AT25-41253: Statistical Machine Learning

Tugas: Eksperimen Hyperparameter Tuning SVM

Program Studi Sains Aktuaria
Institut Teknologi Sumatera

Pengampu: Martin C.T. Manullang

Diterbitkan: 26 Oktober 2025

Informasi Penting

Batas Waktu Pengumpulan: Silakan cek website mata kuliah
Link Pengumpulan: https://mctm.web.id/course/at25-41253
Mode Pengerjaan: Individual (direkomendasikan untuk nilai lebih tinggi) atau Kelompok (maksimal 3 orang)
Yang Harus Dikumpulkan:

- File Jupyter Notebook (.ipynb) - **harus sudah dijalankan sebelum dikumpulkan**
- File PDF hasil export dari notebook (.pdf)

1 Gambaran Umum

Tugas ini dirancang untuk membantu Anda mendapatkan pengalaman hands-on dalam melakukan hyperparameter tuning untuk Support Vector Machine (SVM) pada masalah regresi. Anda akan melakukan eksperimen dengan berbagai konfigurasi hyperparameter, menganalisis dampaknya terhadap performa model, dan memberikan insight dari perspektif aktuaria.

2 Tujuan Pembelajaran

Dengan menyelesaikan tugas ini, Anda akan mampu:

- Memahami dampak berbagai hyperparameter SVM terhadap performa model
- Merancang dan melaksanakan eksperimen hyperparameter tuning secara sistematis
- Menganalisis dan menginterpretasikan metrik evaluasi model (R^2, MAE, RMSE, MAPE)
- Menerapkan critical thinking untuk memilih model optimal dalam aplikasi aktuaria
- Mendokumentasikan eksperimen machine learning secara profesional

3 Terms of Reference

3.1 Dataset

Gunakan Medical Insurance Cost Dataset yang telah disediakan di repository mata kuliah:

- `medical_insurance.csv` - Dataset utama
- `train_data.csv` dan `validation_data.csv` - Data yang sudah di-split (jika tersedia)

Opsi Penggunaan Subset Data:

- Anda **boleh menggunakan 50%-100%** dari dataset untuk training
- Gunakan subset data (50%-70%) untuk eksperimen awal agar training lebih cepat
- **Wajib menggunakan 100%** data untuk model final yang dilaporkan
- Pastikan menggunakan **stratified sampling** agar distribusi data tetap seimbang
- Di notebook, set variable `DATA_PERCENTAGE` sesuai kebutuhan

3.2 Deskripsi Tugas

Tugasmu adalah **meningkatkan performa dari baseline SVM model** yang telah didemonstrasikan dalam materi hands-on (`week10-2.ipynb`) dengan melakukan eksperimen hyperparameter tuning secara sistematis.

3.2.1 Performa Baseline (Referensi)

Model baseline dari sesi hands-on menggunakan:

- Kernel: RBF
- C: 1.0
- Epsilon: 0.1
- Gamma: 'scale'

Tujuanmu adalah mencapai **performa yang lebih baik** dari baseline ini melalui eksperimen yang cermat.

3.3 Eksperimen yang Harus Dilakukan

Anda harus melakukan dan mendokumentasikan eksperimen berikut:

3.3.1 Eksperimen 1: Perbandingan Kernel (20 poin)

- Test minimal **dua kernel berbeda**: pilih dari 'linear', 'rbf', atau 'poly'
- Gunakan hyperparameter default atau nilai yang reasonable untuk setiap kernel
- Bandingkan metrik performa antar kernel
- Berikan analisis kernel mana yang bekerja paling baik dan **mengapa**

3.3.2 Eksperimen 2: Hyperparameter Tuning (30 poin)

- Pilih kernel terbaik dari Eksperimen 1
- Lakukan tuning sistematis untuk minimal **tiga hyperparameter**:
  - C (Regularization parameter) - test minimal 4 nilai
  - epsilon (Epsilon-tube width) - test minimal 3 nilai
  - gamma (untuk kernel RBF/poly) - test minimal 4 nilai
- Coba minimal **3 kombinasi** dan maksimal **6 kombinasi** hyperparameter (di luar Eksperimen 1)
- Gunakan **GridSearchCV** untuk tuning
- Dokumentasikan parameter grid yang digunakan
- Laporkan best parameter yang ditemukan dan cross-validation scores

3.3.3 Eksperimen 3: Evaluasi Model (20 poin)

- Train model final dengan best hyperparameter
- Evaluasi menggunakan **semua** metrik berikut:
  - R^2 Score (coefficient of determination)
  - MAE (Mean Absolute Error)
  - RMSE (Root Mean Squared Error)
  - MAPE (Mean Absolute Percentage Error)
- Bandingkan dengan performa baseline model
- Cek overfitting (bandingkan metrik training vs validation)

3.3.4 Eksperimen 4: Visualisasi & Analisis (15 poin)

- Buat minimal **tiga visualisasi**:
  1. Scatter plot Actual vs Predicted
  2. Residual plot
  3. Chart perbandingan model (baseline vs tuned)
- Berikan interpretasi untuk setiap visualisasi

3.4 Struktur Laporan

Buat Jupyter Notebook dengan struktur berikut (tulis dalam Markdown cells):

1. **Judul dan Pendahuluan (5 poin)**
   - Judul tugas, nama Anda, NIM
   - **Jika pengerjaan kelompok:** Tuliskan nama dan NIM semua anggota, serta **deskripsi kontribusi masing-masing anggota** (siapa mengerjakan apa)
   - Penjelasan singkat tentang masalah yang dihadapi
   - Nyatakan tujuan eksperimenmu
2. **Data Loading dan Preparation (5 poin)**
   - Load dataset
   - Set `DATA_PERCENTAGE` (gunakan 50%-100% sesuai kebutuhan eksperimen)
   - Lakukan preprocessing yang diperlukan (encoding, scaling)
   - Split data jika belum di-split
   - *Catatan: Anda boleh reuse preprocessing code, tapi jelaskan setiap langkah*
3. **Eksperimen 1: Perbandingan Kernel**
   - Code untuk testing berbagai kernel
   - Tabel hasil perbandingan kernel
   - Analisis dan interpretasi (dalam Markdown)
4. **Eksperimen 2: Hyperparameter Tuning**
   - Code untuk GridSearchCV
   - Definisi parameter grid (minimal 3 kombinasi, maksimal 6 kombinasi)
   - Best parameter yang ditemukan
   - Hasil cross-validation
   - Analisis dampak parameter
5. **Eksperimen 3: Evaluasi Model**
   - Code training model final
   - Metrik evaluasi lengkap
   - Tabel perbandingan (baseline vs tuned)
   - Analisis overfitting
6. **Eksperimen 4: Visualisasi & Analisis**
   - Semua plot yang diminta dengan label yang proper
   - Interpretasi dari setiap visualisasi
7. **Insight Aktuarial (10 poin)**

- Diskusikan implikasi praktis untuk pricing asuransi
- Bagaimana Anda akan menggunakan model ini di production?
- Apa keterbatasannya?
- Rekomendasi untuk deployment

8. Kesimpulan (5 poin)

- Rangkum temuan utama
- Nyatakan konfigurasi model terbaik
- Persentase peningkatan dari baseline
- Pelajaran yang didapat

9. Referensi dan Sitasi

- Sitasi sumber eksternal yang digunakan
- Jika menggunakan AI tools, berikan atribusi yang jelas
- Sitasi dataset (sudah disediakan di materi kuliah)

4 Rubrik Penilaian

Komponen	Poin	Kriteria
Eksperimen 1: Perbandingan Kernel	20	Kelengkapan, kedalaman analisis
Eksperimen 2: Hyperparameter Tuning	25	Desain grid, best params, insight
Eksperimen 3: Evaluasi Model	15	Metrik lengkap, perbandingan
Eksperimen 4: Visualisasi	15	Kualitas, interpretasi
Insight Aktuarial	10	Relevansi praktis, kedalaman
Kualitas Laporan	15	Struktur, kejelasan, penggunaan markdown
Kualitas Code	10	Readability, komentar, eksekusi
Bonus Individual	+10	Untuk pengerjaan individual
Total	**110/100**	

Catatan: Pengerjaan individual mendapat bonus poin. Nilai akhir maksimal 100.

5 Persyaratan Teknis

5.1 Persyaratan Code

- Library yang diperlukan: pandas, numpy, scikit-learn, matplotlib, seaborn
- Code harus diberi komentar dengan baik
- Semua cell harus sudah dieksekusi sebelum dikumpulkan
- Tidak ada error dalam notebook
- Gunakan `DATA_PERCENTAGE` untuk mengatur persentase data (50%-100%)

5.2 Persyaratan Markdown

- Gunakan heading yang proper (`#`, `##`, `###`)
- Sertakan penjelasan di antara code cells
- Gunakan bullet point dan numbered list dengan tepat
- Format equation menggunakan LaTeX (jika diperlukan): `$$R^2 = ...$$`
- Sertakan tabel untuk perbandingan hasil

5.3 Struktur Notebook

- Mulai dari **notebook kosong** (jangan copy seluruh hands-on notebook)
- Tulis laporan dan eksperimenmu sendiri
- Anda boleh reuse preprocessing code, tapi jelaskan
- Fokus pada **eksperimenmu** dan **analisismu**

6 Panduan Pengumpulan

6.1 Konvensi Penamaan File

- Notebook: `SVM_Assignment_[NIM].ipynb`
  Contoh: `SVM_Assignment_121450001.ipynb`
- Untuk kelompok: `SVM_Assignment_[NIM1]_[NIM2].ipynb` (untuk 2 orang) atau `SVM_Assignment_[NIM1]_[NIM2]_[NIM3].ipynb` (untuk 3 orang)
  Contoh: `SVM_Assignment_121450001_121450002_121450003.ipynb`
- PDF: Nama yang sama dengan extension .pdf

6.2 Cara Generate PDF dari Notebook

1. **Jalankan semua cell** di notebook
2. Di Jupyter: File → Download as → PDF via LaTeX
3. Atau gunakan command line:
   `jupyter nbconvert --to pdf notebook_Anda.ipynb`
4. Atau di VS Code: Export → PDF

6.3 Yang Harus Dikumpulkan

1. File Jupyter Notebook (.ipynb) - **harus sudah dieksekusi penuh**
2. File PDF hasil export dari notebook (.pdf)
3. Kedua file harus memiliki konten yang identik

6.4 Platform Pengumpulan

Kumpulkan melalui website mata kuliah: https://mctm.web.id/course/at25-41253
Cek deadline di website mata kuliah.

7 Integritas Akademik

7.1 Penggunaan AI Tools

Kebijakan Penggunaan AI

Anda BOLEH menggunakan AI tools (ChatGPT, Copilot, dll.) untuk membantu pembelajaran, TAPI:

- Anda **bertanggung jawab penuh** atas setiap baris code yang Anda kumpulkan
- Anda harus **memahami** cara kerja code tersebut
- Anda harus bisa **menjelaskan** code-mu jika ditanya
- Berikan **atribusi yang jelas** di notebook-mu:
  Contoh: "Code untuk hyperparameter tuning dibantu oleh ChatGPT"
- Jangan copy-paste tanpa memahami

Ingat: AI adalah alat untuk membantumu belajar, bukan menggantikan pembelajaran.

7.2 Kebijakan Kolaborasi

- Pengerjaan individual: Kerjakan semuanya sendiri (nilai lebih tinggi dengan bonus +10)
- Pengerjaan kelompok: Maksimal 3 orang. Semua anggota harus berkontribusi setara
- **Wajib mencantumkan kontribusi setiap anggota** di bagian pendahuluan notebook
- Jangan share code antar kelompok
- Jangan copy dari submission tahun sebelumnya
- Sitasi semua sumber eksternal yang digunakan

7.3 Plagiarisme

Plagiarisme akan berakibat:

- Nilai nol untuk tugas ini

- Dilaporkan ke bagian akademik
- Potensial gagal mata kuliah

8 Tips untuk Sukses

1. **Mulai lebih awal!** Jangan tunggu sampai deadline
2. **Pahami baseline dulu** sebelum mencoba meningkatkannya
3. **Dokumentasi sambil jalan** - tulis penjelasan markdown sambil coding
4. **Gunakan subset data (50%-70%)** untuk eksperimen awal - set `DATA_PERCENTAGE`
5. **Test dengan grid kecil dulu** - GridSearchCV bisa lama
6. **Save pekerjaan secara berkala** - gunakan version control (git)
7. **Cek error** - pastikan semua cell berjalan tanpa error
8. **Tanya jika bingung** - gunakan office hours atau forum mata kuliah

9 Contoh Parameter Grid

Berikut beberapa saran untuk GridSearchCV (Anda tidak harus menggunakan nilai-nilai ini):

9.1 Untuk Eksperimen Awal (Grid Kecil - 3 Kombinasi)

```
param_grid = {
    'C': [1, 10],
    'epsilon': [0.1],
    'gamma': ['scale', 'auto']
}
# Total: 2 × 1 × 2 = 4 kombinasi (cukup untuk test awal)
```

9.2 Untuk Eksperimen Komprehensif (Grid Sedang - 6 Kombinasi)

```
param_grid = {
    'C': [1, 10, 100],
    'epsilon': [0.01, 0.1],
    'gamma': ['scale', 'auto']
}
# Total: 3 × 2 × 2 = 12 kombinasi (reasonable untuk final model)
```

Catatan:

- Gunakan `DATA_PERCENTAGE = 0.5` atau `0.7` untuk grid yang lebih besar
- Untuk model final, gunakan `DATA_PERCENTAGE = 1.0` dengan best parameters
- GridSearchCV dengan 3-fold CV akan training model 3 kali per kombinasi

10 Formula Metrik Evaluasi

Untuk referensimu:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4}$$

Dimana:

- y_i = nilai aktual
- $\hat{y}_i$ = nilai prediksi
- $\bar{y}$ = mean dari nilai aktual
- n = jumlah sampel

11 Sumber Referensi

- Materi kuliah: `week10-1.ipynb` dan `week10-2.ipynb`
- Dokumentasi Scikit-learn: https://scikit-learn.org/stable/
- Dokumentasi SVM: https://scikit-learn.org/stable/modules/svm.html
- GridSearchCV: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- Website mata kuliah: https://mctm.web.id/course/at25-41253

12 Pertanyaan yang Sering Diajukan (FAQ)

Q: Bolehkah saya menggunakan dataset yang berbeda?
A: Tidak, Anda harus menggunakan Medical Insurance Cost Dataset yang disediakan.

Q: Bolehkah saya skip bagian data preprocessing?
A: Anda harus menyertakan preprocessing, tapi boleh reuse code dari materi hands-on dengan penjelasan yang proper.

Q: Berapa lama GridSearchCV akan berjalan?
A: Tergantung ukuran grid dan persentase data. Dengan `DATA_PERCENTAGE = 0.5` dan grid 3-6 kombinasi, biasanya 5-15 menit. Mulai dengan grid kecil untuk test.

Q: Bolehkah saya menggunakan kurang dari 100% data untuk model final?
A: Tidak, model final yang dilaporkan harus menggunakan 100% data (`DATA_PERCENTAGE = 1.0`). Subset data hanya untuk eksperimen awal.

Q: Bagaimana jika model tuned saya performanya lebih buruk dari baseline?
A: Dokumentasikan temuan ini! Jelaskan mengapa hal ini bisa terjadi dan apa yang Anda pelajari.

Q: Bolehkah saya menambahkan eksperimen ekstra di luar yang diminta?
A: Ya! Pekerjaan ekstra yang berkualitas bisa mendapat bonus poin (maksimal +5).

Q: Bagaimana jika saya kerja kelompok tapi ingin kredit individual?
A: Kumpulkan terpisah. Setiap orang harus punya notebook sendiri dengan eksperimen sendiri.

Q: Berapa jumlah anggota kelompok yang diperbolehkan?
A: Maksimal 3 orang per kelompok. Semua anggota harus mencantumkan kontribusi masing-masing di bagian pendahuluan notebook.

Q: Bagaimana cara sitasi AI tools?
A: Tambahkan section di akhir: "AI Tools yang Digunakan: ChatGPT untuk saran code di Bagian X. Semua code sudah direview dan dipahami sebelum dimasukkan."

13 Kontak

Untuk pertanyaan tentang tugas ini:

- Cek website mata kuliah: https://mctm.web.id/course/at25-41253
- Kirim chat di group kelas
- Atau kirimkan email ke: martin.manullang@if.itera.ac.id
- Dosen mungkin akan meminta bantuan asisten untuk menjawab pertanyaan umum

Selamat mengerjakan eksperimen!
Ingat: Tujuannya adalah belajar, bukan hanya mendapat nilai tertinggi.
Pemahaman lebih penting daripada performa.