LAPORAN MACHINE LEARNING DENGAN DATA SET WINE

Disusun oleh

Nashwan Rasyid Muhammad 5025221004

Mu’aafii Putra Ramadhan 5025221135

Ariq Javier Ramadhani Rahim 5025221267

Kelas Pembelajaran Mesin D

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

2024

DAFTAR ISI

BAB I

PENDAHULUAN

1.1 Latar Belakang

Wine adalah minuman yang berasal dari fermentasi alkohol dari buah-buahan atau umumnya anggur. Minuman ini memiliki kandungan alkohol dikarenakan fermentasi dari ragi yang mengonsumsi gula dari anggur. Dalam laporan kami, kami menggunakan data set dari wine portugis yang bernama "Vinho Verde" dan kami mengolah data set tersebut menggunakan metode-metode beragam seperti Supported-Vector Machine, Ordinal Neural Network, dan Decision Tree. Data set ini tidak meliputi data seperti tipe anggur, brand anggur, dan harga anggur namun menggunakan komposisi yang detail tentang *physicochemical* dan *sensory data* yang menjadi data basis dari data set.

1.2 Rumusan Masalah

- Bagaimana model pembelajaran mesin dapat secara efektif memprediksi kualitas wine berdasarkan sifat *physicochemical*?
- Apa prediktor *physicochemical* yang paling signifikan untuk kualitas wine?
- Metode pembelajaran mesin mana-SVM, ONN, atau Decision Tree-yang memberikan prediksi kualitas wine yang paling akurat dan dapat diandalkan dalam dataset ini?
- Bagaimana ketidakseimbangan dalam kelas kualitas dataset mempengaruhi akurasi prediksi dan kinerja model?

1.3 Batasan Masalah

- Eksklusivitas Data: Data yang terkumpul merupakan data yang berisi dengan *physicochemical* dan *sensory data* sehingga tidak mengandung faktor-faktor seperti komersial maupun sejenis anggur

- Batasan Geografis: Data ini hanya mengandung data yang berasal dari merek wine "Vinho Verde" sehingga hanya memiliki data yang berasal dari portugis dan tidak mencakup jenis wilayah wine lainnya.
- Skala Penilaian Kualitas: Kualitas wine dievaluasi semata-mata berdasarkan skala subjektif 0 sampai 10.
- Ketidakseimbangan Data: Data set memiliki ketidakseimbangan dalam kolom kualitas sehingga model-model ini harus secara khusus disesuaikan dengan ketidakseimbangan sehingga dapat membuat prediksi lebih condong ke kelas yang lebih sering diwakili.

1.4 Tujuan dan Manfaat

a. Tujuan

   i. Untuk membuat sebuah model menggunakan SVM, ONN, dan Decision Tree untuk menilai kualitas wine berdasarkan *physicochemical*-nya.

   ii. Untuk mengidentifikasi sifat *physicochemical* yang mana yang paling signifikan yang mempengaruhi kualitas wine

   iii. Untuk mengevaluasi dan membandingkan efektivitas, akurasi, dan keandalan model yang telah dipilih.

   iv. Untuk memperdalam pemahaman tentang bagaimana berbagai faktor *physicochemical* yang mempengaruhi kualitas wine

b. Manfaat

   i. Kontribusi Akademik: Penelitian ini akan berkontribusi pada pengetahuan dalam pengaplikasian ilmu pembelajaran mesin dalam ilmu makanan

   ii. Kontribusi Industri: Penelitian ini akan berkontribusi dalam membantu pembuat wine dan ahli wine dalam menyempurnakan teknik produksi wine mereka untuk meningkatkan kualitas

iii. Wawasan Metodologi: Dengan membandingkan berbagai model yang berbeda, penelitian ini akan menawarkan informasi tentang kesesuaian berbagai algoritma untuk jenis data tertentu.

BAB II

METODOLOGI

2.1 Dataset yang Digunakan

Kami menggunakan dataset Red Wine Quality yang diterbitkan oleh University of California, Irvine Machine Learning Repository di Kaggle karena dataset ini terstruktur dengan baik. Setiap atribut memiliki label yang jelas, meminimalkan waktu yang diperlukan untuk preprocessing dan memaksimalkan waktu untuk pengaplikasian dan analisis. Selain itu, dataset ini memiliki hasil output yang ordinal, yaitu klasifikasi kualitas dari skala 0 sampai 10, hal ini memberikan output yang lebih mendetail daripada nilai “Ya” atau “Tidak”. Juga, ukuran dataset yang berjumlah 1599 data memberikan sampel data yang cukup tanpa memerlukan waktu yang terlalu lama untuk melakukan training. Kombinasi kejelasan atribut, output ordinal, dan ukuran yang pas menjadi alasan kami untuk menggunakan dataset ini untuk di analisa

Dengan output yang ordinal, hal ini memberikan kita tantangan untuk lebih kreatif dalam mengaplikasikan dataset ini terhadap model machine learning, beserta membolehkan pilihan machine learning model yang fleksibel seperti Support Vector Machine, Decision Tree, dan Ordinal Neural Network, yaitu Artificial Neural Network dengan output berbentuk ordinal, atau memiliki orde (seperti 1-10).

2.2 Desain Sistem (Flowchart)

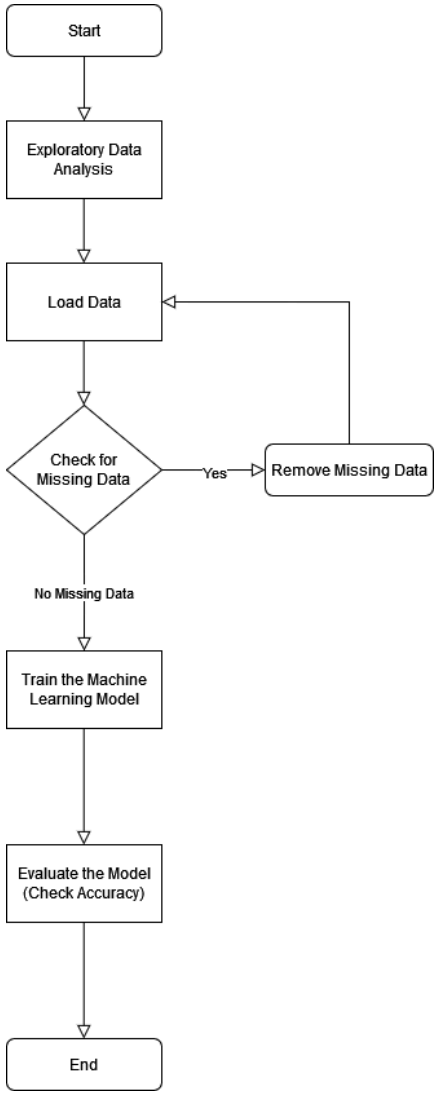


Chart 2 Flowchart of Machine Learning Process

2.3 Analisis Data

Dari dataset yang kami gunakan, data tersebut tidak tersebar secara merata. Setelah dilakukan penganalisis lebih lanjut, data tidak tersebar secara merata. Parameter target memiliki nilai ordinal 3 - 8 dimana persebarannya terpusat pada 5 dan 6. Untuk mengatasinya, diimplementasikan metode SMOTE agar persebarannya merata. SMOTE nantinya akan memeriksa data minoritas (jumlah tersedikit) dan mencoba mencocokkan dengan data lainnya. SMOTE nantinya akan membuat data lain yang memiliki kemiripan mendekati data minoritas

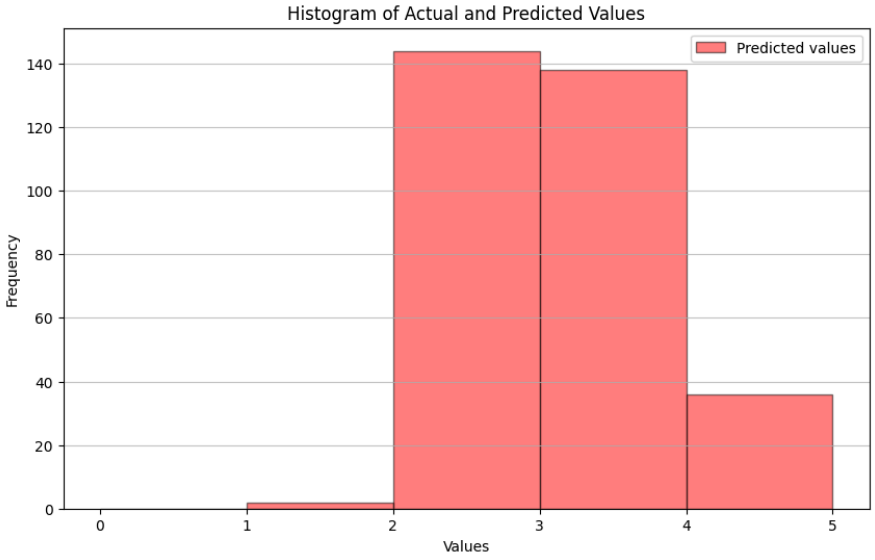
tersebut. Alhasil, data akan memiliki persebaran yang lebih merata dan data yang lebih banyak untuk di *train*.

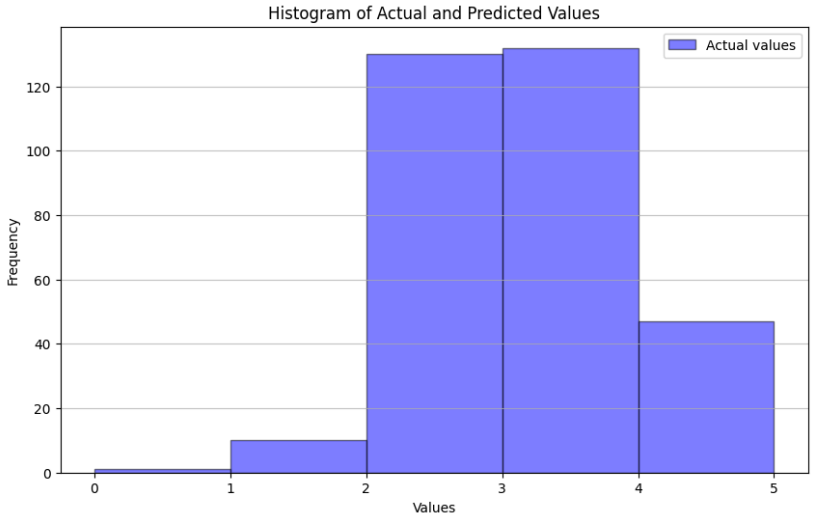
BAB III

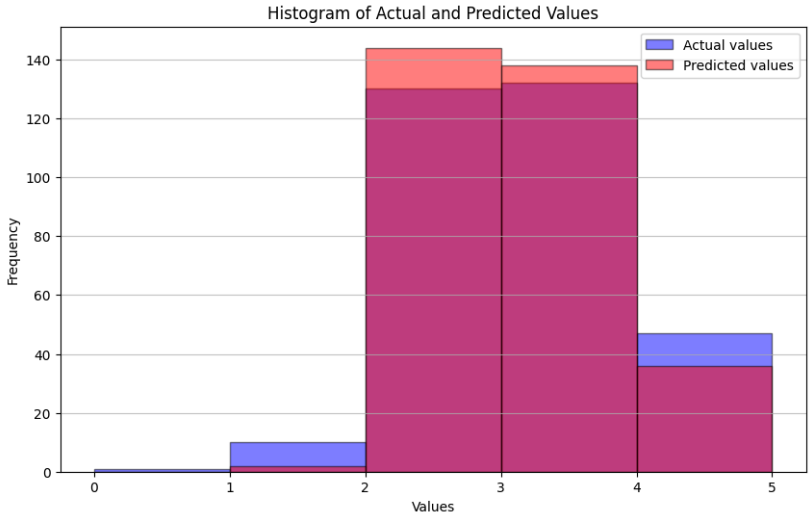
HASIL DAN PEMBAHASAN

3.1 Skenario Pengujian ANN

Pengujian Artificial Neural Network (ANN) terhadap 20% akhir dataset menghasilkan beberapa metrik evaluasi yang signifikan. Model ini mencapai akurasi sebesar 62%. Untuk memvisualisasikan perbandingan antara actual values dan predicted values, kami menggunakan histogram. Histogram menunjukkan distribusi nilai prediksi dan aktual. Dari analisis kesalahan, terlihat bahwa sebagian besar kesalahan prediksi terjadi pada nilai yang lebih rendah, di mana model cenderung overpredict nilai dengan skor tinggi.







Dengan akurasi sebesar 62%, model ini kurang andal untuk digunakan dalam konteks prediksi kualitas anggur. Hasil tabel ini menunjukkan bahwa model dapat secara memberikan prediksi yang mendekati nilai aktual, yang sangat penting dalam industri anggur untuk penilaian kualitas. Namun, ada beberapa limitasi yang perlu diperhatikan. Data yang digunakan mungkin memiliki beberapa keterbatasan seperti ketidakseimbangan kelas yang dapat mempengaruhi hasil. Selain itu, ada kemungkinan beberapa fitur penting belum dimasukkan dalam model, dan model yang lebih kompleks mungkin diperlukan untuk meningkatkan akurasi lebih lanjut.

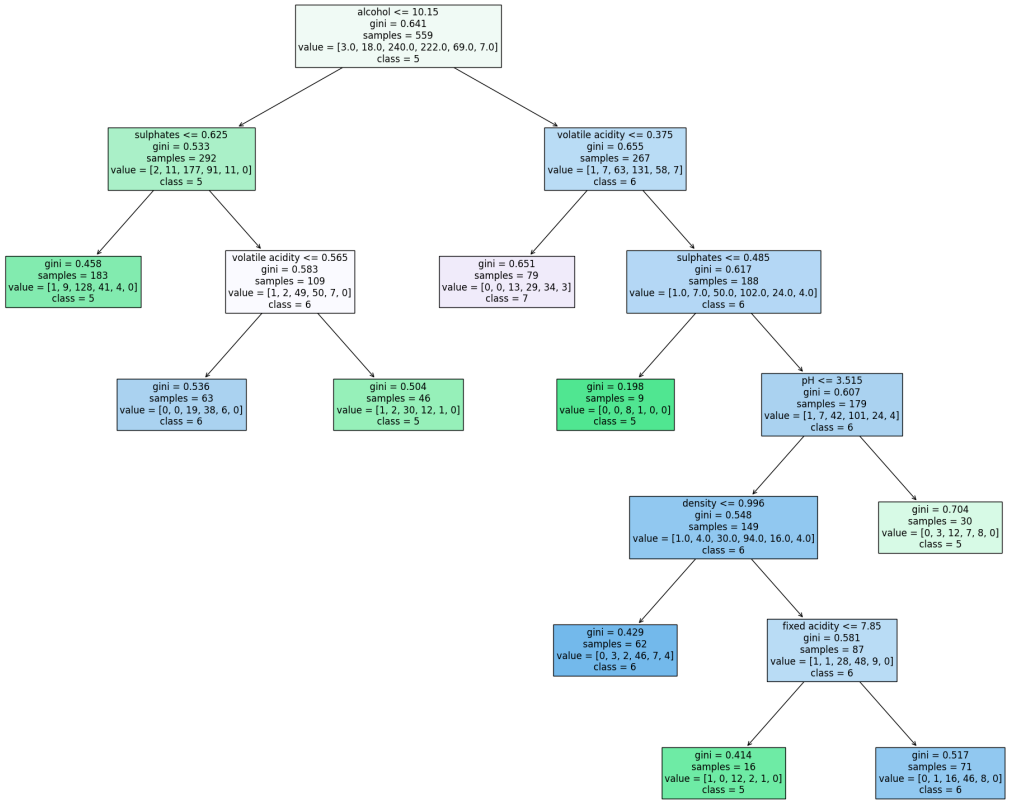
Secara keseluruhan, model ANN yang dikembangkan menunjukkan hasil yang cukup dengan akurasi 60%. Model ini menunjukkan potensi yang kuat untuk digunakan dalam prediksi

kualitas anggur, meskipun ada beberapa area yang dapat ditingkatkan untuk hasil yang lebih optimal.

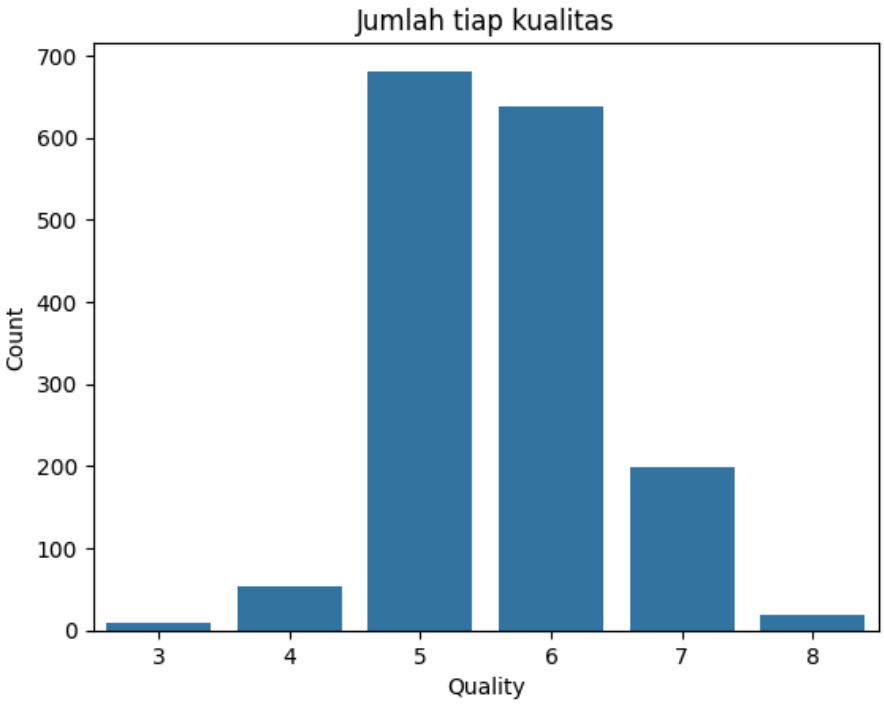
```
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.6570 - loss: 1.0428
Test Accuracy: 62.50%
Random Forest Regressor Test Accuracy (rounded): 0.625
```

3.2 Skenario Pengujian Decision Tree

Pengujian Decision Tree terhadap 20% dari total data pada dataset memberikan hasil yang menarik. Model ini memiliki akurasi dapat mencapai rata-rata 88% ketika menggunakan *cross validation.* Untuk memvisualisasikan decision tree yang sudah dibuat, menggunakan library sklearn dan meng-import tree. Analisis dari tree yang sudah dibuat menunjukkan bahwa data yang jumlahnya signifikan lebih kecil terkadang tidak terdeteksi lokasi penempatannya dalam tree.



Setelah diperiksa lagi datasetnya, memang hasil pada parameter target yang kami cari memusat pada sebuah nilai. Dataset yang digunakan memiliki tipe data ordinal dari 0 - 10. Walau begitu, parameter target yang dicari hanya dari kisaran 3 - 8. Setelah diperiksa lagi, hasil paling banyak ditemukan pada kisaran 5 - 7 dan yang lainnya jauh lebih sedikit. Hal ini membuat datanya tersebar tidak merata

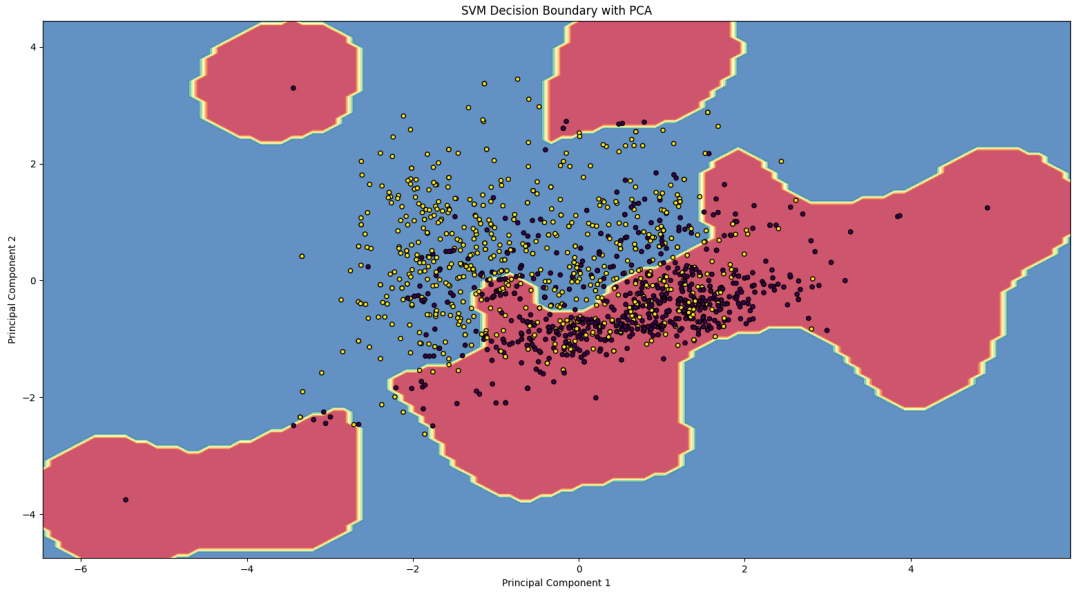


Untuk menangani masalah tersebut, perlu ditambahkan fungsi SMOTE untuk membuat data lebih merata. Sebagai tambahan, digunakan juga metode Random Forest Classifier sebagai metode pengklasifikasian. Metode ini membandingkan beberapa contoh tree yang sudah dibuat untuk mencapai hasil yang lebih memuaskan.

```
Cross-validation scores for each fold: [0.86706949 0.8489426  0.89123867 0.89425982 0.88217523 0
 0.88787879 0.89393939 0.91515152 0.89393939]
Mean cross-validated score: 88.20516341664379 %
```

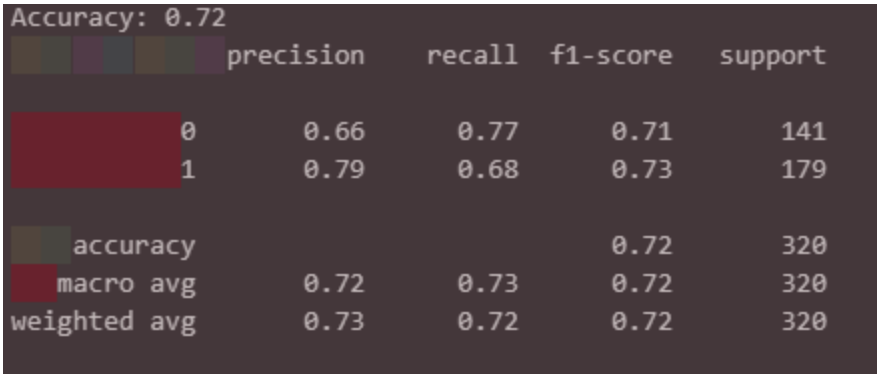
3.3 Skenario Pengujian Supported Vector Machine (SVM)

Data set yang digunakan telah dibagi menjadi training set dan test set dengan masing-masing perbandingan 80/20. Dengan test set yang sebesar 20% memberikan hasil akurasi sebesar 72% walau sudah di-*finetune* dengan SMOTE dan GridSearchCV. Hasil grafik dapat dilihat sebagai berikut:



Grafik 3.5 Hasil SVM

Dari hasil grafik tersebut dapat terlihat bahwa wine yang terklasifikasi sebagai *low quality* wine memiliki bulat yang berwarna kuning dan background biru, dan yang terklasifikasi sebagai *high quality* wine terklasifikasi memiliki bulat yang berwarna maroon dan background merah. Namun terlihat dari hasil grafiknya terdapat banyak noise yang berada di kedua bagian (merah dan biru). Hal ini dapat disebabkan oleh berbagai macam faktor seperti: Variabel yang *inherent* terhadap wine, variabel seperti alkohol dan sulfat dapat menunjukkan rentang yang *overlap* di seluruh variabel kualitas karena adanya faktor lain yang tidak dapat dimodelkan yang mempengaruhi kualitas wine; PCA, walau PCA dapat mengurangi dimensi ke sumbu yang paling informatif, pengurangan dimensi ini dapat membuat variabel-variabel menjadi *overlapping*.



Gambar 3.3 Hasil Akurasi SVM

Hasil SVM memberikan kami akurasi sebesar 72% dari testing set. Metrik ini memberikan bayangan kapabilitasnya SVM untuk mengklasifikasi wine menjadi *high quality* wine dan *low quality* wine. Untuk *low quality* wine kami mendapatkan *precision* dan *recall* masing-masing 66% dan 77%, hal ini menjelaskan bahwa SVM dapat mengidentifikasi wine sebagai *low quality* dengan kebenaran sebesar 66% dan berhasil mengidentifikasi 77% wine sebagai *low quality* wine dari wine keseluruhan. Sedangkan untuk *high quality* wine kami mendapatkan *precision* dan *recall* masing-masing 79% dan 68%, hal ini menunjukan bahwa SVM dapat mengidentifikasi wine sebagai *high quality* dengan kebenaran sebesar 79% dan berhasil mengidentifikasi 68% wine sebagai *high quality* wine dari wine keseluruhan. F1-Score memiliki akurasi masing-masing sebesar 71% dan 73% , hal ini berguna dalam situasi dimana ada sebuah ketidakseimbangan di antara distribusi variabel atau ketika adanya *false positive* atau *false negative*. F1-score membantu kami untuk memastikan bahwa model mempertahankan keseimbangan antara tidak melewatkan wine berkualitas tinggi (recall) dan tidak salah mengklasifikasikan wine berkualitas rendah sebagai wine berkualitas tinggi (presisi).

Untuk file–file model kami dapat diakses di link github kami.

BAB IV

KESIMPULAN

Setelah dipaparkan pembahasan dan analisis dapat disimpulkan bahwa:

- Ordinal Neural Network menghasilkan akurasi sekitar 62%. Model cenderung mengalami overfitting dan lebih sering memprediksi nilai kualitas yang lebih tinggi secara berlebihan. Analisis histogram menunjukkan adanya kecenderungan model untuk memprediksi nilai kualitas tinggi secara berlebihan, yang mengindikasikan distribusi kesalahan yang miring ke arah ekstrem skala kualitas. Meskipun terdapat keterbatasan seperti potensi ketidakseimbangan kelas dan penghilangan fitur penting, model ini menunjukkan kemampuan prediksi yang memadai dalam konteks penilaian kualitas wine dan memiliki potensi apabila dikembangkan lebih lanjut.
- Model Decision Tree kami, yang telah dimodifikasi dengan *cross-validation*, mencapai akurasi rata-rata sebesar 88%. Namun, model ini menghadapi tantangan subset data yang lebih kecil yang tidak ditempatkan secara efektif di dalam pohon. Hal ini menunjukkan adanya keterbatasan dalam menangani rentang data yang jarang dalam data set, terutama terkonsentrasi di antara nilai kualitas 5 dan 7. Untuk mengatasi ketidakseimbangan data tersebut, kami menggunakan SMOTE untuk menyeimbangkan data set, bersama menggunakan Random Forest Classifier untuk meningkatkan akurasi prediksi melalui pembelajaran ensemble.
- Model SVM, yang dioptimalkan dengan SMOTE dan GridSearchCV, menunjukkan akurasi sebesar 72% pada set pengujian. Dari grafis, terlihat model ini menunjukkan no0ise yang cukup besar, yang menunjukkan adanya tumpang tindih di antara variabel-variabel yang dapat dikaitkan dengan pengurangan dimensi oleh PCA dan variabilitas yang melekat pada variabel-variabel wine. Terlepas dari tantangan ini, model ini menunjukkan tingkat *precision* dan *recall* yang wajar, dengan kemampuan yang nyata untuk membedakan antara *high*

quality wine dan *low quality* wine meskipun ada beberapa kesalahan klasifikasi yang terlihat dari F1 *score*.

Dapat ditarik kesimpulan, meskipun setiap model telah menunjukkan potensi dalam memprediksi kualitas wine, namun masih ada beberapa area yang perlu ditingkatkan. Namun dari ketiga model yang menunjukkan yang paling "baik" adalah Decision Tree, dimana Decision Tree menunjukkan akurasi yang paling tinggi dengan benefit dari kesederhanaan dan kemampuan interpretasi yang sangat berharga dalam aplikasi yang membutuhkan pemahaman tentang model decisions

DAFTAR PUSTAKA

Wurz, D. A. (2019). Wine and health: A review of its benefits to human health. *BIO Web of Conferences, 12*, 04001. https://doi.org/10.1051/bioconf/20191204001

Kulkarni, Vrushali Y. (2013). Random Forest Classifiers :A Survey and Future Research Directions. https://adiwijaya.staff.telkomuniversity.ac.id/files/2014/02/Random-Forest-Classifiers_A-Survey-and-Future.pdf