dalam proses cross-validation tetapi harus manual untuk menunjukan score, train dan predict.

STRATIFIED K-FOLD CROSS VALIDATION

Stratified K-fold adalah variasi dari K_fold dimana setiap set mengembalikan stratified fold yaitu setiap set memiliki persentase sample yang sama seperti target class complete setnya. Ini sangat untuk testing dan training data yang tidak imbang

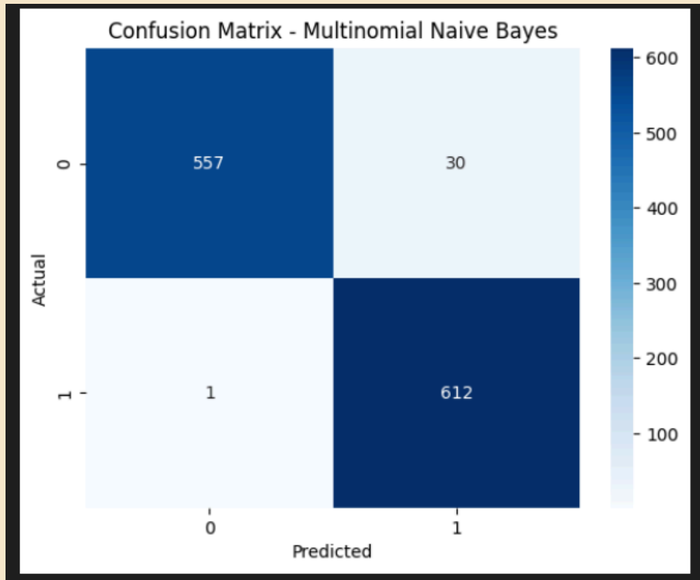
VISUALISASI HASIL

LOGISTIC REGRESSION



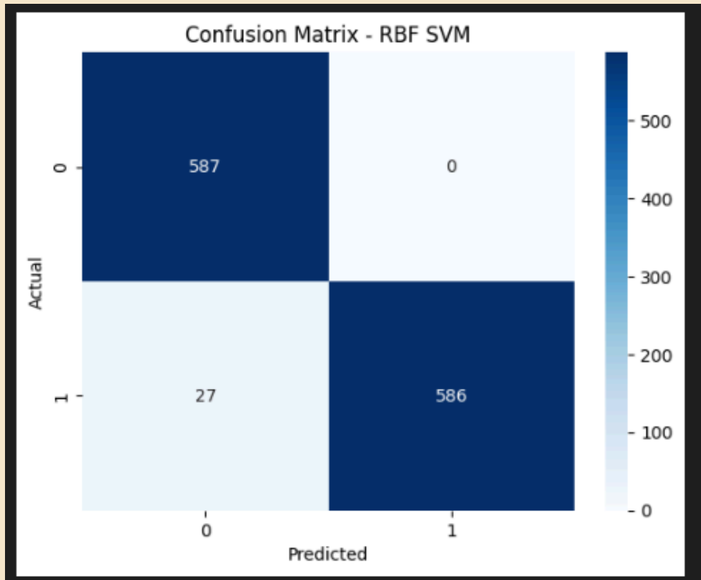

```
Model: Logistic Regression
Test set accuracy: 0.9508333333333333
Test set precision: 0.9561011904761905
Test set recall: 0.9497444633730835
Test set F1 score: 0.9505854024532812
Cross-Validation Scores: [0.94833333 0.96       0.9475     0.94916667 0.96583333]
Mean CV Score: 0.9541666666666666
K-Fold 1 Accuracy: 0.9508333333333333
K-Fold 2 Accuracy: 0.9591666666666666
K-Fold 3 Accuracy: 0.9508333333333333
K-Fold 4 Accuracy: 0.9525
K-Fold 5 Accuracy: 0.9525
Stratified K-Fold 1 Accuracy: 0.9533333333333334
Stratified K-Fold 2 Accuracy: 0.9633333333333334
Stratified K-Fold 3 Accuracy: 0.9483333333333334
Stratified K-Fold 4 Accuracy: 0.9366666666666666
Stratified K-Fold 5 Accuracy: 0.9516666666666667
```

NAIVE BAYES




```
Model: Multinomial Naive Bayes
Test set accuracy: 0.9741666666666666
Test set precision: 0.9757394566710214
Test set recall: 0.9736306766231926
Test set F1 score: 0.9741122844865082
Cross-Validation Scores: [0.9825     0.9875     0.97666667 0.98       0.98333333]
Mean CV Score: 0.982
K-Fold 1 Accuracy: 0.9741666666666666
K-Fold 2 Accuracy: 0.9783333333333334
K-Fold 3 Accuracy: 0.9816666666666667
K-Fold 4 Accuracy: 0.9791666666666666
K-Fold 5 Accuracy: 0.9858333333333333
Stratified K-Fold 1 Accuracy: 0.9841666666666666
Stratified K-Fold 2 Accuracy: 0.9825
Stratified K-Fold 3 Accuracy: 0.9833333333333333
Stratified K-Fold 4 Accuracy: 0.9808333333333333
Stratified K-Fold 5 Accuracy: 0.9766666666666667
```

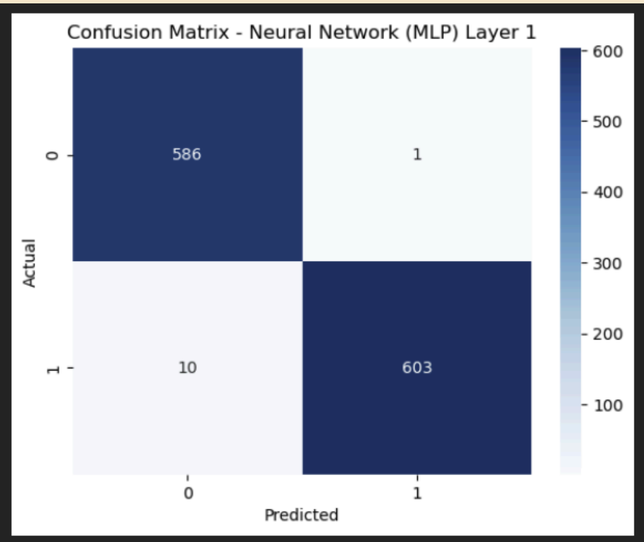
RBF SVM




```
Model: RBF SVM
Test set accuracy: 0.9775
Test set precision: 0.9780130293159609
Test set recall: 0.9779771615008157
Test set F1 score: 0.9774999843749891
Cross-Validation Scores: [0.96583333 0.96333333 0.9675     0.9775     0.96833333]
Mean CV Score: 0.9685
K-Fold 1 Accuracy: 0.9775
K-Fold 2 Accuracy: 0.9666666666666667
K-Fold 3 Accuracy: 0.9691666666666666
K-Fold 4 Accuracy: 0.9566666666666667
K-Fold 5 Accuracy: 0.9708333333333333
Stratified K-Fold 1 Accuracy: 0.9633333333333334
Stratified K-Fold 2 Accuracy: 0.97
Stratified K-Fold 3 Accuracy: 0.9733333333333334
Stratified K-Fold 4 Accuracy: 0.9666666666666667
Stratified K-Fold 5 Accuracy: 0.9766666666666667
```

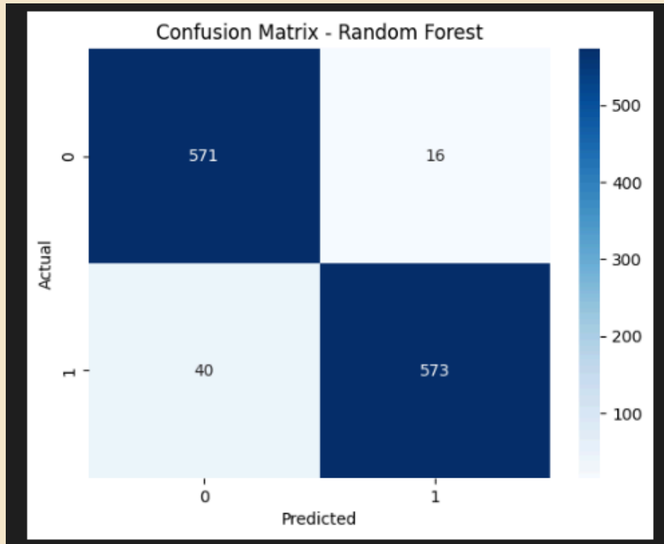
VISUALISASI HASIL

MLP LAYER 1



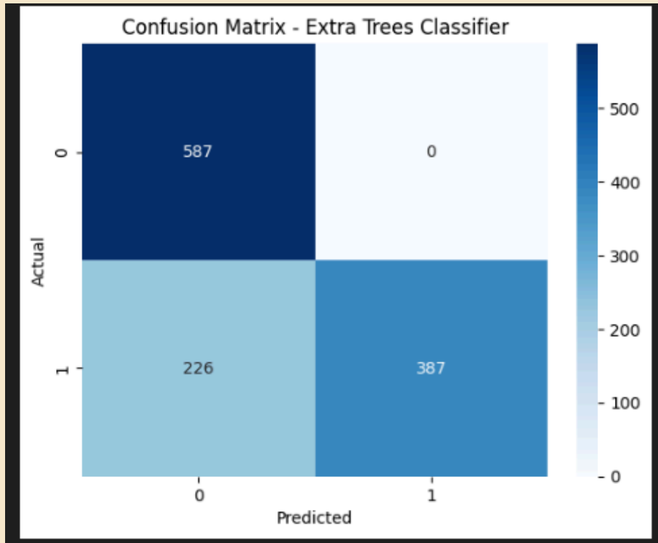

```
Model: Neural Network (MLP) Layer 1
Test set accuracy: 0.9908333333333333
Test set precision: 0.9907829236854971
Test set recall: 0.9909916043920619
Test set F1 score: 0.9908314932649679
Cross-Validation Scores: [0.9875     0.9875     0.99416667 0.9925     0.98916667]
Mean CV Score: 0.9901666666666668
K-Fold 1 Accuracy: 0.9925
K-Fold 2 Accuracy: 0.9916666666666667
K-Fold 3 Accuracy: 0.9958333333333333
K-Fold 4 Accuracy: 0.9858333333333333
K-Fold 5 Accuracy: 0.9875
Stratified K-Fold 1 Accuracy: 0.9916666666666667
Stratified K-Fold 2 Accuracy: 0.9908333333333333
Stratified K-Fold 3 Accuracy: 0.9883333333333333
Stratified K-Fold 4 Accuracy: 0.9883333333333333
Stratified K-Fold 5 Accuracy: 0.995
```

RANDOM FOREST




```
Model: Random Forest
Test set accuracy: 0.9533333333333334
Test set precision: 0.9536844328232545
Test set recall: 0.9537449524915863
Test set F1 score: 0.9533332037033437
Cross-Validation Scores: [0.95333333 0.95583333 0.96083333 0.95916667 0.95166667]
Mean CV Score: 0.9561666666666667
K-Fold 1 Accuracy: 0.9533333333333334
K-Fold 2 Accuracy: 0.9508333333333333
K-Fold 3 Accuracy: 0.9691666666666666
K-Fold 4 Accuracy: 0.96
K-Fold 5 Accuracy: 0.9508333333333333
Stratified K-Fold 1 Accuracy: 0.9441666666666667
Stratified K-Fold 2 Accuracy: 0.9483333333333334
Stratified K-Fold 3 Accuracy: 0.9491666666666667
Stratified K-Fold 4 Accuracy: 0.9566666666666667
Stratified K-Fold 5 Accuracy: 0.9525
```

EXTRA TRESS CLASSIFIER




```
Model: Extra Trees Classifier
Test set accuracy: 0.8116666666666666
Test set precision: 0.8610086100861009
Test set recall: 0.8156606851549755
Test set F1 score: 0.8062857142857143
Cross-Validation Scores: [0.82916667 0.84166667 0.87166667 0.79916667 0.81083333]
Mean CV Score: 0.8305
K-Fold 1 Accuracy: 0.8116666666666666
K-Fold 2 Accuracy: 0.8
K-Fold 3 Accuracy: 0.8591666666666666
K-Fold 4 Accuracy: 0.8391666666666666
K-Fold 5 Accuracy: 0.8666666666666667
Stratified K-Fold 1 Accuracy: 0.8033333333333333
Stratified K-Fold 2 Accuracy: 0.835
Stratified K-Fold 3 Accuracy: 0.8016666666666666
Stratified K-Fold 4 Accuracy: 0.8375
Stratified K-Fold 5 Accuracy: 0.8258333333333333
```

CONCLUSION

Model yang terbaik untuk mendeteksi fake
job posting adalah

MLP LAYER 1

Karena:
Model ini menghasilkan scores yand sangat tinggi dan rata-rata diatas 0.99 semua. Hasil confusion matrix sangat baik dimana hasil False Positive dan Negative sangat sedikit dibandingkan dengan Hasil True Positive dan Negative. Akan tetapi, waktu training model sangat lama dibanding dengan model lain.

THANK YOU

Source: https://www.kaggle.com/code/seifwael123/real-fake-jobs-eda-modelling-99/notebook#Data-Cleaning