2)

برای محاسبه درخت تصمیم با استفاده از الگوریتم ID3 در مجموعه داده ، باید این مراحل را دنبال کنیم:

محاسبه آنتروپی: ناخالصی در مجموعه داده را اندازه گیری می کنیم.

محاسبه سود اطلاعات: کاهش آنتروپی را با تقسیم مجموعه داده بر اساس ویژگی های مختلف اندازه گیری می کنیم.

بهترین ویژگی را انتخاب کنید: برای تقسیم مجموعه داده، ویژگی با بالاترین اطلاعات را انتخاب می کنیم.

تکرار برای زیرمجموعه ها: به صورت بازگشتی فرآیند را به زیر مجموعه های مجموعه داده اعمال می کنیم تا زمانی که معیارهای توقف برآورده شوند.

3)

Precision، نسبت مشاهدات مثبت پیش‌بینی‌شده درست به کل مثبت‌های پیش‌بینی‌شده است.

در یک تنظیمات چند کلاسه، دقت را می توان برای هر کلاس به صورت جداگانه محاسبه کرد و سپس با استفاده از میانگین گیری میکرو یا میانگین گیری کلان، دقت کلی را محاسبه کرد.

Recall، نسبت مشاهدات مثبت پیش بینی شده درست به همه مشاهدات در کلاس واقعی است.

4)

نمرات کامل 1.0 برای دقت، یادآوری و اندازه گیری F نشان می دهد که طبقه بندی کننده درخت تصمیم پیش بینی های دقیقی را در مجموعه آزمون انجام داده است.

این مدل در داده‌های آموزشی و احتمالاً داده‌های آزمایشی عملکرد فوق‌العاده‌ای دارد، اما نمی‌تواند به داده‌های جدید و دیده نشده تعمیم یابد.

5)

overfitt

امتیازات F1 از 0.2857 تا 1.0 متغیر است، که نشان می دهد عملکرد مدل به شدت به زیرمجموعه خاصی از داده هایی که روی آن آموزش داده شده است وابسته است.

این واریانس بالا نشان می دهد که مدل بسیار نزدیک به داده های آموزشی است و به خوبی به داده های دیگر تعمیم نمی یابد.

عمق درخت 5 ممکن است برای مجموعه داده کوچکی از 25 نمونه نسبتاً عمیق باشد. درختان عمیق‌تر می‌توانند الگوهای پیچیده‌تری را در داده‌های آموزشی ثبت کنند، اما به احتمال زیاد overfitt می‌شوند.

مدل‌های overfitt عملکرد فوق‌العاده‌ای در داده‌های آموزشی دارند، اما در تعمیم به داده‌های جدید شکست می‌خورند.

میانگین امتیاز F1 0.496 نشان می دهد که به طور متوسط، عملکرد مدل کمتر از حد بهینه است، که بیشتر از این ایده حمایت می کند که مدل ممکن است نویز را به جای الگوهای اساسی در داده ها ضبط کند.